Can We Undo Our First Impressions? The Role of Reinterpretation in Reversing Implicit Evaluations

Thomas C. Mann and Melissa J. Ferguson
Cornell University

Little work has examined whether implicit evaluations can be effectively "undone" after learning new revelations. Across 7 experiments, participants fully reversed their implicit evaluation of a novel target person after reinterpreting earlier information. Revision occurred across multiple implicit evaluation measures (Experiments 1a and 1b), and only when the new information prompted a reinterpretation of prior learning versus did not (Experiment 2). The updating required active consideration of the information, as it emerged only with at least moderate cognitive resources (Experiment 3). Self-reported reinterpretation predicted (Experiment 4) and mediated (Experiment 5) revised implicit evaluations beyond the separate influence of how thoughtfully participants considered the new information in general. Finally, the revised evaluations were durable 3 days later (Experiment 6). We discuss how these results inform existing theoretical models, and consider implications for future research.

Keywords: implicit evaluations, attitudes, reinterpretation, AMP, IAT

What happens when our initial information about someone turns out to be wrong? Can we change how we feel about someone who upon further examination is nothing like our first impression? Consider the case of a member of the Nazi party in the late 1930s who took over a formerly Jewish owned business to produce supplies for the German war effort, employing Jews as a cheap source of labor. Learning such details about this person would lead most people to detest him for enabling the Germans during the Holocaust. In the end, however, it turned out that this man—Oskar Schindler—used all his money and connections to keep his Jewish workers from being killed, deliberately producing essentially zero materials for the war effort in his factory and ending up destitute from his efforts to protect his workers (Crowe, 2004; Steinhouse, 1994). As such, Schindler represents a dramatic case of an everyday idea: that people sometimes act with ulterior motives that, when revealed, prompt us to reinterpret their earlier actions. In terms of one of the most basic ways we assess the world around us—evaluation—how do revelations such as these influence us? Are we able to "undo" our first impressions and change our minds about the goodness or badness of others?

The ease of this kind of change may depend on which kind of evaluation is meant: explicit or implicit. *Explicit evaluations* are those measured directly, such as when someone endorses a statement about preference (e.g., "I like her."). Explicit evaluations seem to be quite capable of reflecting newly learned truths that override earlier information: one can simply choose to reject an old evaluation and endorse a new one (Gawronski & Bodenhausen, 2006). *Implicit evaluations* are those measured indirectly, which means instead of asking people how they feel about a stimulus, the researcher *infers* it by assessing whether the perception of that stimulus facilitates responses to a different, unrelated stimulus.[1] As it turns out, implicit evaluations are not as easily reversed when prior impressions are found to be false (e.g., Boucher & Rydell, 2012; Gregg, Seibt, & Banaji, 2006; Peters & Gawronski, 2011; see also Wilson, Lindsey, & Schooler, 2000). This work suggests that our implicit first impressions may be relatively hard to undo, persisting even after we learn new information that should override them (see also Rule, Tskhay, Freeman, & Ambady, 2014).

In the present work, we offer a fresh look at this question of whether and how implicit evaluations can be updated to reflect newly learned truths. Given that implicit evaluations uniquely shape and predict behavior (Cameron, Brown-Iannuzzi, & Payne, 2012; Ferguson, 2007; Greenwald, Banaji, & Nosek, 2014; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; McNulty, Olson, Meltzer, & Shaffer, 2013; Perugini, Richetin, & Zogmaister, 2010; Towles-Schwen & Fazio, 2006; cf. Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013), it is important to know whether—or when—they can be reconciled with one's reasoning about what is

This article was published Online First March 23, 2015.

Thomas C. Mann and Melissa J. Ferguson, Department of Psychology, Cornell University.

This research was supported by National Institutes of Health Grant 1R21AG042662-01A1 awarded to Melissa J. Ferguson, and a National Science Foundation Graduate Research Fellowship awarded to Thomas C. Mann.

Correspondence concerning this article should be addressed to Thomas C. Mann, Department of Psychology, Cornell University, G73 Uris Hall, Ithaca, NY 14853. E-mail: tcm79@cornell.edu

[1] Throughout this article, we use the term *implicit evaluation* because it refers to effects—that is, indirectly measured unintentional evaluative responses. The term *implicit attitude*, while often used in this literature, is ambiguous in that it might refer either to behavioral effects or the mental constructs posited to explain them (see discussion in De Houwer, Gawronski, & Barnes-Holmes, 2013).

 http://dx.doi.org/10.1037/pspa0000021

true of the world. In what follows, we describe recent findings that speak to this question, and review what contemporary models of evaluations would predict about this possibility. We then discuss a heretofore little-examined possibility: that, as in the case of Schindler, when new information forces a reinterpretation of the prior impression, reversal of implicit evaluations may be possible.

Can Implicit Evaluations Be Undone?

Implicit evaluations were initially assumed to be difficult to change, let alone completely "undo." They were thought of as the products of long-term exposure to information in one's environment (Greenwald & Banaji, 1995) and were assumed to persist in memory even after new attitudes formed (Wilson et al., 2000). More recent work, however, suggests that implicit evaluations *can* sometimes be altered. For example, some have argued that implicit evaluations are enabled by associative processes that entail the spreading of activation through networks based on spatial or temporal proximity, or semantic similarity (Conrey & Smith, 2007; Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006). Changing these associations has sometimes been assumed to occur through the repeated pairing of the attitude object with counterattitudinal information (Rydell & McConnell, 2006; Rydell, McConnell, Mackie, & Strain, 2006; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007).

In line with these assumptions, some researchers have used extensive evaluative conditioning paradigms to try to modify existing implicit evaluations. For example, Karpinski and Hilton (2001) exposed participants to 200 trials of counterconditioning to try to modify implicit evaluations of the elderly. In this method, change is assumed to depend on the repeated spatial and temporal co-occurrence of the stimulus and evaluative cues (cf. Mitchell, De Houwer, & Lovibond, 2009). If a stimulus, such as a social group, was originally implicitly evaluated as negative, for example, perhaps repeatedly pairing group members with positive cues—without any context, explanation, or reasoning—might nudge the evaluation toward positivity. To be sure, this method has shown that such change of implicit evaluations through evaluative conditioning is possible (e.g., Karpinski & Hilton, 2001; Lai et al., in press; Olson & Fazio, 2006; Rydell et al., 2006).

The question of the current work, however, is whether we can effectively undo implicit evaluations when reasons to doubt our initial impressions come to light. Sometimes we learn new things about the world that immediately transform the meaning of prior knowledge, and to what degree can such revelations alter our impressions? This kind of learning differs from the mere repeated *pairing* of the attitude object with new information, in that it depends on considerations of truth and falsehood. For instance, new details might emerge about a person suggesting that a first impression of him or her was entirely incorrect, and that some different impression is warranted instead. Beyond adding something new to one's representation of that person, such revelations can suggest that other aspects of an impression should be subtracted. As such, processing the new information might entail figuring out whether to endorse the new information as valid and true, and how the new information is related to older information (Gawronski & Bodenhausen, 2006). Below, we describe theory and empirical work addressing the question of whether implicit evaluations can be "undone" through the affirmation of new impressions as true, the negation of old impressions as false, or the combination of the two.

Revision Through the Addition of New Information

Some studies have tried to change implicit impressions by providing new information about a target that is both different in valence from, as well as totally unrelated to, the initial information (Gawronski et al., 2010; Petty et al., 2006, Study 1; Rydell & Gawronski, 2009; Rydell & McConnell, 2006; Rydell et al., 2007). In these studies, participants are typically first presented with a large number of evaluatively consistent statements about a target person (e.g., "Bob donates his time at the soup kitchen"), and subsequently display the expected implicit and explicit evaluation toward that person. Then, researchers attempt to change overall impressions by presenting new statements with the opposite valence (e.g., "Bob refused to help a child fix his bike") that are seemingly unrelated to the initial statements. In this task, participants play an active role in affirming the validity of each new piece of information (Rydell & McConnell, 2006), and so the new information is vetted as accurate. However, this approach of adding new information tends to lead to implicit revision only after considerable amounts of countervailing information is presented, and at a much slower rate than what is needed for explicit revision (e.g., Rydell et al., 2007). To date, the one exception to this is when the new information is extremely negative and rare (Cone & Ferguson, 2014). In these studies, after learning a large amount of mildly positive pieces of information about a new person (e.g., "Bob gave a hitchhiker a ride to a shelter"), participants immediately reversed their implicit evaluations of the person after learning a piece of extremely negative and rare information (e.g., "Bob was recently convicted of molesting children").

Although the evidence for revision through adding new, unrelated information is somewhat mixed, some theories claim that this approach of adding new information is the most likely way in which implicit evaluations can be updated. The Associative–Propositional Evaluation model (APE; Gawronski & Bodenhausen, 2006, 2011) contends that implicit evaluations are generated through associative processing, but can be updated after learning new information that is deemed valid. When people learn about (and believe) new, counterattitudinal information about a target, for instance, this can create a new counterattitudinal association, which might then drive the implicit evaluation. Importantly, this model assumes that adding new information in this way is the most likely route to changing implicit evaluations because it is very difficult to overturn, or silence, older associations that were the basis for the initial impression. However, it is not yet clear when such new information, even if fully believed and affirmed, will have a relatively small or large impact on implicit evaluations. Extreme negativity and diagnosticity (Cone & Ferguson, 2014) may be one criterion that is necessary for new, unrelated information to lead to revision.

These studies that simply add new, unrelated information to the totality of information about the person (or stimulus) might be classified as "addition" studies. They represent cases where we learn new information about someone that is countervailing to, but independent of, our former impressions. This approach assumes that change will emerge incrementally from the totality of information about the stimulus. In this way, even though adding an

extreme piece of information may occasionally swamp out older information (Cone & Ferguson, 2014), new information will tend to be added to older information, often resulting in evaluatively complicated representations (i.e., implicit ambivalence; Petty et al., 2006) or contextualized evaluations (Gawronski, Rydell, Vervliet, & De Houwer, 2010; Rydell & Gawronski, 2009), which can allow for the recovery of the initial association with a shift in context (Gawronski et al., 2010; see Bouton, 2004).

Revision Through the Undoing of Initial Information

What about when new information forces a change in meaning of the initial information? Is it possible to revise implicit impressions by *undoing* the meaning of previously learned information? In other words, can we effectively "erase" the influence of our implicit associations on the basis of new information? Although some theory maintains that this will be very hard to do (Gawronski & Bodenhausen, 2006, 2011; see also Deutsch, Gawronski, & Strack, 2006; Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008), other perspectives claim it is possible. For example, some theoretical work maintains that implicit evaluations are enabled by propositional representations acquired through top-down learning (De Houwer, 2006, 2014; Hughes, Barnes-Holmes, & De Houwer, 2011), or are generated through a variety of processes (see, e.g., Amodio, 2014; Amodio & Devine, 2006; Amodio & Ratner, 2011; Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Payne, 2001; Sherman et al., 2008), each of which may have different capacities for updating (see Amodio & Ratner, 2011). These views either would strongly predict such undoing of implicit evaluations (De Houwer, 2014; Hughes, Barnes-Holmes, & De Houwer, 2011) or are at least open to the possibility (Amodio & Ratner, 2011; Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005).

Another example is the Metacognitive Model (MCM) by Petty and colleagues (Petty & Briñol, 2010; Petty, Briñol, & DeMarree, 2007). In this model, new, countervailing information that overturns older information can indeed change implicit evaluations, but only when the new information is sufficiently elaborated on to *replace* previous associations. Specifically, previously held evaluations that one chooses to reject do not get immediately removed from memory, but instead get tagged as false. These validity tags are initially only weakly associated, which explains why it is difficult to instantly revise (see Petty et al., 2006). However, the model predicts that various factors may moderate the strength of these new "false" tags, such as the extent of elaboration, and the tags may eventually become sufficiently strong so as to prevent the activation of the original evaluation (Petty & Briñol, 2010; Petty et al., 2006).

Despite the theoretical support for implicit revision through the undoing of initial learning, the empirical evidence is mixed. Some studies have attempted to change implicit evaluations by presenting new information about the validity of older information. In such studies, participants first form an initial implicit (as well as explicit) evaluation toward a novel person or group, and then are told that the initial information was true or false (Boucher & Rydell, 2012; Peters & Gawronski, 2011). Asking people to simply "negate" or undo a prior impression in this way has typically been less effective in shifting implicit than explicit evaluations (Deutsch et al., 2006; Gawronski & Bodenhausen, 2011, p. 88; Gawronski et al., 2008; Gregg et al., 2006). The only way that such negation instructions lead to implicit revision is when they are presented nearly simultaneously with the initial information (Peters & Gawronski, 2011), and are sufficiently salient to elicit considerable attention (Boucher & Rydell, 2012).

These studies might be classified as "subtraction" studies because the new information requires people to "unlearn" the prior information, in effect subtracting its influence from their impressions. These might represent cases where we learn that our first impressions were actually based on false rumors, for instance. This kind of subtraction method of telling people to simply reject initial information as false may present several challenges though, which may explain why empirical attempts have yielded only mixed evidence. First, asking people to negate their initial impression may prompt them to try suppressing it, which might ironically keep the rejected thought active (Wegner, 1994). Second, people may be unable to erase all traces of implicit evaluations when instructed to do so, in line with memory work suggesting that intentional forgetting does not erase all traces of a memory (Bjork & Bjork, 2003). Even if participants could, in theory, respond to an instruction to negate everything they have previously learned by thoroughly reinterpreting those details, they may sometimes have low motivation or ability to do so, resulting in insufficiently deep processing of the negation (Boucher & Rydell, 2012; see also Petty et al., 2006). Finally, learning that initial information was false does not necessarily imply the opposite impression. Learning that someone *did not* perform a negative behavior does not mean the person enacted a positive one, and vice versa. Attempting to silence initial information by classifying it as false would seem to face multiple kinds of challenges, and these challenges might explain the mixed empirical record to date.

Joining Forces: Revision Through Subtraction *and* Addition

What happens when we learn new, counterattitudinal information that *also* overrides the initial impression? That is, what about trying to encompass both an addition as well as a subtraction approach to implicit revision? For example, learning that Schindler was actually a hero both adds new information as well as *changes* the meaning of the initial information. The fact that he employed Jewish workers in his factory, for example, now has a completely different meaning (and evaluative connotation). Learning new information that also forces a change in the meaning of older information would seem to possess the advantages of addition and subtraction approaches, whereas avoiding the pitfalls of using either by itself.

To our knowledge, there are only a few studies that provide a test along these lines. Gregg et al. (2006, Studies 3 and 4) attempted this approach in two of their studies. They informed participants that their impressions of two novel groups should be flipped: in one (Study 3), the experimenter had ostensibly made a mistake in informing them of which group was positive and which negative; in the other (Study 4), the groups were said to have changed in their moral character over time, the formerly negative one becoming positive, and the formerly positive one becoming negative. These seem like "subtraction + addition" methods in that the new information states that the evaluation attached to each of the groups should be reversed (e.g., the "Niffites" are no longer

bad, and are instead now good), but, in this case, they did not find any evidence of implicit revision. However, these particular instantiations of a subtraction + addition approach may not have been ideal. First, in Study 3, the new information did not change the *meaning* of the groups' initial behaviors so much as switch the authorship of those behaviors. The behaviors had the same evaluative connotation, but the mapping of behavior to group was supposed to be switched. Second, whether people are able to revise their impressions would seem to depend critically on the believability of such a switch. If people find the notion of a group completely switching its entire moral character unlikely (that is what Study 4 asked participants to believe occurred over time), then we might not see implicit revision even if they had been able to do so (especially given the stickiness of initial immorality; e.g., Knobe, 2006; Malle & Knobe, 1997; Reeder, Pryor, & Wojciszke, 1992). Finally, these concerns might have been especially pronounced because the participants in both studies knew that the groups were fictional and the scenarios hypothetical. In Study 3 for instance, they learn that for some other participants the goodness or badness of the groups is the reverse of what they were told, which might privately undermine their sense that either group is "truly" good or bad. Explicit evaluations, on the other hand, may have reflected participants' perceptions of the expectations of the experimenter.[2]

A better test of the subtraction + addition approach might be to present new countervailing information that truly changes the meaning of the old information, and to do so in a way that has some ecological validity (i.e., use a paradigm that maximizes believability). To our knowledge, there is only one study that has tested such an approach. Wyer (2010, Study 2) showed that implicit evaluations of a novel target were revised in light of new, counterattitudinal information that *changed* the interpretation of prior details—an apparent skinhead who behaved in an off-putting way turned out to be ill with cancer. This changed participants' implicit evaluation of the target, but only if they were able to revisit each one of his initial behaviors once they got the new information. Wyer suggested that this focused rehearsal may have sufficiently strengthened the "false" tags linked to those prior details to allow the implicit revision, in line with the MCM perspective (Petty et al., 2006). Although there remain many questions about why and how this effect emerged, it is intriguing, and raises the possibility that when people learn information that makes them *reassess* their prior knowledge, their implicit evaluations can be updated accordingly. In what follows, we consider this possibility.

