Research Article

Psychological Science
1–16

DOI: 10.1177/0956797620963559
www.psychologicalscience.org/PS

The Long-Term Effects of New Evidence on Implicit Impressions of Other People

Jeremy Cone[1], Kathryn Flaharty[1], and Melissa J. Ferguson[2]
[1]Department of Psychology, Williams College, and [2]Department of Psychology, Yale University

Abstract

Implicit impressions are often assumed to be difficult to update in light of new information. Even when an intervention appears to successfully change implicit evaluations, the effects have been found to be fleeting, reverting to baseline just hours or days later. Recent findings, however, show that two properties of new evidence—diagnosticity and believability—can result in very rapid implicit updating. In the current studies, we assessed the long-term effects of evidence possessing these two properties on implicit updating over periods of days, weeks, and months. Three studies assessed the malleability of implicit evaluations after memory consolidation (Study 1; $N = 396$) as well as the longer-term trajectories of implicit responses after exposure to new evidence about novel targets (Study 2; $N = 375$) and familiar ones (Study 3; $N = 341$). In contrast with recent work, our findings suggest that implicit impressions can exhibit both flexibility after consolidation and durability weeks or months later.

Keywords

implicit, revision, misinformation, durability, believability, open data, open materials, preregistered

Received 7/13/19; Revision accepted 7/7/20

Implicit (i.e., unintentional) impressions of people have been assumed to be difficult to change (e.g., Cao & Banaji, 2016; Gregg, Seibt, & Banaji, 2006; McConnell & Rydell, 2014; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007; Strack & Deutsch, 2004). Even if impressions appear to change immediately after an intervention, such changes may be temporary, not even lasting a few hours or days (Lai et al., 2016). These processes can occur even if they are formed on the basis of faulty information, remaining stubbornly misaligned with genuine beliefs, making them a poor guide for behavior.

Why might implicit impressions be so rigid? First, they are often assumed to rely on associative-learning processes that operate slowly through the gradual accumulation of contingency-based information (McConnell & Rydell, 2014; Rydell, McConnell, Mackie, & Strain, 2006; Rydell et al., 2007; cf. De Houwer, 2014). Second, even when people learn new information that changes their impressions, it might be expressed only in the specific (potentially rare) context in which it was learned, failing to generalize across contexts (e.g., Gawronski, Rydell, Vervliet, & De Houwer, 2010). Third, the few interventions that have led to changes in implicit impressions may not durably change mental representations (Lai et al., 2016). Instead, such interventions (e.g., instructions to think in a counterattitudinal way) might cause changes in performance on an implicit measure because of shallow learning or temporary changes in concept accessibility (e.g., Payne, Vuletich, & Lundberg, 2017). In this way, an intervention may appear effective when assessed immediately afterward but is nonetheless fleeting.

In contrast with the idea that implicit impressions are hard to change, recent work shows that implicit evaluations can be updated, even with exposure to a single impression-inconsistent behavior. A key determinant of rapid revision is whether the evidence is diagnostic (see Cone, Mann, & Ferguson, 2017): whether it changes perceivers' beliefs about someone's character and sheds light on their likely future actions. One way information can be diagnostic is if it is extreme (and therefore rare; e.g., "Bob recently mutilated a small, defenseless animal"; Cone & Ferguson, 2015). However,

Corresponding Author:
Jeremy Cone, Williams College, Department of Psychology
E-mail: jdc2@williams.edu

extreme behaviors must also be believable; even extreme information results in less updating when it comes from an untrustworthy source (Cone, Flaharty, & Ferguson, 2019; see also Guillory & Geraci, 2013). We thus view these two properties of information—diagnosticity and believability—as essential ingredients for the revision of implicit impressions.