Reinterpretation

Although existing theories are open to the possibility of undoing implicit evaluations, in line with some supportive findings, the mechanisms through which it might occur remain largely unknown. We propose that the ability of new information to *recast* the old information on which the initial evaluation was based—such as in the case of Oskar Schindler—is one mechanism of change that may be especially effective. In particular, when new information prompts a reinterpretation (i.e., a change in the evaluative meaning) of old information, we predict that implicit evaluations will be updated accordingly.

Reinterpretation of prior information may be uniquely positioned to produce strong revision of implicit evaluations. This strategy involves not just the invalidation of the initial impression (i.e., subtraction), but also the *replacement* of that impression with a countervailing other (i.e., addition), and it does this in one fell swoop. That is, it introduces an explanation for why previous learning should be reinterpreted, and revised in the opposite evaluative direction. This may often be more effective in producing revision of implicit evaluations than either addition or subtraction approaches alone. If there are reasons to suspect that both rejecting a prior impression and affirming a new one may have limitations when implemented separately (as we previously reviewed), the initial demonstration in Wyer (2010) is a promising sign that reinterpretation—a change in the meaning of earlier details such that an initial impression is both negated *and* replaced with another—may be an effective way to overturn implicit evaluations.

Though the demonstration in Wyer (2010) remains the best test in the literature of the possibility that reinterpretation may produce revision of implicit evaluations, there is much that remains unknown. Most critically, because there were no measures of the nature of the thinking done by participants or comparisons of the effectiveness of different types of counterattitudinal information, it is unclear whether there was any reinterpretation at all; participants given the new information may have simply elaborated on it as they revisited the initial information without changing their understanding of the initial information. In other words, there was no evidence about the process leading to the revision, and whether it involved any reassessment of the meaning of the initial behaviors. It may be that elaborating on *any* counterattitudinal new information, without a rejection of the earlier impression, would have been sufficient.

In the current work, we examine whether implicit evaluations can be fully and durably reversed when new information *changes the meaning of a previous impression*. This would demonstrate that implicit evaluations can be fully reversed after reasoning about a prior impression. In addition, we identify the process by which this kind of change occurs, and reveal its operating characteristics: it requires more than the simple pairing of the old attitude object with counterattitudinal information, and is not reducible to more extensive *general* thinking about that information; rather, it occurs through reinterpretation specifically.

Overview of Current Work

We developed a new paradigm tailored to this research topic. In each experiment, participants read a story, presented one sentence at a time, about an individual named Francis West who is described as breaking into and causing damage to two homes. Participants' implicit evaluations toward Francis West are then measured. Afterward, they read a final piece of information which either maintains the gist of what they already read (control condi-

[2] Note that a few studies similar to those reported by Gregg et al. (2006) did not technically measure implicit evaluations. Wilson and colleagues (2000) used a measure requiring rapid explicit judgments, and compared these to more conventional slower explicit judgments. Petty et al. (2006) employed a study design (in Experiment 2) asking participants to switch the targets of two sets of information, but the implicit measures tapped associations between the targets and confidence versus doubt, rather than evaluations. As such, very few studies have attempted "addition + subtraction" designs while measuring changes in implicit evaluations.

tions in which Francis remains negative) or dramatically reverses it (experimental condition in which Francis becomes positive) by offering a reinterpretation of what was previously learned: The houses were on fire, and Francis was searching for two young children who he knew were inside.

With this paradigm, we accomplish a number of goals across the studies. First, we use a paradigm in which participants are learning about an ostensibly real person through a cohesive narrative that is meant to be engaging and credible. In this way, we hope to maximize participants' motivation and attention while in the learning paradigm. Second, we demonstrate fast revision in the direction of negative to positive, which has been shown to be especially difficult in recent work (Cone & Ferguson, 2014) possibly because of negativity dominance (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Cacioppo, Gardner, & Berntson, 1997; Rozin & Royzman, 2001). Third, the paradigm leads participants to form implicit evaluations toward a novel person, which enables us to test the ways in which implicit evaluations can be *changed* (Fazio, 2007; Ferguson & Fukukura, 2012; Gregg et al., 2006). That is, we can ensure that we measure learning, rather than reactivation of previously learned material (e.g., see Gregg et al., 2006, p. 16). Presenting counterattitudinal information about familiar objects might simply activate prior learning, the characteristics of which (how long it took to learn it, etc.) would be unknowable.[3]

In Experiments 1a and 1b, we demonstrate that reversal occurs in this paradigm using two different implicit measures of evaluations. Experiment 2 shows that the relevance of the new information in recasting the old is essential to this reversal by comparing it with a condition containing equally positive but irrelevant information. Experiment 3 examines whether the revision occurs through an active thought process of reappraising the old information, by manipulating cognitive load to test whether a reduction in cognitive resources would undermine revision. Experiment 4 demonstrates that participants' self-reported belief that the new information changes the meaning of the prior story predicts the degree of revised implicit evaluations of Francis West, even when controlling for more general measures of the speed of thinking and extensiveness of thinking. In Experiment 5 we show that self-reported reinterpretation of the prior story mediates the effect of the new information on implicit evaluations, even when controlling for the extent to which participants reported thinking about the new information in general. Finally, in Experiment 6 we demonstrate the durability of the revised implicit evaluations over 3 days. For all studies, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study (Simmons, Nelson, & Simonsohn, 2012).

Experiment 1a

We presented participants with a story in which a novel person ("Francis West") is depicted breaking into and causing great damage to two houses on his street, followed by a revelation (in the experimental condition) that his behavior was motivated out of a desire to save children inside from a fire spreading through the houses. Thus, this revelation implies a reinterpretation of the prior story. In a control condition, participants instead read one additional piece of information that does not contradict, but rather is consistent with, the information learned before (see Appendix). In addition to measuring implicit evaluations, we assessed explicit evaluations and a variety of participants' reactions to the story, including story comprehension, confusion, and how deeply they reported thinking about the story's details.

Method

Participants. There were 200 workers on Amazon's Mechanical Turk Web site (www.mturk.com) who participated in the study for \$0.75 (55% male; M_{age} = 37 years, SD = 14). We selected this number a priori so as to collect ~100 per between-participants condition. Because this was our first attempt at testing the effect, we collected enough data to be able to detect a moderately sized effect.

Materials. To induce an initially negative implicit evaluation toward a novel target person, participants were led through a story detailing supposedly true events centering around an individual named Francis West. The story was presented in a linear piecemeal fashion, across 26 screens. The described events portray Francis West as a man who ransacks the homes of his neighbors, destroying their property (e.g., throwing a pot of water onto a laptop) and taking "precious things" from the bedrooms. After initial assessment of implicit evaluations toward Francis, participants were then presented with a single screen of additional information that varied by condition. In the control condition, the new information continued the thread of the story: Francis began to chuck rocks at the windows of the houses he had just pillaged. In the other condition (henceforth dubbed the "fire rescue" condition), participants instead read that Francis broke into the houses because he saw that they were on fire, and the only precious things he removed from the bedrooms were the young kids of the families.

As participants read each statement about Francis in both the initial and subsequent story periods, an image of the upper body of a White male labeled Francis West was displayed on the screen. Each participant was randomly assigned an image of one individual to serve as Francis West, out of a set of 11 such images drawn from previous research (Minear & Park, 2004). The men in all photographs had neutral expressions, and ranged in age from 20 to 33 years.

Implicit evaluations. Implicit evaluations were measured twice, once after reading the initial statements (Time 1) and once after reading the final information (Time 2). The Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) was administered at both instances (see the Procedure section). Each AMP consisted of 40 trials, with separate sets of Chinese characters used at Time 1 and Time 2. The order in which the sets were administered was counterbalanced across participants.[4] The image of Francis West

[3] Presenting participants with counterattitudinal information about a known social group might indeed result in changed implicit evaluations of the group. However, this does not mean that participants *learned something new*. Thus, the evidence for the context-dependence of implicit evaluations (e.g., Blair, Ma, & Lenton, 2001; Dasgupta & Greenwald, 2001; Ferguson & Wojnowicz, 2011; Wittenbrink, Judd, & Park, 2001; see Blair, 2002, for a review) does not address the topic of how easily implicit evaluations can reflect new information.

[4] The order in which the two sets of ideographs were administered will not be discussed further; order affected only one analysis of interest across all six studies. In Experiment 3, the interaction of time, prime person, and story condition was significantly moderated by ideograph order, such that the revision effect was stronger in one counterbalance condition than in the other (but the revision effects, and all simple effects of interest, were still significant in both). Ideograph set order had no similar effects in any other analysis.

that was assigned to the participant served as the prime on half of the trials, and each of the other 10 images of unknown individuals served as a prime on two of the other trials. Participants were instructed that they would sequentially view a set of Chinese ideographs, and that their task was to determine for each whether it was more or less pleasant than the average ideograph by pressing the *k* and *d* keys on their keyboard, respectively. On each of the 40 trials, participants were first presented with a prime photograph of Francis or an unknown individual for 75 ms, followed by a blank screen for 125 ms, an ideograph for 100 ms, and finally a pattern mask of black and white noise, which remained on the screen until the participant responded. Participants were told that though the images that precede the ideographs may sometimes be positive or negative, they were to prevent these images from affecting their ratings, and instead should evaluate the ideographs solely on their own merits. Previous research suggests that this measure taps evaluative reactions toward the primes that are misattributed to the relatively neutral targets; thus, providing a measure of spontaneously elicited, unintentional evaluations of the primes (Payne et al., 2013; Payne et al., 2005; cf. Bar-Anan & Nosek, 2012).

Explicit evaluations. Explicit evaluations toward Francis were measured at Time 2 using measures adapted from previous research (Rydell & McConnell, 2006; Rydell, McConnell, Mackie, & Strain, 2006). Participants indicated how likable Francis is from 1 (*very unlikable*) to 7 (*very likable*), and completed 7-point semantic differential scales in random order on the dimensions of bad-good, mean-pleasant, disagreeable-agreeable, uncaring-caring, and cruel-kind. These six items were reliable, $\alpha = .994$, and were combined into a single scale for the analyses.

Questionnaire measures. Participants completed three multiple choice questions asking questions about the story that have different probable answers in the control and fire rescue conditions: why Francis threw water around the house (e.g., to ruin items, to put out a fire), why the cat died (e.g., Francis stepped on it, smoke inhalation), and what he removed from the houses (e.g., jewelry, children). They then identified Francis in a lineup of the 11 faces presented in the study (1 Francis and 10 control), indicated their level of confusion about what happened in the story on a scale from 1 (*not confused at all*) to 7 (*completely confused*), the extent to which they thought about the Time 1 story elements after reading the new information at Time 2 on a scale from 1 (*not at all*) to 7 (*a lot*), and how hard it was to make sense of how the Time 2 information fit with the rest of the story on a scale from 1 (*not at all hard*) to 7 (*very hard*).

Procedure. After providing informed consent, participants were instructed to read each statement in the story depicting the initial events surrounding Francis West. They were told to pay careful attention to the details that unfolded, as they would be asked questions about their perspective on the events later in the study. Participants proceeded through the screens containing the descriptions at their own pace, but spent a minimum of 3 s on each screen.

After reading the initial story about Francis West, participants took the first AMP. Then, they were informed that they would read a final piece of information about the events described previously, and were shown either the control or fire rescue information. They were instructed to think about how this information relates to what they learned before, and were required to wait at least 15 s before advancing. They then completed the Time 2 AMP. Next, they indicated whether they knew Mandarin and/or Cantonese, completed the explicit evaluation scale, and answered the other questionnaire items. Finally, they completed demographic questions, a set of measures unrelated to the present study regarding political evaluations, and were debriefed and compensated.

Results

Data preparation. Following Payne et al. (2005), data from four participants were excluded for indicating familiarity with Mandarin or Cantonese (2% of cases) and 18 participants for using a single key on every trial of at least one AMP, indicating a disregard for the instructions (9% of cases). This left 178 cases for the final analysis.

Comprehension checks. Participants in both conditions showed good comprehension of the story details, with 93% in the control condition and 84% in the fire rescue condition answering the three interpretative questions in a manner fully consistent with the condition to which they had been assigned. This difference was marginally significant, $\chi^2 = 2.94$, $p = .086$. In addition, every participant correctly identified Francis West in the lineup of 11 photographs.[5]

Implicit evaluations toward Francis West. Implicit evaluations were assessed using a 2 (Measurement Time: Time 1 and Time 2) × 2 (Prime Person: Francis West and control faces) × 2 (Story Condition: control or fire rescue) mixed design, with the first two factors manipulated within-participants and the third manipulated between-participants. The proportion of trials in each cell of this design on which participants indicated that the Chinese ideograph was more pleasant than average served as the dependent variable in a repeated-measures analysis of variance (ANOVA).

Every effect in the model was significant, including our predicted three-way interaction between measurement time, prime person, and story condition, $F(1, 176) = 36.551$, $p < .001$, $\eta_p^2 = .172$. Decomposition of this interaction revealed that in the control story condition, in which Francis is depicted as negative at both Time 1 and Time 2, no interaction between time and prime person emerged, $F(1, 176) = 1.55$, $p = .214$. Instead, as predicted, there was solely a main effect of prime, such that implicit positivity toward Francis was lower ($M = .40$, $SD = .27$) than implicit positivity toward the neutral faces ($M = .63$, $SD = .19$), $F(1, 176) = 50.51$, $p < .001$, $\eta_p^2 = .223$. In the fire rescue condition, however, there was the predicted significant interaction between measurement time and prime person, $F(1, 176) = 51.90$, $p < .001$,

[5] For descriptive purposes, we continued to collect information on how frequently participants completed the inferential items in a manner fully consistent with their story condition, and whether they could correctly identify the face of Francis West out of a lineup, throughout all of the subsequent studies. From Experiment 1b onward, we also always included a simple manipulation check item asking participants to identify from a short list which final information had been presented to them about Francis West. We never made a priori predictions about these measures, and so do not discuss them further in this article, though the correlations between implicit liking and the inferential items in the fire rescue condition are available in Table 1. We chose in advance to include participants regardless of their performance on these measures. Across the samples of all of our experiments, after setting aside excluded participants (reasons reported in text), 98.73% were correct in identifying the final information they had been shown, 99.58% correctly identified the image of Francis West, and 82.70% responded to the inferential questions in a manner fully consistent with their story condition.

η_p^2 = .228. At Time 1, Francis was less positive (M = .40, SD = .30) than the neutral faces (M = .60, SD = .19), $F(1, 176)$ = 23.53, $p < .001$, η_p^2 = .118, whereas at Time 2, Francis was more positive (M = .72, SD = .22) than the neutral faces (M = .54, SD = .23), $F(1, 176)$ = 17.42, $p < .001$, η_p^2 = .09. Viewed another way, implicit positivity toward Francis did not show a shift from Time 1 to Time 2 in the control story condition, $F(1, 176)$ = 1.43, p = .234, but showed a significant increase from Time 1 to Time 2 in the fire rescue condition, $F(1, 176)$ = 84.58, $p < .001$, η_p^2 = .325 (see Figure 1).

Explicit evaluations toward Francis West. Explicit evaluations were measured after the second AMP, and differed significantly by story condition, unequal variances $t(103.36)$ = 34.77, $p < .001$, Cohen's d = 6.84. Participants had more positive explicit evaluations toward Francis West in the fire rescue (M = 5.99, SD = 1.21) versus control condition (M = 1.20, SD = .41).

Finally, rather than reporting the correlations between implicit and explicit evaluations (as well as their relations to other measures) in the main text for each study, we summarize this information in Table 1.

Story condition effects on other questionnaire measures. Participants reported significantly more confusion with the story in the control (M = 3.64, SD = 2.05) versus fire rescue condition (M = 1.93, SD = 1.57), unequal variances $t(169.50)$ = 6.27, $p < .001$, Cohen's d = .96. This pattern was the same for reported difficulty making sense of the final story information, unequal variances $t(175.96)$ = 2.95, p = .004, Cohen's d = .44. This result makes sense, given the lack of resolution available to participants in the control condition regarding the motivation for the destruction perpetrated by Francis West. Also expected, there was a strong story condition effect on the self-reported extent to which participants thought about the overall story upon reading the final information, with those in the fire rescue condition reporting more thinking (M = 6.26, SD = 1.20) than those in the control condition (M = 5.12, SD = 1.72), unequal variances $t(163.02)$ = 5.14, $p < .002$, Cohen's d = .80. However, none of the above questionnaire items moderated or mediated implicit evaluation revision, and the three-way interaction between time, target person, and story condition remained even when controlling for any of these (mean centered; Aiken & West, 1991).

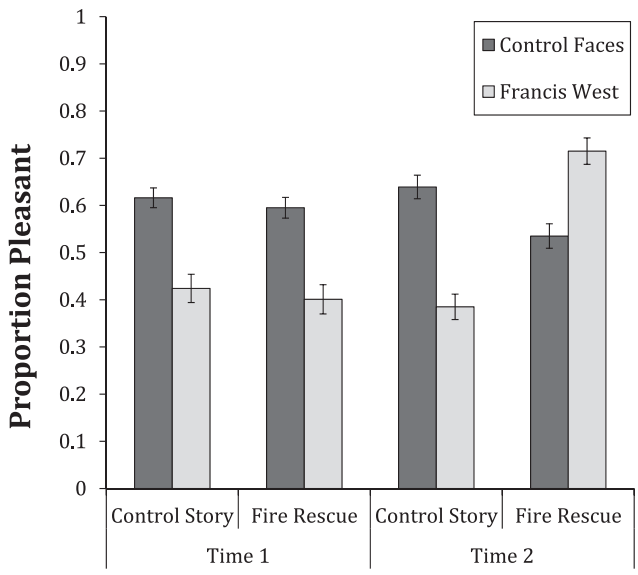

Figure 1. Mean proportion of ideographs judged more pleasant than average in Experiment 1a, by face prime, time, and story condition. Error bars are *SE*s.

Discussion

Experiment 1a shows that participants strongly revised their implicit evaluations toward a novel target person, once given a reinterpretation of the original information. Whereas those in the control condition showed a persistence of their initial, negative evaluations, those in the fire rescue condition switched to significant implicit positivity toward Francis West after reading the revelation. Experiment 2 will begin to examine mechanism, but first Experiment 1b replicates the basic pattern with a different implicit evaluation measure.

Experiment 1b

In Experiment 1b, we sought to replicate the revision effect using the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). In addition, to induce an even stronger initial negative evaluation (and hint at a possible motive in the control condition), we added a new detail to suggest in the first part of the story that Francis' actions were a hate crime against the town's first interracial families in the neighborhood. We anticipated that the revelation that he actually was saving the children from a fire would effectively counter this suspicion and produce reversal just as it had in Experiment 1a. As a final addition, we measured explicit evaluations after both IATs. This allows us to show how implicit liking of Francis West is indeed changing in a similar way to explicit liking.

Method

Participants. There were 301 participants recruited from Amazon's Mechanical Turk Web site (www.mturk.com) who were paid $1.00 to participate in this study (55.6% male; M_{age} = 32.0 years, SD = 11.8). We sought to collect data from ~150 participants for each of the two between-participants story conditions, and so intended to collect data from 300 participants. The number is more than in Experiment 1a because we thought that the IAT, which relies on response times, might produce noisier data than the AMP used in Experiment 1a. One additional participant completed the study without subsequently submitting it for compensation.

Materials. The story that participants read at Time 1 was identical to that from Experiment 1a, except for the addition of a new event at the very start: "Francis' small town was about 99% White, but recently the Griffins and Wards, the town's first ever interracial families, had moved in." By providing the hint of a motive for Francis' actions, we thought that this might reduce the persistent difference in expressed confusion between the two conditions, as well as provide even more initial negativity toward Francis that must later be overcome.

IAT. The IAT included the same 11 faces used in the AMP in Experiment 1a, with one of the 11 having been randomly assigned to be used as Francis West for each participant. Positive adjectives

Table 1
Correlations Between Implicit and Explicit Evaluations and Questionnaire Measures in Fire Rescue Condition at Time 2, Experiments 1–6

	Study 1a	Study 1b	Study 2	Study 3	Study 4	Study 5	Study 6 Time 2	Study 6 Time 3
Implicit evaluations								
Perfect comprehension	.12	.10	−.07	.27**	−.03	.08	−.06	.14
Confusion	−.04	−.12	.06	.02	−.03	−.24**	−.03	−.14
Difficulty making sense	.04		−.02					
Extent of thinking on story	−.03		.18†					
Extent of thinking (new info)				.19*				
Extent of thinking (old info)				.02				
Subjective meaning change					.31*	.23**		
Rapid vs. gradual thinking					−.10			
Extensiveness of thinking					.01	−.09		
Positive mood							.28**	.15
Belief that story is real							.09	.14
Explicit evaluations								
Perfect comprehension	.32**	.48***	.54***	.52***	.40***	.30***	.19†	.22*
Confusion	−.33**	−.58***	−.55***	−.45***	−.41***	−.45***	−.26*	−.23*
Difficulty making sense	−.31**		−.65***					
Extent of thinking on story	.26*		.18†					
Extent of thinking (new info)				.32***				
Extent of thinking (old info)				.05				
Subjective meaning change					.71***	.58***		
Rapid vs. gradual thinking					−.40***			
Extensiveness of thinking					−.26*	−.10		
Positive mood							.50***	.42***
Belief that story is real							.19†	.16
Implicit evaluations	.30**	.16†	.14	.32***	.24†	.48***	.27**	.35**

Note. Cell values are Pearson correlations. In Experiments 1a and 2–6, the correlations involving implicit evaluations are partial correlations with the proportion of ideographs on Francis trials judged more pleasant than average, controlling for the proportion of ideographs on Control trials judged more pleasant than average. The exception is the IAT used in Experiment 1b, which is a relative measure of positivity toward Francis (vs. control faces); thus, the *D* scores were used, without any covariates. On all implicit and explicit liking measures, higher scores indicate more positive evaluations toward Francis West.
† $p < .1$. * $p < .05$. ** $p < .01$. *** $p < .001$.

presented during the IAT included *wonderful*, *excellent*, *good*, *great*, *appealing*, *outstanding*, *lovely*, *fantastic*, *beautiful*, and *amazing*. Negative adjectives included *horrible*, *terrible*, *awful*, *disgusting*, *offensive*, *hideous*, *revolting*, *bad*, *dreadful*, and *nasty*.

During the critical blocks, participants quickly sorted stimuli into one of four categories: *Francis West*, *Other People*, *Good Words*, and *Bad Words*. At any one time, two of these category labels (one person category and one adjective category) were displayed on the left side of the screen, and the other two were presented on the right side. A single stimulus from one of the four categories appeared on each trial. Half of the person trials consisted of an image of Francis West, whereas the other half displayed randomly chosen images from the other 10 images (control faces). Likewise, on half of the adjective trials a positive adjective was randomly selected from the list, whereas on the other half a negative one was. Each stimulus was displayed until the participant registered a response by pressing one of the two keys. If the response was correct, the next trial began; if it was incorrect, a red "X" appeared on the screen until the participant gave the correct response. The intertrial interval was 250 ms. The IAT consisted of seven blocks, with 20 trials in practice Blocks 1, 2, 3, and 6, and 40 trials in test Blocks 4 and 7 as well as transition practice Block 5 (for details, see Greenwald, Nosek, & Banaji, 2003). The order of the Francis West + Bad and Francis West + Good blocks was counterbalanced across participants.

Procedure. Participants first completed the story procedure in the same fashion as in Experiment 1a. Then, they took the first IAT and the same explicit evaluation items from Experiment 1a. Participants were next presented with the Time 2 story information in the same manner as before, either the fire rescue or control information. Immediately thereafter they completed the second IAT, the explicit evaluation scale, the story comprehension items, the photo identification, the confusion items, a new multiple-choice manipulation check question asking them to directly identify the final story information that they had been presented with, and demographic questions. They were then debriefed, thanked, and compensated for their time.

Results

Data preparation. We calculated implicit positivity toward Francis West for both IATs according to the *D1* scoring algorithm (Greenwald, Nosek, & Banaji, 2003). The differences between blocks were computed so that higher scores meant faster responding during the "Francis West + Good" pairing. On this measure, more positive scores are taken to suggest more implicit positivity

toward Francis West, and more negative scores are taken to suggest more implicit negativity toward him, relative to the control faces. Because of server error, IAT data were not recorded for eight participants. Eleven participants were excluded for responding faster than 300 ms on over 10% of trials, following scoring recommendations (Greenwald et al., 2003), leaving the final sample with 282 cases (54.6% male; M_{age} = 32.2 years, SD = 11.8).

Implicit evaluations toward Francis West. We analyzed implicit positivity toward Francis West by submitting D scores to a 2 (Measurement Time: Time 1 and Time 2) × 2 (Story Condition: control or fire rescue) mixed ANOVA, with the first factor manipulated within-participants. This analysis revealed a main effect of time, qualified by the time by story condition interaction, $F(1, 280) = 10.65$, $p = .001$, $\eta_p^2 = .037$. Simple effects tests demonstrated that, as expected, implicit evaluations toward Francis did not vary by story condition at Time 1, $F(1, 280) = 1.11$, $p = .292$, but were significantly negative ($M = -.066$, $SD = .394$) across the sample, one-sample $t(281) = -2.82$, $p = .005$, Cohen's $d = .336$. At Time 2, however, implicit positivity of Francis was higher in the fire rescue condition ($M = .197$, $SD = .34$) than the control story condition ($M = -.021$, $SD = .33$), $F(1, 280) = 45.44$, $p < .001$, $\eta_p^2 = .140$. Although implicit evaluations toward Francis were significantly greater than zero in the fire rescue condition at Time 2, one sample $t(147) = 7.04$, $p < .001$, $d = 1.16$, they were no longer significantly below zero in the control condition, one sample $t(133) = -.72$, $p = .474$. Thus, the significant change in positivity across time in the fire rescue condition paralleled the shift observed in Experiment 1a, $F(1, 280) = 45.44$, $p < .001$, $\eta_p^2 = .140$, whereas a marginal shift in the positive direction also occurred in the control condition that was not present in the prior study, $F(1, 280) = 3.64$, $p = .057$, $\eta_p^2 = .013$ (see Figure 2 for the mean D scores in each condition).

Auxiliary analysis. The shift toward neutral implicit evaluation of Francis West from Time 1 to Time 2 in the control condition might suggest that a portion of the sample is relatively unconvinced of Francis West's badness at Time 1 and, after being similarly unconvinced at Time 2, lose whatever tenuous negativity toward him they might have possessed at Time 1, producing the condition's overall shift to neutral at Time 2. To show reason-based revision, we need to demonstrate that for those individuals that *are* induced with an initial evaluation, new information can quickly and significantly reverse it. Otherwise, a finding that the group as a whole exhibits reversal at Time 2 could be a product of shifts in these unconvinced participants, rather than shifts in those who acquired the initial evaluation. As such, we undertook a more conservative auxiliary test of the revision hypothesis by examining the means of Time 2 IAT scores solely among those participants who displayed initial negativity at Time 1 (D scores less than zero). In this subsample, IAT scores at Time 2 differed between story conditions, $t(159) = 5.73$, $p < .001$, Cohen's $d = .91$, such that scores were still significantly below zero in the control story group ($M = -.14$, $SD = .31$), one sample $t(76) = -3.80$, $p < .001$, and were now significantly above zero in the fire rescue group ($M = .16$, $SD = .34$), one sample $t(83) = 4.32$, $p < .001$.[6] Although like any measure D scores are not a process-pure measure of underlying evaluations (Conrey et al., 2005) and their absolute values and even rank order are subject to extraneous variation, this Time 2 pattern of only those individuals below zero on the IAT distribution at Time 1 corroborates our prediction and demonstrates continued deviation from zero in the control story. This even more conservative test of our revision hypothesis builds support for the theory and corroborates the pattern from Experiment 1a (the results reported in Experiment 1a hold as well when submitted to this same analysis).

Explicit evaluations toward Francis West. Explicit evaluations toward Francis West were measured at Time 1 and Time 2. Analysis using a 2 (Measurement Time: Time 1 and Time 2) × 2 (Story Condition: control or fire rescue) mixed ANOVA revealed that a main effect of measurement time was qualified by the expected Time × Story condition interaction, $F(1, 280) = 1030.93$, $p < .001$, $\eta_p^2 = .786$. Simple effects tests revealed that explicit liking of Francis West increased from Time 1 ($M = 1.30$, $SD = .71$) to Time 2 ($M = 5.93$, $SD = 1.47$) in the fire rescue condition, $F(1, 280) = 2129.35$, $p < .001$, $\eta_p^2 = .884$. No such change from Time 1 ($M = 1.17$, $SD = .46$) to Time 2 ($M = 1.12$, $SD = .43$) took place in the control group, $F(1, 280) = .17$, $p = .680$.

Story condition effects on other questionnaire measures. As in Experiment 1a, confusion with the story was higher in the control condition ($M = 2.79$, $SD = 1.67$) than in the fire rescue condition ($M = 2.19$, $SD = 1.62$), $t(281) = 3.07$, $p = .002$, Cohen's $d = .365$. The addition of the hate-crime element to the story did not appear to stem this trend of higher confusion in the control condition. But once more, controlling for (centered) confusion did not reduce the Time × Condition interaction to nonsignificance, $p = .007$.

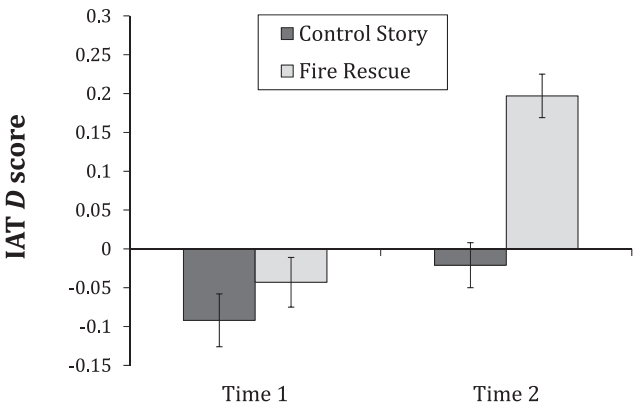

Figure 2. Mean D scores in Experiment 1b, by measurement time and story condition. Error bars are *SE*s.

Discussion

We replicated reason-based revision using a different implicit measure. Whereas participants' negative implicit evaluations to-

[6] This analysis differentially includes participants who completed the compatible-first order (Francis + bad) over the incompatible-first order (Francis + good), as the compatible-first order tends to produce more negative scores. However, we find the same results when splitting by order: D scores become positive in the fire rescue condition in both orders, $M = .16$, $SD = .33$, $t(43) = 3.21$, $p = .003$, and $M = .16$, $SD = .35$, $t(39) = 2.86$, $p = .007$, whereas D scores are still negative in the control story condition in both orders, $M = -.12$, $SD = .31$, $t(50) = -2.78$, $p = .008$, and $M = -.16$, $SD = .32$, $t(25) = -2.60$, $p = .015$, respectively.

ward Francis West were not qualified by time in the control condition, those in the fire rescue condition showed a significant reversal from negative to positive. Those positive evaluations in the fire rescue condition after receiving the final information were also significantly more positive than the neutral point; however, the final negative evaluations in the control condition did not differ from neutral overall, departing from the results in Experiment 1a with the AMP. A follow-up analysis, however, showed that among those participants most critical for the demonstration of revision—those who *did* show an initial negative implicit evaluation—negative evaluations persisted in the control group and were reversed in the fire rescue group.

As of yet, our results do not yet address the *how* or *why* of implicit evaluation revision. Our argument is that it is the relevance of the fire rescue information to the initial story that prompts revision through reinterpretation. The story presented in the Francis West paradigm appears to fit this bill, but the work to demonstrate the operative mechanisms has yet to be done. In fact, the failure of self-reported extent of thinking to moderate revision in Experiment 1a could be seen as an initial point against this idea, if reinterpretation of prior story events requires any amount of deliberation about the story, as we suspect that it might. However, this single questionnaire item may not adequately tap the degree to which reinterpretation took place, either because introspective access to this mechanism is weak (Nisbett & Wilson, 1977), or the reinterpretation that is required is easy to execute in this particular case. As such, we examine the conditions under which reason-based implicit evaluation change occurs in Experiment 2.

Experiment 2

To identify whether reinterpretation of the initial information is critical for revision, we compared evaluation change in the fire rescue condition with another condition in which extremely positive information about Francis is presented, but does not prompt a reinterpretation of Francis' initial, negative actions. To the extent that these conditions differ, our account that a change in the meaning of the initial information plays a crucial role in revision is supported.

Method

Participants. There were 299 participants recruited on Amazon's Mechanical Turk Web site (www.mturk.com) for this study in return for $0.75 in compensation (50.3% male; M_{age} = 33.9 years, SD = 11.7). We intended to collect data from 300 participants, but a technical error prevented one participant from completing the study, and that person was still compensated; thus, the study received data from only 299 individuals. As in Experiment 1a, this number of participants allowed us to fill each of our between-participants story conditions with data from ~100 people.

Materials. The story was identical to that used in Experiment 1a, except with an additional between-participants story condition added that served as a positive control condition. This was meant to present participants with a piece of information about Francis West at Time 2 that would be equally as positive as the fire rescue, but not providing an explanation for his initial, seemingly negative behaviors. Thus, the positive control condition would associate strong positive information with Francis just as the fire rescue condition does, while not justifying the recasting of the prior negative information. This comparison could illuminate whether the effect observed in Experiments 1a and 1b is because of the addition of an extremely positive piece of information to the participant's corpus of knowledge about Francis West to such a degree that it "swamps out" the previously learned negative information without any revision of the prior associative expressions per se, or as we predict, depends on a reinterpretation of the prior information.