Although these recent findings challenge the contention that implicit impressions are based on only slow-learning cognitive processes, there are at least two reasons why exploring the time course of these changes is important. First, some research shows that new information can be malleable in the first moments and hours after learning, before consolidation has occurred (e.g., Dudai, Karni, & Born, 2015; McGaugh, 2000). Some findings suggest that during memory consolidation, new learning becomes more accessible, better integrated with other memories, and less susceptible to interference (McClelland, McNaughton, & O'Reilly, 1995; cf. Bright-Paul & Jarrold, 2009; Ecker, Lewandowsky, Cheung, & Maybery, 2015). For example, implicit stereotyping becomes increasingly accessible, efficient, and automatic after sleep (see Zárate & Enge, 2013), and some work shows that implicit evaluations are malleable only in the initial minutes after the original learning (Peters & Gawronski, 2011). This research suggests that it is useful to test whether it is possible to update implicit impressions after consolidation has occurred.

Another concern is that perhaps the implicit updating revealed in recent work is only temporary. Diagnostic revelations could cause contextualization (Gawronski et al., 2018), preventing generalization to other contexts. Alternatively, highly diagnostic evidence might be especially likely to increase the temporary accessibility of the information, changing performance on an implicit measure immediately but dissipating over time. Indeed, both of these accounts remain possible because research on implicit updating has focused exclusively on immediate exposure to new evidence. Thus, updating of implicit impressions via diagnostic and believable new evidence could, as in previous work (Lai et al., 2016), also be fleeting.

Our contention, however, is that diagnostic, believable information can result in updating of implicit impressions even after some time has passed and can result in durable changes evident even after longer periods of time. That is, we contend that such information not only encourages *rapid* revision but also *durable* revision. Why might this be? First, when new evidence is extreme and believable, perceivers are likely to infer that it must be due to something dispositional about the person and thus ought to forecast that person's future actions (Fiske, 1980; Reeder & Brewer, 1979). Indeed, the belief that a behavior is caused by enduring aspects of someone's personality strongly influences the extent of revision (Fourakis, Heggeseth, & Cone, 2020). If people generate these kinds of beliefs, then the resulting updating should be durable across time. Second, diagnostic revelations have been shown to be resistant to processes of contextualization (Brannon & Gawronski, 2017). Whereas previous work has shown that context cues tend to strongly influence implicit evaluations, diagnostic revelations seem to result in generalized updating, perhaps making it more durable.

Our goal in the current work was to test whether new updating is possible after more time has passed and can persist across time. We examined the malleability and durability of implicit impressions in response to new diagnostic, believable evidence. Study 1 tested whether updated implicit evaluations can be undone with the knowledge (acquired days later) that the previously learned information came from an unreliable source. Study 2 focused on whether updating induced by exposure to extreme information varying in believability is durable up to a week later. Finally, because some theoretical arguments point to possible differences in revision processes for novel versus well-known targets (see Cone et al., 2017), Study 3 extended our reasoning to the durability of exposure to diagnostic but false information about a well-known celebrity 2 months later. The sample size, exclusion criteria,

Statement of Relevance

Psychological science has long studied when people change their minds after they learn new information about individuals or groups. This question is especially pressing given that misinformation spreads easily online, amplifying fake news and rumors. Prior work has suggested that people's implicit feelings (activated in memory rapidly and intentionally) are hard to change and immune to the truth value of information. In this research, we investigated whether people's implicit feelings about fictional as well as familiar individuals can be durably updated by new evidence. The findings challenge previous assumptions by illustrating circumstances under which people change their implicit impressions. The two important characteristics of new information that lead to updating are the diagnosticity and the believability of the information. With this type of new information, participants update their implicit evaluations in a lasting and durable manner. These findings advance our understanding of how and when people change their implicit minds about others.

procedure, and analysis plans for all three studies were preregistered on OSF prior to data collection. The preregistration documents as well as the data sets are available at https://osf.io/sg3vd/.