In the positive control condition at Time 2 participants read the following statement, pretested in a separate sample to be equally as positive to the action of saving two children from a raging fire[7]: "At a different point in time, Francis West was in the news because he was at a subway station when he noticed that a baby had crawled and fallen onto the tracks below. Seeing a rapidly approaching train, Francis jumped down onto the tracks, grabbed the baby, and climbed up to safety a split-second before the train came roaring past." Besides this difference, the new positive control condition proceeded in an identical fashion to the other two.

Procedure. The procedure was the same as in Experiment 1a, except for the addition of the positive control condition, and the manipulation check asking participants to identify the final information they read about Francis, with three answer choices reflecting the Time 2 information presented in the three story conditions. Besides these changes, participants viewed the same stories and took the same AMPs and questionnaire measures as used in Experiment 1a. At the end of the study, participants completed measures for an unrelated investigation.

Results

Data preparation. In line with Payne et al. (2005), 14 participants familiar with Mandarin or Cantonese (4.7% of cases), and 17 participants who used a single key on every trial of at least one of the two AMPs (5.7% of cases), were excluded, leaving 268 cases for analysis.

Implicit evaluations toward Francis West. We assessed implicit evaluations toward Francis West by analyzing judgments of ideographs in a 2 (Measurement Time: Time 1 and Time 2) × 2 (Prime Person: Francis West and neutral) × 3 (Story Condition: control, fire rescue, or subway rescue) mixed design, with the first two factors manipulated within-participants and the third manipulated between-participants.

Every effect in the design was statistically significant, but of most interest, the three-way interaction between time, prime person, and story condition obtained, $F(2, 265) = 20.004$, $p < .001$, $\eta_p^2 = .131$. There was once again no interaction between time and prime person in the control condition, $F(1, 265) = .16$, $p = .692$, with only a main effect of prime person, $F(1, 265) = 57.17$, $p <$

[7] Fifty Mturk workers read eight heroic actions that a hypothetical individual might do (e.g., "Waded into the water above Niagara Falls to save two stranded kids") along with a description similar to the fire rescue condition ("Ran into two burning homes to save two children from a fire"). These nine actions were presented in random order on a single screen, and participants evaluated how positively they would view a person who did the action on a 1-100 scale (*not positive at all* to *as positive as possible*). The fire rescue behavior (M = 92.4, SD = 10.5) and the subway rescue scenario (M = 91.7, SD = 16.1) did not differ, $t(49) = .54$, $p = .595$.

.001, $\eta_p^2 = .177$, such that Francis West was evaluated less positively ($M = .40$, $SD = .26$) than neutral faces ($M = .64$, $SD = .20$).

Once more, in the fire rescue condition there was a significant interaction between time and prime person, $F(1, 265) = 76.87$, $p < .001$, $\eta_p^2 = .225$. At Time 1, Francis was significantly less implicitly positive ($M = .39$, $SD = .27$) than the neutral faces ($M = .64$, $SD = .19$), $F(1, 265) = 41.43$, $p < .001$, $\eta_p^2 = .135$. However, at Time 2, Francis was significantly more positive ($M = .67$, $SD = .23$) than the neutral faces ($M = .53$, $SD = .21$), $F(1, 265) = 14.78$, $p < .001$, $\eta_p^2 = .053$.

In the subway condition, there was also a significant interaction between time and prime face, $F(1, 265) = 12.65$, $p < .001$, $\eta_p^2 = .046$. At Time 1, Francis was significantly less implicitly positive ($M = .36$, $SD = .27$) than the neutral faces ($M = .64$, $SD = .21$), $F(1, 265) = 54.83$, $p < .001$, $\eta_p^2 = .171$. At Time 2, Francis was still significantly less positive ($M = .50$, $SD = .30$) than the neutral faces ($M = .62$, $SD = .24$), $F(1, 265) = 11.45$, $p = .001$, $\eta_p^2 = .041$. The increase in implicit positivity of Francis West from Time 1 to Time 2 was significant, $F(1, 265) = 17.74$, $p < .001$, $\eta_p^2 = .063$. Thus, though the subway rescue condition significantly attenuated the implicit negativity of Francis relative to neutral faces, only the fire rescue condition evidenced a significant revision of a negative evaluation into a positive one. Figure 3 displays the mean proportion of ideographs judged more pleasant than average at Times 1 and 2 for both Francis and the neutral faces within each story condition.

Explicit evaluations toward Francis West. The effect of story condition on explicit evaluations toward Francis West at Time 2 was assessed using a one-way ANOVA (control, fire rescue, or subway rescue). As before, explicit evaluations toward Francis West at Time 2 varied between conditions, $F(2, 265) = 430.42$, $p < .001$, $\eta_p^2 = .765$. Participants explicitly liked Francis the most in the fire rescue condition ($M = 5.86$, $SD = 1.35$), followed by the subway rescue condition ($M = 2.45$, $SD = 1.18$), and finally the control condition ($M = 1.21$, $SD = 0.60$). Each of the simple comparisons between conditions was significant, all ps $< .001$. In addition, one-sample t tests revealed that each group mean significantly diverged from the midpoint, all ps $< .001$; thus, as with implicit evaluations, the subway rescue information did not make Francis West explicitly positive, though he was seen as less negative than in the control condition.

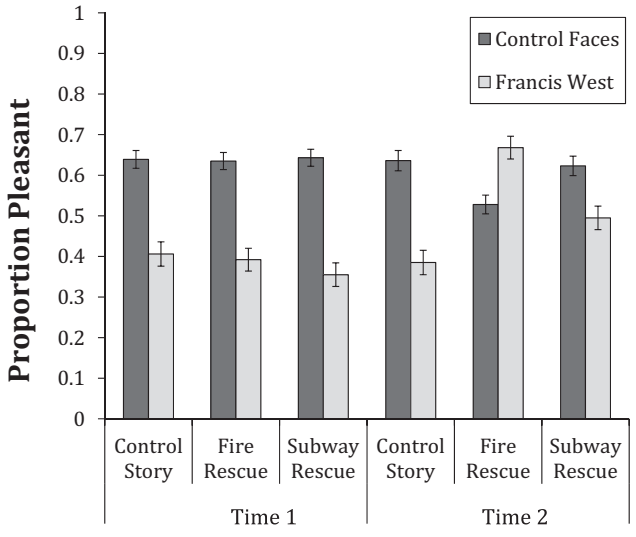

Figure 3. Mean proportion of ideographs judged more pleasant than average in Experiment 2, by measurement time, story condition, and face prime. Error bars are *SE*s.

Story condition effects on other questionnaire measures. Subjective confusion with the story was lowest in the fire rescue condition ($M = 2.01$, $SD = 1.58$), higher in the control condition ($M = 3.15$, $SD = 1.95$), and highest in the subway rescue condition ($M = 3.62$, $SD = 1.96$), $F(2, 265) = 18.85$, $p < .001$, $\eta_p^2 = .125$. Including subjective confusion in the analysis of implicit evaluations did not produce any interactions or change the significance of any effects.

As in Experiment 1a, there was also a story condition effect on the reported extent to which participants thought about the new information that they were presented with at Time 2, $F(2, 265) = 3.03$, $p = .05$, $\eta_p^2 = .022$. Participants thought the most about the information in the fire rescue condition ($M = 5.96$, $SD = 1.40$), and less in both the control condition ($M = 5.56$, $SD = 1.34$), and the subway rescue condition ($M = 5.49$, $SD = 1.42$). Extent of thinking was greater in the fire rescue condition relative to the others, $F(1, 265) = 5.93$, $p = .016$, $\eta_p^2 = .022$, but the subway rescue and control conditions did not differ, $F(1, 265) = .10$, $p = .758$. This suggests that participants thought most about the information in the fire rescue condition.

Moderation by extent of thinking. Although the revision effect was not moderated by the self-reported "extent of thinking about the story" measure in Experiment 1a, when we added this as a covariate in the implicit evaluations analysis in the present experiment, the four-way interaction (between time, prime person, story condition, and extent of thinking) was marginal, $F(2, 262) = 2.74$, $p = .066$, $\eta_p^2 = .021$. Examining the story conditions separately revealed that the Time × Prime person × Extent of thought interaction was not significant in the control or subway rescue conditions, both ps $> .5$, but was in the fire rescue condition, $F(1, 93) = 6.14$, $p = .015$, $\eta_p^2 = .062$. The effect was such that the Time × Prime person interaction (the revision effect) was stronger at high levels of thinking about the story (+1 *SD*) ($F(1, 93) = 55.00$, $p < .001$, $\eta_p^2 = .372$) than at low levels (−1 *SD*) ($F(1, 93) = 12.03$, $p = .001$, $\eta_p^2 = .115$, though as the numbers show, the revision effect was still significant with less reported thinking.

Discussion

Building on the initial studies, the control positive condition appeared markedly different from the fire rescue condition. Whereas reading about the fire rescue produced significant reversal, reading about the subway story reduced, but did not eliminate, participants' initial, negative impression. These results are consistent with our account that the ability of new information to prompt reinterpretation of prior information is key to full revision in this paradigm. The reduction, but not reversal, of negative implicit evaluations in the subway condition is consistent with the idea that the subway information was simply added to, but did not reverse, the initially learned negative information. In this sense the subway control condition bears some similarity to learning in the Bob paradigm, in which counterattitudinal learning proceeds by presenting participants with unconnected statements about Bob that are opposite in valence, but do not contradict the previously learned information in any other way; in this

paradigm, implicit evaluation revision generally proceeds slowly (e.g., Rydell & McConnell, 2006; Rydell et al., 2007) unless the initial impression is positive, and the new behavior is extreme and negative (Cone & Ferguson, 2014). Although Cone and Ferguson (2014) found that a single, sufficiently extreme negative behavior was enough to significantly alter initial positive implicit evaluations toward a novel target, they too found that a single, extreme positive behavior was *not* enough to overturn an initial negative implicit impression. Together with the findings of the current study, this suggests that revision might be especially effective if new information is added in such a way that also changes the meaning of the original information.

These findings also speak to some theoretical assumptions about how propositional information might modify implicit evaluations. The APE model (Gawronski & Bodenhausen, 2006) assumes that the affirmation of new, counterattitudinal propositional information can create a new counterattitudinal association, which could then affect implicit evaluations. However, here, participants in both the fire and subway conditions presumably affirmed (i.e., believed, processed) the new information, but reversal only happened in the former and not the latter case. This finding illustrates the (theoretical) importance of identifying *when* new propositional learning can modify implicit evaluations.

We still, however, do not know much about the process of reinterpretation. Are participants effortfully reinterpreting the previous information? Though self-reported extent of thinking marginally moderated the overall revision effect, the phrasing of this single question ("When you read that final piece of information about Francis West, to what extent did you think about the details you read earlier in the study?") made it specific to how much the participants felt that they explicitly revisited the previous details. As such, it likely did not adequately tap other forms of thinking, such as extent of thinking about the *new* information or rapid comprehension that did not require deliberate revisiting of the old information. The item also does not indicate what participants did with the old information when revisiting it (reinterpret, rehearse, reject, etc.). Therefore, we take this marginal effect as only suggestive evidence that some type of active thinking is involved in revision in these studies, and in the following studies we examine which aspects of active thinking are most important in producing revision.

In our next study, we tested whether some minimal degree of effortful processing is required to produce revision of implicit evaluations. To do so, we examine if this revision requires effort, and therefore, will be reduced when one is under high cognitive load.

Experiment 3

Method

Participants. There were 451 individuals who were recruited from Amazon's Mechanical Turk Web site (www.mturk.com) to participate in exchange for $1.00 (47.9% male; M_{age} = 34.86, SD = 11.8). A priori, we wanted to collect data from 450 participants so as to fill each of six between-participants conditions with 75 participants each. Given the results of our previous experiments, we determined via power analysis that a full 100 per cell was not necessary to achieve our effects, and reduced this amount to 75 for reasons of cost (while remaining above 90% power). Data from an additional participant were collected because one person completed the study without submitting the request for payment on Mturk.

Materials. The story used in Experiment 1a was used, and the stimuli were identical to those used in the previous studies. From the questionnaire, we dropped the item gauging difficulty making sense of the story after learning the final information (because of redundancy with the general confusion question) and the item asking the extent to which participants thought about the prior story details when learning the final information. That item had seemed to affect the revision results in one of the two studies in which it had been included (marginal interaction in Experiment 2) but not the other (Experiment 1a). We suspected that given the unburdened ability of all participants to consider the final information for as long as they wished, this extent of thinking measure might not effectively tap into meaningful variation in effortful thinking. To get more specifically at the type of thinking that might matter to the revision effect, we used the following two items, solely in the low- and high-cognitive load conditions: One asking participants how much they went back to think about the *new* story details after they no longer had to remember the number, and a second item asking the same about the *old* story details, both on a scale from 1 (*not at all*) to 7 (*very much*). We did not ask participants to self-report their extent of thinking in the no-load conditions, as we chose to focus here on which type of information participants would selectively choose to mentally revisit once given the opportunity to after the relieving of cognitive load.

Procedure. All participants completed the first story session and the first AMP, and were then told that they would be reading one final piece of information about Francis West. Participants in the no load condition were then presented with this information and moved on to the second AMP as was done in Experiments 1 and 2. In the two cognitive load conditions, however, participants were informed that they would do an additional task while considering this information: they would need to maintain a number in memory, to be reported immediately after moving on from the new story information. Such a cognitive load induction technique has been used successfully in prior research (see, e.g., Gilbert & Osborne, 1989). When they understood these directions and were ready to proceed, participants were presented with either a random two-digit number (low load) or random eight-digit number (high load) for 20 s. Then, the page automatically advanced to the final information about Francis West (either fire rescue or control). When participants were ready to advance, they were then presented with a textbox in which they had up to 15 s to enter the number they were previously presented with. Following this, the page automatically advanced to the second AMP. All participants then completed the explicit questionnaire items, and finally an unrelated experiment.

Results

Data preparation. Following the same procedure as our previous studies, six participants familiar with Mandarin or Cantonese were excluded from all analyses (1.3% of cases), as were 34 additional participants who used only one of the two response keys on all of the trials of at least one of the two AMPs (7.5% of cases), following established procedure (Payne et al., 2005). This left 411 cases for analysis.

Memory for the number in the low and high load conditions. We coded the numbers that participants recalled in the low and high load conditions for errors (omitted or extra digits, digits out of order). Predictably, perfect recall of the number occurred more frequently in the low load (99.25%) than high load condition (80.42%), $\chi^2(1) = 26.18$, $p < .001$. To ensure that all data from the low and high load conditions are drawn from participants who had engaged in the cognitive load task (and thus, experienced the manipulation as intended), we included in all analyses only those participants who perfectly recalled the number (Gilbert & Osborne, 1989). Getting the number correct had no effect on comprehension score, $F(1, 407) = .072$, $p = .788$, nor did its interaction with story condition, $F(1, 407) = .064$, $p = .800$. Where instructive, we also report findings for those participants who failed to correctly recall the number assigned to them.

Implicit evaluations toward Francis West. We submitted the proportion of ideographs judged as more pleasant than average to a 2 (Measurement Time: Time 1 and Time 2) × 2 (Prime Person: Francis West and Neutral) × 2 (Story Condition: Control or Fire Rescue) × 3 (Cognitive Load: No Load, Low Load, or High Load) mixed ANOVA, with the first two factors manipulated within-participants and the latter two manipulated between-participants.

Replicating the previous studies, the three-way interaction between measurement time, person, and story condition was significant, $F(1, 376) = 36.55$, $p < .001$, $\eta_p^2 = .089$. However, this effect was significantly moderated by cognitive load, $F(1, 376) = 3.43$, $p = .034$, $\eta_p^2 = .018$. Planned follow up analyses indicated that the contrast of high load versus the other conditions (no load and low load) moderated the revision effect, $F(1, 376) = 6.39$, $p = .012$, $\eta_p^2 = .017$, while the comparison of no load versus low load did not, $F(1, 376) = .46$, $p = .498$. In the high load condition, there was no significant interaction between time, person, and story, $F(1, 376) = 1.70$, $p = .193$, whereas in the other two conditions there was, $F(1, 376) = 43.75$, $p < .001$, $\eta_p^2 = .104$. As a result, in the high load condition, Francis West did not exceed the neutral faces in positivity at Time 2 in the fire rescue condition, $F(1, 376) = .245$, $p = .621$, while he did in the other two load conditions combined, $F(1, 376) = 25.102$, $p < .001$, $\eta_p^2 = .063$. Figure 4 displays the mean proportion of ideographs judged pleasant for both Francis West and the neutral faces within each story and cognitive load condition at Time 2.[8] The interaction between the revision effect and the high load versus other conditions contrast remains significant even if including only those participants who had perfect comprehension, correctly identified Francis in the photo lineup, and/or correctly identified the final story detail they had been presented with, all ps $< .05$.

Although the exclusion of those not perfectly remembering the number during the cognitive load task reduces the sample size by 28 individuals in the high load condition to 115 (80.42%), this is not likely to be responsible for the lack of a Time × Person × Story effect in the high load condition. This is because when we conducted our revision analysis on solely those 28 participants in the high load condition who failed to recall the number, we found significant revision among this sample, evidenced by a significant interaction between time, person, and story, $F(1, 26) = 6.78$, $p = .015$, $\eta_p^2 = .207$. If the 28 participants who failed to recall the number in the high load condition show significant revision, then reduced sample size is unlikely to be the explanation for why the 115 participants who did remember the number failed to show it.

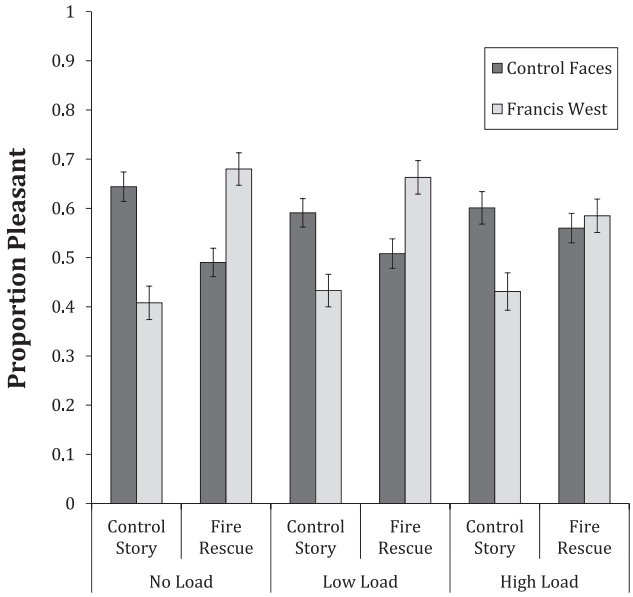

Figure 4. Mean proportion of ideographs judged more pleasant than average in Experiment 3 at Time 2 by cognitive load condition, story condition, and face prime. Error bars are *SE*s.

Moderation by extent of thinking. We conducted two additional analyses to determine whether, in the load conditions, the extent to which participants reported that they went back to consider the new information and the old information moderated the size of the revision effect. In the first analysis, we examined the moderating effect of thinking about the new information. The three-way effect of time, person, and story condition was moderated by the amount of thinking about the new information, $F(1, 244) = 4.27$, $p = .04$, $\eta_p^2 = .017$. Probing this interaction 1 *SD* above and below the mean level of thinking about the new information, we found that the revision effect (Time × Person × Story condition) was significant at high levels of going back to think about the new information in the story, $F(1, 244) = 15.76$, $p < .001$, $\eta_p^2 = .061$, but not at low levels of going back to think about the new information, $F(1, 244) = .89$, $p = .347$. Adding cognitive load condition to this model (low vs. high load) did not qualify the effect, $F(1, 240) = .12$, $p = .731$. Thus, the experience of going back to think on the new information appears to moderate the revision effect, and this was the case regardless of the amount of cognitive load.

In the second analysis, we examined the moderating effect of thinking about the old information. The three-way revision effect was not qualified by the amount of thinking about the old infor-

[8] The lack of interaction between time, person prime, and story in the high load condition is not because of differences at Time 1. At Time 1, there was no interaction between person, story condition, and cognitive load, whether the latter is coded as the original three-level factor, $F(2, 376) = .71$, $p = .492$, or a contrast of the high load group versus the other two: no load and low load; $F(1, 378) = 1.33$, $p = .250$. Indeed, in the high-load condition at Time 1, Francis was more negative than neutral faces in both the control and fire rescue conditions (both ps $< .001$).

mation, $F(1, 244) = .03$, $p = .854$. However, when cognitive load was added to the model, there was a significant interaction between time, person, story condition, cognitive load, and extent of thinking about the old information, $F(1, 240) = 9.52$, $p = .002$, $\eta_p^2 = .038$. Investigating this interaction revealed that for low levels of thinking about the old information, cognitive load did not moderate the revision effect, which was strong. However, at high levels of thinking about the old information, cognitive load moderated the revision effect, $F(1, 240) = 12.403$, $p = .001$, $\eta_p^2 = .049$. This effect is such that in the fire rescue condition, there is reversal only for low load ($F(1, 240) = 19.103$, $p < .001$, $\eta_p^2 = .074$), but not high load, $F(1, 240) = .608$, $p = .436$. In high load, Francis is only neutral at Time 2, with unfamiliar faces and Francis not differing, $F(1, 240) = .989$, $p = .321$, even though he's significantly negative at Time 1. Thus, when participants reported thinking a lot about the *old* information after no longer needing to remember the number, this tended to worsen the revision effect in the high cognitive load condition. Although one might have plausibly supposed that such thinking indicates active efforts at reinterpretation, it seems that such thinking is not an indicator of successfully making sense of the story. If thinking about the new information but not the old information seems to be linked with strong revision, this may mean that changing one's interpretation of the story is experienced as extensive thought about the *new* information, but not the old.

Explicit evaluations toward Francis West. We assessed explicit liking of Francis West in a 2 (Story Condition: Control vs. Fire Rescue) × 2 (Cognitive Load: No Load, Low Load, or High Load) ANOVA. Results indicated a main effect of story condition, $F(1, 376) = 1281.76$, $p < .001$, $\eta_p^2 = .773$, such that liking was higher in the fire rescue condition ($M = 5.69$, $SD = 1.59$) than in the control condition ($M = 1.25$, $SD = 0.61$). There was no main effect of cognitive load condition, $F(2, 376) = 1.66$, $p = .192$, or interaction between story condition and cognitive load condition, $F(2, 376) = 2.17$, $p = .116$. Thus, while cognitive load affected implicit evaluations, it did not affect explicit evaluations toward Francis.

Discussion

Consistent with our account that active thinking about the information is central to our reinterpretation effects, these results show that the availability of at least minimal cognitive resources is a necessary condition for full revision. Those in the high load condition, in comparison with those in low load or no load conditions, showed no moderation of their implicit evaluations by condition. Participants with less or no load, meanwhile, showed the pattern of results observed in our previous studies: significant reversal in the fire rescue condition and no change in the control condition. The failure of those in the high load condition to show the revision effect was not attributable to a failure to learn the new story information, or even an inability to come to the correct conclusions about the story when answering the comprehension questions. Instead, it seems that the high load burden interrupted not the *ability* to process the new information, which participants were able to do when queried, but rather their *tendency* to do so. The moderation of the revision effect by extent of going back to think about the new information suggests that among those who did, the reversal effect occurred. However, the failure of revision overall in the high load condition does suggest that on the whole, participants in this condition moved on to the AMP without doing this.

One interesting question is why did the self-report measure of going back to think about the new information produce a stronger reversal effect, while thinking about the old information seemed to do the opposite (at least under high cognitive load)? This question raises the issue of the type of active thinking required for revision in our studies. It is hard to interpret what participants might mean when they report thinking about the "old" versus the "new" information. These questions are too broad to pinpoint whether participants are rehearsing, reinterpreting, elaborating, rejecting, and so forth. Thus, in the next study we more precisely measured the type of thinking that we predict should produce revision: recognition that the new information produced *reinterpretation* of what had been previously learned about Francis.

Experiment 4

The comparison of the fire rescue and subway rescue conditions in Experiment 2 suggested that the revision effect emerged because of reinterpretation. In other words, revision occurred in the fire but not in the subway condition because of the degree to which the new, counterattitudinal information was able to *explain away* the initial information only in the former. The first goal of Experiment 4 was to extend the evidence in support of this argument by measuring the degree to which participants report reinterpreting the earlier story information after learning the fire rescue information about Francis West, to show that such subjective reinterpretation does indeed predict the revised implicit evaluations of Francis.

A second goal of Experiment 4, however, was to begin to address how reinterpretation might relate to, or differ from, other forms of thinking about the new information that could be responsible for our effects. In particular, to this point it is unclear how reinterpretation relates to research on elaboration. Much research has suggested that the extent to which one thoughtfully processes new information is an important predictor of its effect on explicit evaluations (for reviews, see Barden & Tormala, 2014; Petty, Haugtvedt, & Smith, 1995). When persuasive information is more thoroughly elaborated, it has a more powerful impact on impressions. A few studies have found similar effects of elaboration on established (Briñol, Petty, & McCaslin, 2009; Horcajo, Briñol, & Petty, 2010) or novel (Smith, De Houwer, & Nosek, 2013; Wyer, 2010) implicit evaluations. Of most relevance to our present purposes, Wyer (2010) found in one study that new information that suggested a reinterpretation of prior details did not produce revision of implicit evaluations *unless* participants were able to revisit all of the prior information upon which they had based their first impressions. Wyer (2010) argued, in line with Petty et al. (2006), that this suggested that for the old evaluation to be effectively tagged as false such that it would no longer impact implicit evaluations, participants needed this opportunity to carefully elaborate on the new revelation.

Is the reinterpretation in our studies effective because it forces participants to engage in extensive amounts of elaboration on the new information, but not necessarily reinterpretation in particular?

That is, perhaps the critical ingredient in reinterpretation in our paradigm is that it simply is an effective way to get people to do a lot of thinking about the new information, regardless of whether those thoughts are specifically about reinterpreting the meaning of Francis' earlier actions. From this perspective, the reason we have not seen revision in the subway condition (Experiment 2), for instance, is because that information, for whatever reason, was not sufficiently surprising or interesting to trigger enough elaboration (on that new information) to produce revision.[9] Furthermore, the findings from Experiment 3—that high cognitive load prevented full revision of implicit evaluations—show that some amount of active thinking is necessary to produce revision, but do not disambiguate what sort of elaboration participants are engaging in (reinterpreting the earlier details or otherwise).

If the critical aspect of our paradigm is that it produces a large amount of general elaboration on the new information, such that change is driven by the degree to which people think carefully about the new information in general rather that reinterpretation of the earlier details in particular, then just the degree of thinking about the new information should predict revision, and any self-reported reinterpretation would not independently contribute to the effect. However, if reinterpretation is the *specific form* of elaborative thinking that drives the effects, then we should find that the belief that new information changes the meaning of the old information predicts revision, even when the extent of thinking more generally is controlled. Such a finding would imply that the proximal mechanism in our studies is the recognition that the new information changes the meaning of the old. We test these two accounts in the next study.

We included only the fire rescue condition, and added three items: an item about reinterpretation (how much does the new information change the meaning of the prior events), and two items gauging degree of thinking about the new information more generally (how rapidly vs. gradually one's thinking proceeded, and how extensively one deliberated about the new information). We predicted that degree of thinking carefully (either gradually and/or extensively) might significantly predict revision, in line with previous research (Briñol et al., 2009; Petty et al., 2006; Wyer, 2010). However, we also predicted that the reinterpretation item would uniquely predict revision even while controlling for careful thinking. This would suggest that the reinterpretation happening in our studies—that is responsible for the revision effects—is a more specific mechanism than general elaboration on the new information.

Method

Participants. We recruited 75 participants from Amazon's Mechanical Turk Web site (www.mturk.com) to participate in the current study in exchange for $1.75 (36% male; M_{age} = 36.56 years, SD = 11.47). This smaller sample size was determined a priori based on the lack of between-participants conditions in the current study.

Materials. To assess whether revised implicit evaluations of Francis West would be predicted by the degree to which participants reinterpreted the earlier story details, we asked participants to respond to the following 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?" on a scale from 1 (*not at all*) to 9 (*a large amount*).

To begin to address whether reinterpretation in particular is predictive of revised implicit evaluations of Francis West, rather than elaborative thinking more generally, we added two further questions. The first was designed to measure the sense participants had of how quickly or gradually their thoughts about Francis came together after learning the new information. Specifically, they read:

> Sometimes, our thoughts come together quickly. At other times, our thoughts come together more gradually. When you got the new information about Francis West a moment ago, *did your thoughts about the meaning of Francis' actions come together quickly or more gradually*?

Participants responded on a scale from 1 (*quickly*) to 9 (*gradually*). Our second question aimed at tapping into the degree to which participants elaborated more generally (vs. reinterpretation in particular) focused on how extensive participants felt their thinking to be. They read:

> Sometimes, we deliberate a lot, and our thinking is very extensive. At other times, we deliberate less, and our thinking is less extensive. When you got the new information about Francis West a moment ago, *how much thinking did you do–not much deliberation or a lot of deliberation*?

and responded on a scale from 1 (*not much deliberation*) to 9 (*a lot of deliberation*). All of the other explicit measures from Experiment 3 were included, except for those dealing with the cognitive load manipulation from that study (including the extent to which they went back to think about the new and old information after no longer needing to remember the number), because the cognitive load task was not included here.

Procedure. All participants completed the fire rescue condition. They first read the Time 1 information, followed by the first AMP, and then were presented with the Time 2 fire rescue information as presented in Experiments 1a, 2, and 3, with one alteration to their instructions: To encourage more variability in the extent of reinterpretation, we simply asked participants to think about the final information, rather than specifically to think about how it relates to what they had previously read.

Next, right after reading the Time 2 (fire rescue) information, but before the second AMP, participants responded to the three questions regarding the nature of their thoughts at the time of learning the final details about Francis West. To increase the chance that participants would discriminate between these three (potentially highly related) questions, we presented all three ques-

[9] Indeed, we found that participants in the subway rescue condition in Study 2 reported thinking about the prior story details less than those in the fire rescue condition. However, when we went back and compared the fire rescue and subway rescue conditions with only those participants selecting the highest value of "7" on the thought extent scale, story condition still moderated the interaction between time and person, $F(1, 71) = 11.66$, $p = .001$, $\eta_p^2 = .141$. At Time 2, Francis was still significantly more negative than control faces in the subway condition ($F(1, 71) = 5.29$, $p = .024$, $\eta_p^2 = .069$) but significantly more positive than control faces in the fire condition ($F(1, 71) = 14.52$, $p < .001$, $\eta_p^2 = .170$). Therefore, reporting the max value on the measure of how much they thought about the story does not fully account for the difference in revision between the two conditions.

tions on the same screen and required participants to read all of them in advance, for at least 30 s, before moving on to answer them. To reduce potential noise from different question orders, we fixed the order of the questions, such that the meaning change question came first, followed by the thought speed question, and finally the deliberation extent question. Participants then completed the second AMP, the rest of the explicit measures, and were thanked, debriefed, and compensated.

Results

Data preparation. Following our procedure from the previous studies, we excluded from all analyzes the data from those participants who reported that they knew Mandarin or Cantonese (two participants; 2.67%), and any additional participants who used a single response key on all trials of at least 1 AMP (four participants; 5.33%). This left 69 participants in the analysis.

Implicit evaluations toward Francis West. Implicit evaluations were once again measured from the average proportion of pleasantness judgments of ideographs following the different face primes on the AMP. These AMP judgments were analyzed in a 2 (Measurement Time: Time 1 and Time 2) $\times$ 2 (Prime Person: Francis West and control faces) repeated-measures ANOVA. The anticipated interaction between measurement time and prime person was significant, $F(1, 68) = 23.33$, $p < .001$, $\eta_p^2 = .255$. At Time 1, Francis West was significantly less implicitly positive than control faces: Ideographs following Francis primes were judged to be more pleasant than average significantly less often ($M = .37$, $SD = .28$) than ideographs following control face primes ($M = .57$, $SD = .20$), $F(1, 68) =$, $p < .001$, $\eta_p^2 = .260$. At Time 2, however, implicit evaluations had reversed. Ideographs following Francis primes were judged to be more pleasant than average significantly more often ($M = .66$, $SD = .27$) than ideographs following control faces primes ($M = .53$, $SD = .23$), $F(1, 68) = 9.30$, $p = .003$, $\eta_p^2 = .120$.

Subjective thought measures and implicit evaluations. Next, to examine the relationship between the type of thinking that participants reported doing when reading the final information about Francis West and revised implicit evaluations toward him, we conducted a planned multiple linear regression analysis. The dependent variable was the average proportion of "pleasant" judgments following Francis West primes at Time 2 for each participant, with average pleasantness judgments of ideographs following Francis primes at Time 1, and control face primes at Time 1 and Time 2 entered as three covariates. The three measures of thought type (extent to which the new information changed the meaning of the old, speed of thought, and extent of deliberation) were the key predictors. All six of the model predictors were entered simultaneously in a single step. This allowed us to examine the potential for each predictor to have independent influences on final implicit evaluations of Francis West.