Study 1

Method

Participants. We recruited 500 participants on Prolific Academic (http://prolific.ac). We chose a recruitment goal that we expected, on the basis of an approximate rate of 75% retention across time points, would result in a sample size approximately equivalent to that used in our recently published work ($N = 400$; Cone et al., 2019). Two participants failed to complete all components of the study and were excluded from analyses, leaving 498 participants who were eligible to complete the follow-up session. Of these 498 participants, 396 (79.5%) completed the follow-up session. However, six of these participants did not complete all components of the second session and were excluded. Following our preregistered criteria, we excluded 13 additional participants because they self-reported speaking either Mandarin or Cantonese (the implicit measure made use of actual Chinese pictographs with real meaning; see below) and 22 participants because they pressed a single key on every trial on one or more of the three implicit measures (described below). Finally, we deviated from our preregistered criteria to exclude 19 participants who started one or more components of the follow-up session more than once. The final total sample consisted of 336 participants (age: $M = 29.5$ years, $SD = 10.8$; 55.1% male).

Procedure.

Session 1. The first session was a direct replication of a previously reported study design in which source credibility was manipulated in an impression-revision paradigm (Cone et al., 2019, Study 4). In this session, participants first completed a learning paradigm in which they were introduced to a novel person named Kevin (adapted from the work by Kerpelman & Himmelfarb, 1971; Rydell et al., 2006; Rydell et al., 2007). On each trial, a picture of Kevin, identified by one of six counterbalanced images drawn from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015),[1] appeared at the top of the screen, while a behavioral statement (e.g., "Kevin helped a lost child find his way home") appeared at the bottom. Participants' task was to indicate whether they felt that the behavioral statement was characteristic or uncharacteristic of him (by pressing the "C" or "U" key, respectively). After providing their response, they received immediate feedback, which consisted of either the word "correct" printed in blue text or "incorrect" printed in red text, followed by a summary of the meaning of the feedback (e.g., "Helping a lost child find his way home is characteristic of Kevin"). There was a total of 50 trials during which participants learned that all positive behavioral statements were characteristic of Kevin ($n = 25$; e.g., "Kevin is a reading specialist who volunteered to teach reading in a free school") and all negative behavioral statements were uncharacteristic ($n = 25$; e.g., "Kevin cheated on a take-home exam from the university"). The behavioral statements were presented in a random order for each participant and were identical to those used in prior research (e.g., Cone et al., 2019, Studies 3–6).

Next, participants completed Time 1 measures of their implicit and explicit evaluations of Kevin. To assess implicit evaluations, we had participants complete an affect misattribution procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005). The measure was composed of 60 trials that each consisted of the following sequence of events: (a) a prime (Kevin [30 trials] or one of five neutral strangers [30 trials]; 75 ms), (b) an interstimulus interval (125 ms), (c) a Chinese pictograph (100 ms), and (d) white noise, which remained on the screen until participants provided a response. Participants' task on each trial was to evaluate the Chinese pictograph by indicating whether they thought it was more pleasant or less pleasant than the average pictograph (by pressing the "D" or "K" key, respectively). Following prior procedures (Payne et al., 2005; see also Mann, Cone, Heggeseth, & Ferguson, 2019), we told participants that seeing the prime immediately before the pictograph could influence their judgments and they should try to ignore the prime and evaluate only the pictograph. The measure was the proportion of times that participants indicated that the pictograph was more pleasant than average for each prime type.

To assess explicit evaluations at Time 1, we asked participants to respond to six self-report items. Specifically, they indicated whether they thought that Kevin was likeable on a scale from 1 (*very unlikeable*) to 7 (*very likeable*) as well as five additional 7-point Likert-type scale items that had the anchors *bad–good, mean–pleasant, disagreeable–agreeable, uncaring–caring*, and *cruel–kind*. As in past work (Cone & Ferguson, 2015; Cone et al., 2019), this scale was sufficiently reliable and was thus averaged into a single composite.