Results showed that self-reported extent to which the new information changed the meaning of the prior details had a uniquely predictive relationship with final implicit positivity of Francis West, $\beta = .287$, $t(62) = 2.30$, $p = .025$. However, the measure of whether participants felt their thoughts came together rapidly versus gradually had no relationship with Time 2 implicit evaluations of Francis, $\beta = .001$, $t(62) = .009$, $p = .993$, and neither did the measure of how extensive participants reported their thinking to be, $\beta = .052$, $t(62) = .420$, $p = .676$. To check the robustness of the relationship between the measure of meaning change and implicit evaluations of Francis, as well as to determine whether either of the other two thought measures would predict implicit evaluations if the other thought measures were omitted, we conducted a series of exploratory follow-up regressions. Specifically, we examined regressions that included all possible subsets of the three thought measures (with the same covariates of other AMP trial types), and found that in none of these models did either the thought speed or deliberation extent measure produce a significant effect, all ps $> .5$. Additionally, the meaning change measure never became nonsignificant, all ps $< .05$.

The sample as a whole also strongly endorsed the view that the new information changed the meaning of the prior events ($M = 8.59$ out of 9, $SD = .99$), that their thinking proceeded quickly rather than gradually ($M = 2.41$ out of 9, $SD = 2.10$, where higher values indicate more gradual thinking), and that their deliberation was not very extensive ($M = 2.87$ out of 9, $SD = 2.44$), suggesting that revision here tends to be experienced as relatively easy (provided that cognitive resources are not maximally strained as in Experiment 3). Indeed, the reported extent to which the meaning of the initial story had changed was negatively correlated with both the degree to which thinking proceeded gradually, $r(67) = -.41$, $p < .001$, and extent of deliberation, $r(67) = -.36$, $p = .002$. The degree to which thinking was gradual (vs. fast) correlated positively with the extent of deliberation, $r(67) = .45$, $p < .001$.

Explicit evaluations toward Francis West. We once again used an average of responses on the six questions gauging liking of Francis West to assess changes in explicit liking over time. In a paired-samples t test we found that, unsurprisingly, explicit liking of Francis West was much higher at Time 2 ($M = 6.27$, $SD = .97$) than at Time 1 ($M = 1.21$, $SD = .51$), $t(68) = 38.83$, $p < .001$.

Further, we conducted a similar multiple regression analysis to that performed on implicit evaluations toward Francis West. The index of explicit liking at Time 2 was regressed on liking at Time 1, as well as the belief that the new information changed the meaning of the old, the gradualness of thought, and the extent of thought. Paralleling the results with implicit evaluations, we found that the extent to which participants felt the new information changed the meaning of the old information predicted Time 2 explicit liking of Francis, $\beta = .664$, $t(64) = 6.87$, $p < .001$. However, both the thought speed measure ($\beta = -.178$, $t(64) = -1.70$, $p = .095$) and the deliberation extent measure ($\beta = .070$, $t(64) = .69$, $p = .494$) did not.

Discussion

The results supported our account that reinterpretation in the Francis West paradigm operates through a separate mechanism from general elaborative thinking. Although greater contemplative thought (in terms of either self-reported thought speed or extensiveness of deliberation) did not correlate significantly with greater revision, there was a unique, strong impact of belief that the new information changed the meaning of the prior story. Additionally, the distributions of responses on the measures and their negative correlation suggested that recognition of the new information's explanatory value was linked with having thoughts come together quickly rather than gradually, and with less extensive thinking. Reinterpretation seems to require at least a brief revisit of the prior

story details so as to reframe their meaning, which requires the availability of at least some cognitive resources (Experiment 3), but it is not akin simply to extensive, general elaboration.

Experiment 5

Experiment 2 demonstrated that information that reinterprets the prior events produced much stronger change (indeed, a reversal) than equally positive information that does *not* reinterpret the prior events. Building the case that reinterpretation is the operative mechanism in driving the revision effect in the fire rescue condition, Experiment 4 showed that the extent to which participants believed that the final information altered the meaning of the previous events was significantly correlated with final implicit evaluations of Francis West, while more general measures of elaboration were not. However, we have not yet demonstrated that reinterpretation per se *mediates* the greater revision in the fire rescue condition relative to other conditions.

In this next experiment, we tested mediation by including solely the fire rescue and subway rescue control conditions from Experiment 2. In both conditions, we expected that the extent to which participants reported the new information to alter the meaning of the prior story events should predict the degree of their revision (some reinterpretation might occur among participants in the subway condition if they, say, suspected that his heroics might suggest that there was some unknown good reason behind his seemingly negative actions in his neighbors' homes). Furthermore, because the fire rescue information was expected to prompt this reinterpretation to a larger degree than the subway rescue information, we predicted that reinterpretation would mediate the effect of story condition on implicit evaluations. In addition, we retained the more direct measure of general elaboration from Experiment 4 (extent of thinking) to demonstrate that only reinterpretation, and not elaborative thinking more generally, would mediate the effect of story condition on revised implicit evaluations of Francis West.

Method

Participants. There were 296 participants recruited on Amazon's Mechanical Turk Web site (www.mturk.com) who participated in return for $1.75 (49% male; M_{age} = 33.72 years, SD = 10.58). We intended to recruit 300 participants (150 per between-participants condition), but a transient server error interrupted the experiment for four individuals, making it impossible for them to continue the study. They were compensated for their time, but their incomplete data were not included in any analyses.

Materials. Participants viewed the events from the same initial story used in Experiments 1a, 2, 3, and 4. The Time 2 information consisted of that from either the fire rescue or subway rescue stories conditions used in Experiment 2. To measure the extent to which participants subjectively reinterpreted the earlier story events in light of the new information, we asked participants to respond to the following single item adapted from Experiment 4: "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?" on a scale from 1 (*not at all*) to 9 (*completely*). To measure extent of more general elaborative thinking, participants responded to the same deliberation question from Experiment 4. All of the other explicit measures from Experiment 4 were included in this study, except for the item that gauged the speed with which thoughts came to mind.

Procedure. Participants were assigned to either the fire rescue or subway rescue condition, and completed the study in an identical fashion to Experiment 4. Immediately after reading the Time 2 information, participants responded to the subjective meaning change measure and the deliberation extensiveness measure before moving on to the second AMP. The order of these two questions was counterbalanced. To reduce the chance that the order in which the two questions were asked might influence participant responses, we again presented the questions on the same screen and asked participants to read both for at least 20 s before answering them. (The order of the two questions on the screen produced no significant effects in any analyses and is thus not discussed further.)

After answering the two questions about their thoughts when presented with the fire or subway rescue information, participants completed the second AMP, the rest of the explicit measures, and were thanked, debriefed, and compensated.

Results

Data preparation. In keeping with the exclusion criteria from our previous studies, we dropped all data from nine participants for familiarity with Mandarin and/or Cantonese (3.0%) and 21 more for using a single key on every trial of at least one of the two AMPs, thus failing to follow instructions (7.1%). This left 266 cases for analysis.

Implicit evaluations toward Francis West. We analyzed average pleasantness judgments of ideographs on the AMP in a 2 (Measurement Time: Time 1 and Time 2) × 2 (Prime Person: Francis West and control faces) × 2 (Story Condition: fire rescue or subway rescue) mixed ANOVA, with the first two factors varying within-participants and the third between-participants. The anticipated interaction between time, prime person, and story condition obtained, $F(1, 264) = 44.41$, $p < .001$, $\eta_p^2 = .144$. Simple effects tests revealed that at Time 1, ideographs following the image of Francis West were rated more negatively than those following control faces in both the fire rescue condition ($M_{Francis} = .38$, $SD_{Francis} = .28$; $M_{Control} = .61$, $SD_{Control} = .20$; $F(1, 264) = 56.69$, $p < .001$, $\eta_p^2 = .177$) and the subway rescue condition ($M_{Francis} = .36$, $SD_{Francis} = .28$; $M_{Control} = .65$, $SD_{Control} = .23$; $F(1, 264) = 75.84$, $p < .001$, $\eta_p^2 = .223$). At Time 2 in the fire rescue condition, implicit evaluations toward Francis West had reversed: ideographs following Francis primes were rated significantly more positively ($M = .70$, $SD = .26$) than those following control primes ($M = .51$, $SD = .24$), $F(1, 264) = 27.52$, $p < .001$, $\eta_p^2 = .094$. However, at Time 2 in the subway rescue condition, the initial implicit evaluations were attenuated but not reversed. Ideographs following Francis primes were still rated significantly more negatively ($M = .41$, $SD = .28$) than ideographs following control face primes ($M = .66$, $SD = .25$), $F(1, 264) = 48.35$, $p < .001$, $\eta_p^2 = .155$. The interaction between time and prime person was significant in the fire rescue condition, $F(1, 264) = 108.58$, $p < .001$, $\eta_p^2 = .291$, but not in the subway rescue condition, $F(1, 264) = .55$, $p = .459$.

Mediation by reinterpretation. Next, we turned to our central interest in this experiment: Examining whether subjective degree of change in the meaning of the old story details in light of the new

information would uniquely mediate the fire rescue versus subway rescue condition difference in final positivity toward Francis West, even when controlling for extent of general elaboration. Using the PROCESS tool for SPSS (Hayes, 2013), we conducted a bias-corrected bootstrap mediation analysis using 10,000 samples. The dependent variable was the proportion of ideographs judged pleasant following Francis West primes at Time 2, with proportions following neutral primes at Time 1 and 2 and Francis primes at Time 1 entered as covariates. The independent variable was story condition (fire = 1, subway = 0) and the mediators were self-reported extent of change in the meaning of the initial story information and amount of deliberation. The two potential mediators were entered into a single model in parallel, but the interpretation of the results does not change in significance or direction if the mediators are run in separate analyses. Table 2 shows the zero-order correlations among all variables included in the model.

The analysis yielded a significant indirect effect for the mediation of story condition through reinterpretation, estimate = .146, 95% CI: [.0361, .2438], Sobel Z = 2.56, $p = .011$. There was no parallel indirect effect through extent of deliberative thinking, estimate = .004, 95% CI: [−.0034, .0177], Sobel Z = .78, $p = .436$. Figure 5 illustrates the mediation model.

Explicit evaluations toward Francis West. We assessed changes in explicit evaluations of Francis West in a 2 (Measurement Time: Time 1 and Time 2) × 2 (Story Condition: fire rescue or subway rescue) mixed ANOVA, with measurement time varying within-participants and story condition varying between-participants. The dependent variable was the average of the six explicit liking scales used in each of the prior experiments. Story condition significantly affected the change in explicit liking over time, $F(1, 264) = 826.91$, $p < .001$, $\eta_p^2 = .758$. Francis increased in explicit positivity in the fire rescue condition from Time 1 ($M = 1.26$, $SD = .78$) to Time 2 ($M = 6.29$, $SD = 1.15$), $F(1, 264) = 2488.31$, $p < .001$, $\eta_p^2 = .904$, as well as in the subway rescue condition from Time 1 ($M = 1.22$, $SD = .53$) to Time 2 ($M = 2.06$, $SD = 1.00$), $F(1, 264) = 62.43$, $p < .001$, $\eta_p^2 = .191$, but was significantly more positive at Time 2 in the fire rescue condition than the subway rescue condition, $F(1, 264) = 1017.61$, $p < .001$, $\eta_p^2 = .794$. Furthermore, in a mediation analysis similar to that performed on implicit evaluations, we found that change in meaning of the initial story details mediated the effect of story condition on Time 2 explicit liking of Francis West (controlling for Time 1 explicit liking), estimate = 2.75, 95% bias-corrected bootstrap CI (10,000 samples): [2.1452, 3.2683], Sobel $Z = 12.61$, $p < .001$. On the other hand, there was no significant indirect effect through extent of deliberation, estimate = −.01, 95% bias-corrected bootstrap CI (10,000 samples): [−.0391, .0107], Sobel $Z = -.45$, $p = .650$.

Discussion

As predicted, the extent to which participants reported that the new information changed the meaning of the earlier details of the story mediated the effect of the new information (fire rescue vs. subway rescue) on revised implicit evaluations of Francis West. That is, participants tended to engage in more reinterpretation in the fire rescue condition, and the extent to which they did so predicted the greater revision in that condition. More important, we also found no evidence that a more general measure of elaborative thinking—that asked participants to report whether they deliberated over the new information more or less extensively—mediated the condition difference in revision. Even when controlling for a potential indirect effect of general extent of thinking, the degree to which participants reported the meaning of the old information was changed by the new information significantly mediated the difference between the fire rescue and subway conditions.

Experiment 6

For our final study, we addressed a potential concern that the revised implicit evaluations produced in this work might not be durable, "real" change, but perhaps a transient effect in which the powerful new information is especially salient. This might produce a brief shift that masquerades as real change, only to revert back to the initial evaluation after the passing of time.

We sought to demonstrate the longevity of the revised implicit evaluations. Indeed, a changing temporal context has been noted as a potential source of spontaneous recovery of conditioned responses (Bouton, 1993; Bouton, Westbrook, Corcoran, & Maren, 2006), and so demonstrating no return of the initial negative implicit evaluation of Francis West in the revision condition would be informative. In showing the endurance of the revised implicit evaluations, we can suggest that there is nontrivial durability and, thus, "realness" to these evaluations.

To examine the effects of time on implicit evaluation revision, we repeated our basic revision procedure. Then, participants were

Table 2
Zero-Order Correlations Between All Variables in the Mediation Model in Experiment 5

Measure	1	2	3	4	5	6
1. Story condition						
2. Time 2 Francis pleasantness	.47***					
3. Time 2 control pleasantness	−.29***	−.36***				
4. Time 1 Francis pleasantness	.03	.29***	.04			
5. Time 1 control pleasantness	−.08	.01	.44***	−.08		
6. Meaning change	.89***	.49***	−.35***	.02	−.11	
7. Extent of deliberation	−.12†	−.07	−.07	−.07	.00	−.03

Note. Cell values are Pearson correlations. Story condition is coded 0 = subway rescue, 1 = fire rescue. Pleasantness covariates refer to the average portion of ideographs judged to be more pleasant than average, by time and prime type.
† $p < .1$. *** $p < .001$.

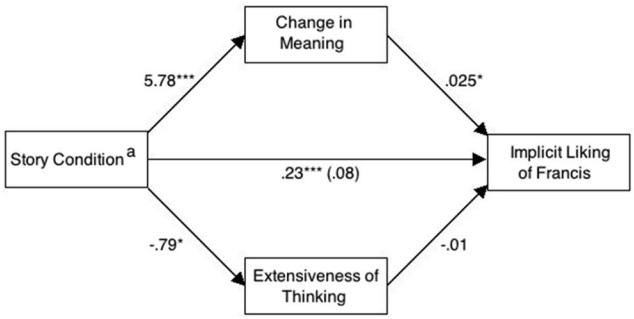

Figure 5. Mediation of story condition effect on Time 2 judgments of ideograph pleasantness following Francis West primes, through subjective change in meaning of the Time 1 information and general extensiveness of thinking in Experiment 5. Slopes are unstandardized regression coefficients. Model controls for judgments of ideographs following Francis West primes at Time 1 and control face primes at Times 1 and 2. [a] Coded 0: subway rescue, 1: fire rescue. *** $p < .001$, * $p < .05$.

invited to return for a follow-up study 3 days later. At that time, they were told simply that they might remember reading a story about a man named Francis West, but that before they would be asked about this they were to complete a different task (a third implicit measure). To show the durability of revision, we expected to observe relatively no change in revised implicit evaluations across the delay. To keep things simple, we used only the fire rescue and control conditions.

Method

Participants. There were 301 participants who were recruited from Amazon's Mechanical Turk Web site (www.mturk.com) to take part in this two-session study in return for $1.50 paid compensation (53.5% male; M_{age} = 32.02 years, SD = 10.08). Because we were uncertain about how much attrition there would be between the two study sessions, and about whether the effect size would be much reduced after a delay, we opted a priori to collect data from 300 participants so as to fill each of two between-participants conditions with 150 participants. Data from an additional participant were recorded because one person completed the study without submitting the request for payment on Mturk.

Materials. The story materials and AMP stimuli used in this study for Time 1 and Time 2 were identical to those used in the previous experiments, including the control and fire rescue conditions, with one exception: At Time 2, participants in that condition now read that he had a criminal history, yelled at children, and broke into the houses in revenge against them as well as to steal valuables (rather than that he started throwing rocks at the houses). This change was made to better equate the two conditions on the degree to which the final information provided motive for his actions. At Time 3, the AMP used the same prime images as the preceding AMPs and one of two new randomly chosen sets of 40 ideographs as targets. Thus, the same ideographs were never rated twice by an individual participant. The explicit evaluation toward Francis at Time 3 was measured using the same scale as used at Times 1 and 2. For exploratory purposes, we added a single item right before the demographic questions at Time 2 asking participants to self-report their mood ("Indicate how you feel right now, that is, at the present moment") on a scale from 1 (*very bad*) to 7 (*very good*), and an item gauging the extent to which they thought the story depicted "real" events ("To what extent do you believe that the Francis West story is based on real events?") from 1 (*not at all*) to 7 (*completely*).

Procedure. After being assigned to either the fire rescue or control story conditions, participants completed the same procedure as in Experiment 3 no cognitive load condition, without the questions about thought type. To minimize noise, all participants completed the AMP before the explicit evaluation scale at each measurement instance. After completing the various questionnaire items at Time 2 (explicit evaluation scale, comprehension checks, manipulation checks, mood, belief the story was real, and demographic questions), participants were then informed that they had the option of entering an email address so that we could contact them in 3 days with a short follow-up study, which they would receive extra compensation for completing. They were told that not doing so would have no impact on their compensation for what they had already completed; all but six did so. Approximately 3 days later, participants received an email inviting them back for the short final session of the study, and were given a window of 24 hr in which to complete it; 63.1% of participants returned and completed the final session. Attrition was equally likely in the two story conditions, $\chi^2(1) = .12$, $p = .725$.

Upon beginning the final session, all participants were told, "Three days ago you read a story about a man named Francis West. You will be asked to answer questions about him in a few moments; please do your best to answer these questions regardless of how much you remember. But first, there is another task to do." They then completed the third AMP, followed by the third administration of the explicit evaluation scale (dubbed "Time 3" hereafter).

Results

Data preparation. Implicit evaluations on each of the three AMPs were computed in the same manner as done in previous studies, as were the three repetitions of the explicit evaluation scale. All analyses were conducted solely on those participants who completed the second session of the study. In addition, 10 participants were dropped for using one key on every trial of at least one AMP, thus, disregarding instructions, and one more was dropped for familiarity with Mandarin or Cantonese. This left 179 cases for analysis.

Implicit evaluations toward Francis West. Implicit positivity on the AMP toward Francis West was assessed in a 2 (Story Condition: Fire rescue or control) × 2 (Prime Person: Francis West vs. control faces) × 3 (Measurement Time: Time 1, Time 2, and Time 3) Mixed ANOVA, with the first factor varying between-participants and the latter two within-participants. All effects were significant, including the crucial three-way interaction, $F(2, 175) = 25.60$, $p < .001$, $\eta_p^2 = .226$. In Figure 6 we show the means and *SE*s of positivity toward the Francis and control primes in each of the story conditions and each of Times 1, 2, and 3. More important, Francis West was significantly more negative than the control faces in the control condition at all measurement instances, as well as the fire rescue condition at Time 1, all ps < .001; additionally, he was more positive than the control faces in the fire rescue condition at Times 2 and 3, both ps < .001. Thus,

with only the barest of reminders about the study, implicit evaluations toward Francis persevered in both the fire rescue and control conditions (positive and negative, respectively) for the 3 days between Session 1 and Session 2 of the experiment. Time effects showed that in the control story condition, positivity toward Francis did not shift between any two measurement times, all $ps > .1$. In the fire rescue condition, Francis West was significantly more positive at Time 2 ($M = .70$, $SD = .23$) than at Time 1 ($M = .41$, $SD = .27$), $p < .001$, and was marginally less positive at Time 3 ($M = .66$, $SD = .23$) relative to Time 2, $p = .071$. However, even at Time 3 he was still more positive than at Time 1, $p < .001$.

Consistent with prior testing, neither the degree of belief that the study depicted true events nor subjective mood moderated these results. Furthermore, neither reduced the significance of the key interaction when added to the model.

Explicit evaluations toward Francis West. Explicit liking of Francis West was analyzed in a 2 (Story Condition: fire rescue or control) × 3 (Measurement Time: Time 1, Time 2, and Time 3) mixed ANOVA, with the former factor varying between-participants and the latter varying within-participants. Both main effects were significant, as was the hypothesized interaction, $F(2, 175) = 617.26$, $p < .001$, $\eta_p^2 = .876$. Simple effects tests showed that explicit liking of Francis did not differ between story conditions at Time 1, $F(1, 176) = .127$, $p = .722$, but that Francis was significantly more liked at Time 2 in the fire rescue condition ($M = 6.18$, $SD = .98$) than in the control condition ($M = 1.11$, $SD = .27$), $F(1, 176) = 2181.37$, $p < .001$, $\eta_p^2 = .925$. At Time 3, Francis was also more liked in the fire rescue ($M = 6.00$, $SD = 1.21$) than the control condition ($M = 1.43$, $SD = .99$), $F(1, 176) = 760.82$, $p < .001$, $\eta_p^2 = .812$. Figure 7 illustrates the explicit liking of Francis West at each measurement time in both story conditions.

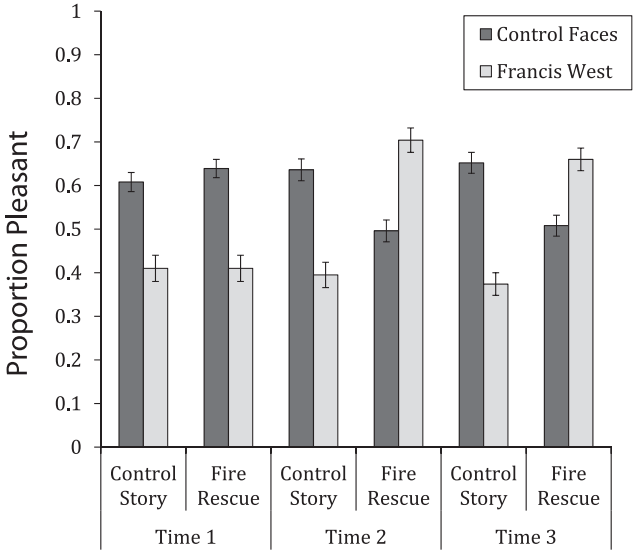

Figure 6. Mean proportion of ideographs rated as more pleasant than average in Experiment 6 by measurement time, story condition, and face prime. Error bars are *SE*s.

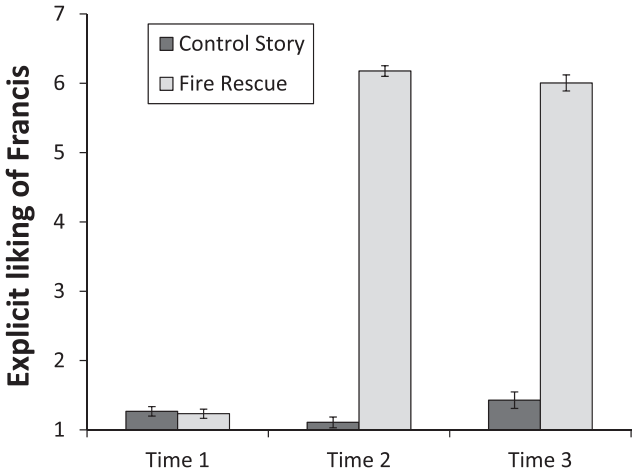

Figure 7. Mean explicit liking of Francis West in each story condition at each measurement time in Experiment 6. Error bars are *SE*s.

Discussion

After 3 days, implicit evaluations were still significant and positive (and revised from Time 1) in the fire rescue condition, and still significant and negative (i.e., unchanged) in the control condition. Participants were not even re-exposed to his image until it was primed during the final AMP, and his name was not a prime on the task. This durability of the implicit evaluations supports our claim that these revised evaluations are not merely transient effects of the testing situation that will revert back to their initial levels after the passing of a bit of time.

General Discussion

Across seven studies, we found strong and consistent evidence that initial implicit evaluations can be undone. More important, we identified one factor capable of prompting such revision: reinterpretation of the knowledge that formed the basis of the initial evaluation. In Experiments 1a and 1b, participants who had formed an initial negative implicit evaluation toward Francis West fully reversed those evaluations when they learned a reason for his bad behavior.[10] Experiment 2 showed that this reversal occurred only when participants read that his prior actions were aimed at saving children from a fire, but not when they read about a similarly

[10] As one anonymous reviewer pointed out, in most of our experiments in the fire rescue condition there was not only a significant increase in the proportion of "pleasant" responses on trials with Francis West primes from Time 1 to Time 2, but a significant *decrease* in that proportion on trials with neutral face primes. We see two possibilities for why this occurs. One is that this may be simply an artifact: only one prime is highly relevant (Francis West), and the others are distractors of a sort. This may amplify the tendency to respond on those trials in an opposite way compared with the Francis West trials, and would be especially pronounced if a participant is trying to use the two response keys relatively evenly (see Scherer & Lambert, 2009, for an investigation of such contrast effects in priming tasks). Another possibility is that this effect demonstrates revision because of implicit social comparison. We view this as an exciting theoretical possibility. However, regardless, this effect does not pertain to our main finding of interest, which is the implicit positivity or negativity of Francis West relative to the control trials.

positive action that did not explain his prior behavior (saving a baby from an oncoming train). This suggests that this revision depends on a recognition of the relationship between this new information and the old. Experiment 3 showed that this recognition requires at least minimal cognitive resources by demonstrating that high cognitive load at the time of reading the new information, relative to low load or no load, prevented full revision. Experiment 4 showed that subjective change in meaning of the earlier details of the story predicted more positive final implicit evaluations of Francis West in the fire rescue condition, and also showed that change in this paradigm is specific to the degree to which participants reinterpreted the earlier story rather than how much they thought about the information more generally. Experiment 5 demonstrated that reported change in the meaning of the earlier story details mediated the enhanced revision in the fire condition relative to the subway condition, which presented equally positive, but nonreinterpreting, information. Once again, a more general measure of elaborative thinking was not a significant mediator. Finally, Experiment 6 showed that the implicit evaluations formed in both the control and fire rescue conditions were not fleeting. After 3 days, implicit positive and negative evaluations were still apparent in the fire rescue and control conditions, respectively.

Collectively, these studies represent a closer examination of the conditions under which durable revision of implicit evaluations is possible, and identify one mechanism: reinterpretation. The empirical record to date has suggested that implicit evaluations are often resistant to efforts to undo a prior impression, even mere minutes after their initial formation (Gregg et al., 2006; Wilson et al., 2000). When they have shifted at all, it has been toward neutrality or ambivalence (Boucher & Rydell, 2012; Petty et al., 2006, Study 1; Peters & Gawronski, 2011, Experiment 3), or not occurred unless participants were compelled to engage in substantial elaboration (Wyer, 2010). Recent work by Cone and Ferguson (2014) found that initial positive implicit evaluations can indeed be overturned by new and highly diagnostic negative information, but the cognitive mechanism through which such change occurs, as well as whether revision could overturn an initial negative implicit evaluation, remained to be examined. Our work both shows negative-to-positive change in implicit evaluations and identifies a mechanism driving it: not only must new information imply the opposite evaluation of the target, but the initial information must also be reframed. In other words, the reinterpretation that we used both invalidated the initial learning, and replaced it with new meaning.

In combination, our studies suggest that reversal of implicit evaluations can occur through reinterpretation, we have characterized some of the features of this process, finding it to be deliberate enough to require at least minimal cognitive resources but not to be interchangeable with just *any* extensive thinking about the story information. Although the task of identifying the complete set of requirements for implicit evaluation change through reinterpretation will extend beyond this set of studies, we now turn to a discussion of theoretical implications.

Reconciliation: What Was Different This Time?

The present studies identify the role of reinterpretation in implicit evaluation revision. However, one might note that in prior studies that failed to find full reversal, researchers similarly attempted to make appeals to the irrelevance of the initial information (e.g., Gregg et al., 2006; Peters & Gawronski, 2011, Experiment 3; Petty et al., 2006, Study 2). There are numerous possibilities for why our paradigm showed reversal whereas prior similar attempts did not. For one, the instructions in these prior studies prompted a reinterpretation of the old information that was somewhat different from our version of reinterpretation. Whereas in our studies the actions of Francis are no longer negative in light of the revelations about the situation, in those prior studies they suggested that the *targets* of the old information should be changed. In other words, in prior studies, the actions of aggressors are still negative, but just do not correspond to the group or person one initially thought they did. This type of reinterpretation may not be as easily implemented as in our case where an understanding of the new information basically compels the overturning of the initial impression (i.e., If Francis West was trying to save those kids, then his earlier actions were highly likely to have been enacted for that effort). It also may be that trying to realign behavior with new targets is ineffective at eliminating all traces of the initial information, as is the case with directed forgetting (Bjork & Bjork, 2003) or adding new information to the old but not replacing it, as in the formation of contextualized evaluations (Gawronski & Cesario, 2013) and implicit ambivalence (Petty et al., 2006). Lastly, studies in which the targets change over time might also not produce as unified an impression as the paradigm we used here (e.g., Rydell et al., 2007). Ultimately, however, the task of identifying the differences between our paradigm and previous paradigms remains to be taken up.

Our findings offer new empirical support for the theoretical claim that reason-based routes to implicit evaluation revision should be possible. For instance, under the APE model (Gawronski & Bodenhausen, 2006, 2011), propositional reasoning is assumed to be capable of changing associative structure when new information is validated, but the parameters of this route of change have not been fully specified. Some studies have assumed that this change operates through the associative pairing of information contained within the propositions (Peters & Gawronski, 2011). However, the APE model does not outline when some kinds of affirmations will be more effective than others. Our results suggest that the effect of affirmation of new information will be especially strong when the new information *recasts* prior details, and that this process can produce full revision to the point where little evidence of the prior evaluation remains. Indeed, this recasting is able to force revision even in the case of an initial negative implicit evaluation turning to positive, an effect that has been particularly difficult to obtain (Cone & Ferguson, 2014).

Likewise, under the metacognitive model (Petty et al., 2007), new "false" tags on old associations are assumed to be able to negate beliefs, but the conditions under which this occurs remain unspecified. Although the results in Wyer (2010) suggested that revisiting the prior details in light of the new information was necessary for the initial evaluation to be undone, the mechanism of change was unclear. Our results suggest that reinterpretation can lead to revision without representing the initial information, and pinpointed reinterpretation in particular as the type of reasoning that drives change in this paradigm, beyond more general elaborative thinking. As such, our results can be read as expanding the routes of revision of implicit evaluations under current theories, providing evidence of when and how such change occurs.

Implications for Established Evaluations

Despite our results, a consistent finding in research on implicit evaluations is that people's implicit evaluations can be at odds with what they explicitly believe (Banaji & Greenwald, 2013). Our results do not imply that any and all implicit evaluations can be easily and rapidly changed through reasoning (reinterpretation or otherwise), just that there may be routes through which established implicit evaluations can be changed that have not been highlighted before, and which future investigations may profitably explore. Our view is that although we have explored implicit evaluation change in a specific scenario, we suspect that the mechanisms it illuminates operate in a variety of more mundane settings. That is, any time that a person *construes* new information to require a reinterpretation of prior knowledge about an evaluation object, we would expect shifts to occur in implicit evaluations proportionate to the amount and extremity of reinterpretation. An important future direction for this line of work is to examine situations in which this will be true. We examine some of these considerations below.

Insight into the basis of the initial implicit evaluation. In our studies, it is safe to assume that participants are quite aware of the basis for their evaluative feelings toward the target person, Francis West. However, with evaluations formed over a long period of time, people may have relatively poor introspective access to which content in memory actually affects their implicit evaluations. Insight into which information shapes judgments and impressions is often poor and based on inaccurate inferences or selective sampling of reasons (Wilson, Dunn, Kraft, & Lisle, 1989). Although the claim that implicit evaluations are inaccessible to consciousness (e.g., Greenwald & Banaji, 1995) has been challenged (Gawronski, Hofmann, & Wilbur, 2006; Gawronski, LeBel, & Peters, 2007; Hahn, Judd, Hirsh, & Blair, 2014), it remains likely that similar to the lack of source awareness of explicit evaluations (Bornstein, 1989; Hovland, Lumsdaine, & Sheffield, 1949; Kumkale & Albarracín, 2004; Wilson et al., 1989; Zajonc, 1968), people are not always aware of or may not remember the sources of their implicit evaluations (e.g., Dijksterhuis, 2004; Olson & Fazio, 2001, 2002; Rydell et al., 2006; see Gawronski et al., 2006). Thus, the reasons for an evaluation that people bring to mind when reflecting on their impressions of a person or group may differ from those that actually guide behavior. If rejection or reinterpretation of the basis of an initial evaluation is capable of revising even established implicit evaluations, as we posit here, this may be difficult if those sources in memory are forgotten, were never known, or are inaccurately identified (see also Lane, Ryan, Nadel, & Greenberg, 2014). To the extent that they are able to identify reasons that *do* contribute to their current implicit evaluation, they may be able to enact a greater amount of revision then they could otherwise, a possibility for future investigation.

Lack of information that prompts a full reinterpretation. In the Francis West paradigm, the scenario is designed to allow for a single piece of new information to completely recast everything that was previously learned about the target person. One explanation for why implicit evaluation change might be more difficult for many established evaluations is that this type of new information may not be encountered. However, such cases exist, such as when a Nazi seems like he is supporting the holocaust when in fact he is saving over 1,000 people. In fact, any situation in which ulterior motives come to light is a potential case for revision of first impressions to occur.

Motivational considerations. Another relevant consideration is that people are not always willing or able to process information in an unbiased way (Frey, 1986; Kunda, 1990; Lord, Ross, & Lepper, 1979). Under such circumstances, even when information that contradicts a current position is attended, it is often held to a higher bar than information that supports one's position (Ditto & Lopez, 1992; Ditto et al., 2003; Eagly et al., 2000).

The Francis West paradigm all but compels participants to change their minds about the target person; preserving the initial evaluation is quite indefensible. Though this may sometimes be the case in real life (e.g., discovering someone to be the victim of false and malicious accusations), it may be more typically the case that the effect of new information on an existing impression is a function less of the properties of that information itself than the manner and extent to which it is elaborated (Greenwald, 1968; Petty & Cacioppo, 1986; Petty, Ostrom, & Brock, 1981). Often, sweeping recalibration of past beliefs about an evaluation object may primarily occur when one is motivated to construe new information as prompting such revision, rather than new information inherently *requiring* such recalibration to occur.

Implicit Versus Explicit Evaluations: How Do They Relate to One Another?

The central theoretical contribution of the present work is the demonstration that changes to the meaning of prior information can lead to a full reversal of previously learned implicit evaluations. And yet, what does this mean for the presumed relations between implicit and explicit evaluations? After all, claims about (two) different processes underlying each type of evaluation have been used to explain the many examples of dissociation among them (e.g., Gawronski & LeBel, 2008; Gawronski & Strack, 2004; Petty et al., 2006; Rydell et al., 2006). If the processes underlying them are not as distinct as assumed—as our findings might imply—then why do we see so much evidence for dissociation elsewhere? One potential explanation for such dissociations concerns the lack of "structural fit" among explicit versus implicit measures, including features such as format, stimuli, instructions, and so forth (Payne, Burkley, & Stokes, 2008). These differences could explain dissociation without necessarily invoking any claims about underlying processes. Although there is a small amount of data suggesting that implicit and explicit evaluations differ even when controlling for fit (see Payne et al., 2008), this remains an open empirical question.

Another consideration is the tendency to assume that differences in behavior are because of dissociated processes. It is difficult (for us) to think of any behavioral evidence that would alone adjudicate between propositional versus associative processing, because any behavioral evidence can always be explained by boot-strapped versions of one's favorite propositional or associative account (see Ferguson, Mann, & Wojnowicz, 2014; Moors, 2014). What we can do, however, is specify the circumstances under which implicit and explicit evaluations form, change, and predict behavior. One can then create computational models that formally test theories of associative versus propositional processing, as has been done frequently in cognitive psychology (e.g., Botvinick & Plaut, 2006; Read & Montoya, 1999; Sun, Slusarz, & Terry, 2005). For now, we have demonstrated one

way in which implicit evaluations can be completely undone through a propositional, or reason-based, route. What remains is the work of figuring out what these findings mean for how implicit and explicit evaluations relate to one another.

Conclusion

Implicit first impressions are not immune to revision through reason. Far from being "stuck" in dogged opposition to our reasoned conclusions about the validity of prior impressions, our implicit evaluations can reflect our updated interpretations of the world. Our findings suggest that to change unwanted implicit evaluations, we may marshal reason to undermine the *bases* of our evaluations, if we can identify and edit them. Future work can examine other routes to implicit evaluation change and identify the conditions under which they are successful.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage Publications.

Amodio, D. M. (2014). Dual experiences, multiple processes: Looking beyond dualities for mechanisms of the mind. In J. S. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 560–576). New York, NY: Guilford Press.

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

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

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

Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality and Social Psychology Bulletin, 38,* 1194–1208. http://dx.doi.org/10.1177/0146167212446835

Barden, J., & Tormala, Z. L. (2014). Elaboration and Attitude Strength: The new meta-cognitive perspective. *Social and Personality Psychology Compass, 8,* 17–29. http://dx.doi.org/10.1111/spc3.12078

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

Bjork, E. L., & Bjork, R. A. (2003). Intentional forgetting can increase, not decrease, residual influences of to-be-forgotten information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 524–531. http://dx.doi.org/10.1037/0278-7393.29.4.524

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

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,* 828–841. http://dx.doi.org/10.1037/0022-3514.81.5.828

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin, 106,* 265–289. http://dx.doi.org/10.1037/0033-2909.106.2.265

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review, 113,* 201–233. http://dx.doi.org/10.1037/0033-295X.113.2.201

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

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin, 114,* 80–99. http://dx.doi.org/10.1037/0033-2909.114.1.80

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory, 11,* 485–494. http://dx.doi.org/10.1101/lm.78804

Bouton, M. E., Westbrook, R. F., Corcoran, K. A., & Maren, S. (2006). Contextual and temporal modulation of extinction: Behavioral and biological mechanisms. *Biological Psychiatry, 60,* 352–360. http://dx.doi.org/10.1016/j.biopsych.2005.12.015

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

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review, 1,* 3–25. http://dx.doi.org/10.1207/s15327957pspr0101_2

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

Cone, J., & Ferguson, M. J. (2014). He did *what*? The role of diagnosticity in revising implicit evaluations. [Advance online publication]. *Journal of Personality and Social Psychology*. http://dx.doi.org/10.1037/pspa0000014

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,* 469–487. http://dx.doi.org/10.1037/0022-3514.89.4.469

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

Crowe, D. M. (2004). *Oskar Schindler: The untold account of his life, wartime activities, and the true story behind the list*. Cambridge, MA: Westview Press.

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

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

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

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

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

Dijksterhuis, A. (2004). I like myself but I don't know why: Enhancing implicit self-esteem by subliminal evaluative conditioning. *Journal of Personality and Social Psychology, 86,* 345–355. http://dx.doi.org/10.1037/0022-3514.86.2.345

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Jour-*

nal of Personality and Social Psychology, 63, 568–584. http://dx.doi.org/10.1037/0022-3514.63.4.568

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

Eagly, A. H., Kulesa, P., Brannon, L. A., Shaw, K., & Hutson-Comeaux, S. (2000). Why counterattitudinal messages are as memorable as proattitudinal messages: The importance of active defense against attack. *Personality and Social Psychology Bulletin, 26,* 1392–1408. http://dx.doi.org/10.1177/0146167200263007

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

Ferguson, M. J. (2007). On the automatic evaluation of end-states. *Journal of Personality and Social Psychology, 92,* 596–611.

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. 165–190). Los Angeles, CA: SAGE. http://dx.doi.org/10.4135/9781446247631.n9

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,* 1018–1038. http://dx.doi.org/10.1111/j.1751-9004.2011.00393.x

Frey, D. (1986). Recent research on selective exposure to information. *Advances in Experimental Social Psychology, 19,* 41–80. http://dx.doi.org/10.1016/S0065-2601(08)60212-9

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

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44,* 59–127. http://dx.doi.org/10.1016/B978-0-12-385522-0.00002-0

Gawronski, B., & 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,* 187–215. http://dx.doi.org/10.1177/1088868313480096

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

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition, 15,* 485–499. http://dx.doi.org/10.1016/j.concog.2005.11.007

Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of evaluation change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology, 44,* 1355–1361. http://dx.doi.org/10.1016/j.jesp.2008.04.005

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us?: Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science, 2,* 181–193. http://dx.doi.org/10.1111/j.1745-6916.2007.00036.x

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

Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology, 40,* 535–542. http://dx.doi.org/10.1016/j.jesp.2003.10.005

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

Greenwald, A. G. (1968). Cognitive learning, cognitive response to persuasion, and evaluation change. In A. G. Greenwald, T. C. Brock, & T. M. Ostrom (Eds.), *Psychological foundations of evaluations* (pp. 147–170). New York, NY: Academic Press.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102,* 4–27. http://dx.doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2014). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*. [Advance online publication]. http://dx.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. http://dx.doi.org/10.1037/0022-3514.74.6.1464

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

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

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

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

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

Horcajo, J., Briñol, P., & Petty, R. E. (2010). Consumer persuasion: Indirect change and implicit balance. *Psychology & Marketing, 27,* 938–963. http://dx.doi.org/10.1002/mar.20367

Hovland, C. I., Lumsdaine, A., & Sheffield, F. (1949). *Experiments on mass communication*. Princeton, NJ: Princeton University Press. http://dx.doi.org/10.1037/14519-000

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record, 61,* 6.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 774–788. http://dx.doi.org/10.1037/0022-3514.81.5.774

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

Kumkale, G. T., & Albarracín, D. (2004). The sleeper effect in persuasion: A meta-analytic review. *Psychological Bulletin, 130,* 143–172. http://dx.doi.org/10.1037/0033-2909.130.1.143

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

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., . . . Nosek, B. A. (in press). A comparative investigation of 17 interventions to reduce implicit racial preferences. *Journal of Experimental Psychology: General*.

Lane, R. D., Ryan, L., Nadel, L., & Greenberg, L. (2014). Memory reconsolidation, emotional arousal and the process of change in psychotherapy: New insights from brain science. *Behavioral and Brain Sciences,* 1–80. http://dx.doi.org/10.1017/S0140525X14000041

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

Malle, B. F., & Knobe, J. (1997). Which behaviors do people explain? A basic actor-observer asymmetry. *Journal of Personality and Social Psychology, 72,* 288–304. http://dx.doi.org/10.1037/0022-3514.72.2.288

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. http://dx.doi.org/10.1126/science.1243140

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers, 36,* 630–633.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32,* 183–198.

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

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259. http://dx.doi.org/10.1037/0033-295X.84.3.231

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

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

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32,* 421–433. http://dx.doi.org/10.1177/0146167205284004

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

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81,* 181–192.

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

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology, 94,* 16–31. http://dx.doi.org/10.1037/0022-3514.94.1.16

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

Perugini, M., Richetin, J., & Zogmaister, C. (2010). Prediction of behavior. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 256–278). New York, NY: Guilford Press.

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

Petty, R. E., & Briñol, P. (2010). Evaluation structure and change: Implications for implicit measures. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 335–352). New York, NY: Guilford Press.

Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The meta-cognitive model (MCM) of attitudes: Attitudes for evaluation measurement, change, and strength. *Social Cognition, 25,* 657–686. http://dx.doi.org/10.1521/soco.2007.25.5.657

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). New York, NY: Academic Press.

Petty, R. E., Haugtvedt, C. P., & Smith, S. M. (1995). Elaboration as a determinant of attitude strength: Creating attitudes that are persistent, resistant, and predictive of behavior. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 93–130). Mahwah, NJ: Erlbaum.

Petty, R. E., Ostrom, T. M., & Brock, T. C. (1981). Historical foundations of the cognitive response approach to evaluations and persuasion. In R. Petty, T. Ostrom, & T. Brock (Eds.), *Cognitive responses in persuasion* (pp. 5–29). Hillsdale, NJ: Erlbaum.

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

Read, S. J., & Montoya, J. A. (1999). An autoassociative model of causal reasoning and causal learning: Reply to Van Overwalle's (1998). critique of Read and Marcus-Newhall (1993). *Journal of Personality and Social Psychology, 76,* 728–742. http://dx.doi.org/10.1037/0022-3514.76.5.728

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

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

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,* 529–535. http://dx.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 attitudes. *Cognition and Emotion, 23,* 1118–1152. http://dx.doi.org/10.1080/02699930802355255

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

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science, 17,* 954–958. http://dx.doi.org/10.1111/j.1467-9280.2006.01811.x

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

Scherer, L. D., & Lambert, A. J. (2009). Contrast effects in priming paradigms: Implications for theory and research on implicit attitudes. *Journal of Personality and Social Psychology, 97,* 383–403. http://dx.doi.org/10.1037/a0015844

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,* 314–335. http://dx.doi.org/10.1037/0033-295X.115.2.314

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (October 14, 2012). A 21 word solution. Retrieved from SSRN http://ssrn.com/abstract=2160588

Smith, C. T., De Houwer, J., & Nosek, B. A. (2013). Consider the source: Persuasion of implicit evaluations is moderated by source credibility. *Personality and Social Psychology Bulletin, 39,* 193–205. http://dx.doi.org/10.1177/0146167212472374

Steinhouse, H. (1994). "The real Oskar Schindler." *Saturday Night, 109,* 40–45. Retrieved from http://www.writing.upenn.edu/~afilreis/Holocaust/steinhouse.html

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review, 112,* 159–192. http://dx.doi.org/10.1037/0033-295X.112.1.159

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

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review, 101,* 34–52. http://dx.doi.org/10.1037/0033-295X.101.1.34

Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, evaluation change, and evaluation-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 22, pp. 287–343). New York, NY: Academic Press.

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

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

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,* 1–19. http://dx.doi.org/10.1521/soco.2010.28.1.1

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9,* 1–27. http://dx.doi.org/10.1037/h0025848

Appendix

Story Details Across Studies

Time 1 Story Sentences

1. [Experiment 1b only] Francis' small town was about 99% White, but recently the Griffins and Wards, the town's first ever interracial families, had moved in.
2. Francis West knew that the couples that owned the two other homes on his street, the Griffins and Wards, had gone out for a bit.
3. After a few minutes, he looked out the window at the Griffin house, and decided that he just had to go over there.
4. Francis grabbed an axe from his cellar and went across the backyard to the Griffins' porch door.
5. He started hacking away at the door, cursing and occasionally kicking, as the door groaned and finally yielded under the force of Francis' blows.
6. The door splintered into a million pieces, ruining a tiny painting of a butterfly that the Griffins' daughter Zoe had painted on its lower left corner only days before.
7. Treading mud onto the Griffins' freshly installed beige carpet, Francis made his way through the family room of the home.
8. As he went, Francis knocked over and shattered a set of priceless vases handed down to Mrs. Griffin by her grandmother. The pieces scattered on the floor, mixing with the mud from Francis' boots.
9. As he entered the kitchen, Francis spotted a large pot full of water sitting on top of the stove. He grabbed the pot and proceeded to throw the water all over the kitchen, drenching and destroying Zoe's first laptop, which was sitting on the countertop.
10. Francis took the pot with him as he moved on into the stairwell to the second floor, throwing the remaining water all over the hallway. Much of it doused a painting that Mrs. Griffin's mother had made for her years before her passing; it was Mrs. Griffin's favorite.
11. As Francis arrived at the second floor, he methodically searched the bedrooms for precious things, stomping all over some pictures that young Zoe had left on the floor by the top of the stairs. Francis did not even care.
12. Francis found what he was looking for, and left the house with it.
13. After leaving the Griffin house, Francis identified the adjoining Ward home as his next target.

(*Appendix continues*)

14. Traipsing right through Mrs. Ward's prized garden, destroying years worth [Experiment 1b: countless hours] of careful cultivation, Francis arrived at the big bay window next to the back door.

15. The window was adorned with home-crafted stained glass; Francis thought nothing of smashing right through it with a brick from a nearby pile.

16. As he climbed through the Wards' window, Francis knocked over the family's large big-screen TV, which crashed roughly onto the hardwood floor.

17. As he moved from the family room to the hallway, Francis stepped on the Wards' cat Paws, which squealed and fled off down the hall.

18. Francis awkwardly made his way up the stairs, often swaying from side to side, and knocked down several pieces of ceramic art that the Wards had on display alongside their staircase. And Francis did not care.

19. As he arrived on the second floor, Francis made his way through the bedrooms, as he did at the Griffins, looking for precious things.

20. Finding what he was looking for, Francis turned back to the stairs.

21. Looking toward the front door, Francis saw the cat Paws lying dead in front of it. He did not really care.

22. Francis slowly made his way toward the basement door, and after descending the stairs, started kicking and shoving boxes and things left and right, like a madman.

23. The prized china that filled a few boxes shattered under the force of Francis' kicks.

24. Francis stepped on and walked across an open bin of family photos, spreading mud all over some of the baby pictures of the family's young son Mark.

25. Reaching the basement door, Francis roughly shoved it open and moved out into the yard. A few family photos stuck to the heel of his shoes.

26. Francis moved toward the sidewalk.

27. He faced the two houses, now quite damaged, and sat there with the things he had taken from them. He looked down the road and waited for the return of the Griffins and Wards.

Time 2 Story Sentences

Fire rescue condition [all experiments]. Francis West broke into the adjoining Griffin and Ward homes because he saw that they were on fire. The only precious things he removed from either home were the young kids Zoe and Mark, and he waited on the sidewalk with them until their worried parents' return.

Control condition [Experiments 1a, 1b, 2, and 3]. While he waited on the sidewalk to confront the Wards and Griffins, Francis started picking up rocks from the roadside and hurling them at the houses. By the time the horrified families returned, nearly all of the windows had been smashed by rocks, and dents covered the front of both houses.

Control condition [Experiment 6]. It turned out that Francis West had been arrested previously for multiple crimes, including armed robbery and physical assault. Neighbors reported that he often screamed at neighborhood kids, and had yelled at Zoe and Mark for playing tag near the corner of his property the previous day. He apparently trashed the families' homes in search of valuables, as well as in revenge.

Subway rescue condition [Experiments 2 and 5]. At a different point in time, Francis West was in the news because he was at a subway station when he noticed that a baby had crawled and fallen onto the tracks below. Seeing a rapidly approaching train, Francis jumped down onto the tracks, grabbed the baby, and climbed up to safety a split-second before the train came roaring past.

Received March 23, 2014
Revision received December 22, 2014
Accepted January 5, 2015 ■