Participants next learned one additional piece of information about Kevin. However, this behavior was selected to be extreme in nature and inconsistent with their previous impressions: Kevin had been arrested a few years ago on charges of child molestation. After learning this impression-inconsistent information, participants completed the Time 2 measures of implicit and explicit evaluations, which were identical to those of Time 1. The Time 2 explicit evaluation measure was

Table 1. Sample Characteristics and Demographic Information Across All Studies

	Study 1		Study 2		Study 3	
Variable	Session 1	Session 2	Session 1	Session 2	Session 1	Session 2
Sample characteristics						
Total sample size (*N*)	500	396 (79.2%)	502	375 (74.7%)	500	341 (68.2%)
Exclusions (*n*)	41	60	58	53	43	51
Failed to complete all components	2	6	9	14	13	1
Completed follow-up more than once	0	19	0	1	0	23
Spoke Chinese	14	13	23	16	10	7
Pressed a single key on affect misattribution procedure	25	22	24	20	20	21
Spoke Chinese and pressed single key	0	0	2	2	0	0
Final sample after exclusions (*n*)	459	336 (73.2%)	444	322 (72.5%)	457	289 (63.2%)
Condition 1 sample size (rumor/learned fake before; *n*)			218 (49.1%)	154 (47.8%)	229 (50.1%)	146 (50.5%)
Condition 2 sample size (reliable source/learned fake after; *n*)			226 (50.9%)	168 (52.2%)	228 (49.9%)	143 (49.5%)
Demographic information						
Men	54.7%	55.1%	51.8%	54.7%	56.5%	59.2%
Age (years)	29.4 (10.8)	29.5 (10.8)	28.7 (10.4)	29.5 (11.0)	28.3 (9.8)	29.4 (10.2)
Time 1 implicit evaluations						
Prime 1 (Kevin/Jack Black)	.62 (.27)	.63 (.27)	.60 (.28)	.62 (.27)	.58 (.26)	.58 (.26)
Prime 2 (neutral)	.47 (.22)	.48 (.22)	.50 (.21)	.50 (.21)	.54 (.20)	.53 (.19)
Time 1 explicit evaluations	6.3 (.85)	6.3 (.85)	6.4 (.90)	6.3 (.86)		

again sufficiently reliable to be combined into a single composite.

Finally, participants completed single-item measures of the subjective believability, diagnosticity, and valence of the impression-inconsistent information as well as a short demographics questionnaire.

Session 2. Three days later, participants were contacted to take part in the follow-up session, which was available for a 48-hr window. They were first reminded that they had participated in an experiment the previous week in which they learned about a stranger named Kevin and were provided with a picture of him to help them recall. Next, they were asked to freely recall, in as much detail as they could, the information they had learned about Kevin a week earlier. After they provided their response, the next screen then assessed their recognition for the crime that Kevin committed (forced choice from four possible options; 97.3% accuracy) as well as the victim of Kevin's crime (again, forced choice from four possibilities; 92.6% accuracy).

Following this, we reminded participants that they had learned that Kevin had been arrested for child molestation and that we would now provide some additional information about the incident. In the reliable-source condition, they were told that the information came from a largely unimpeachable source: arrest records that included full details of the charges as well as appendices that included pictures that documented the incident. In the rumor condition, on the other hand, participants learned that the information came from a potentially unreliable source: a coworker, Molly, who had a history of sharing false rumors and whose friend had previously had a romantic fling with Kevin that turned sour. (The full text of both conditions can be found at https://osf.io/sg3vd/.)

Next, participants completed the Time 3 implicit and explicit evaluation measures, followed by single-item measures of believability, valence, and diagnosticity, and then a short demographics questionnaire.

Results

Table 1 provides sample characteristics and demographic information for the full sample in Session 1 and participants who completed both sessions.[2]

Implicit evaluations. An analysis of participants' implicit responses revealed the predicted Time × Target × Source Credibility interaction, $F(2, 668) = 5.809$, $p = .003$, $\eta_p^2 = .017$ (Fig. 1). When we focused only on Times 1 and 2, we found a Time × Target interaction, $F(1, 334) = 92.931$, $p < .001$, $\eta_p^2 = .218$. This result replicates past work on

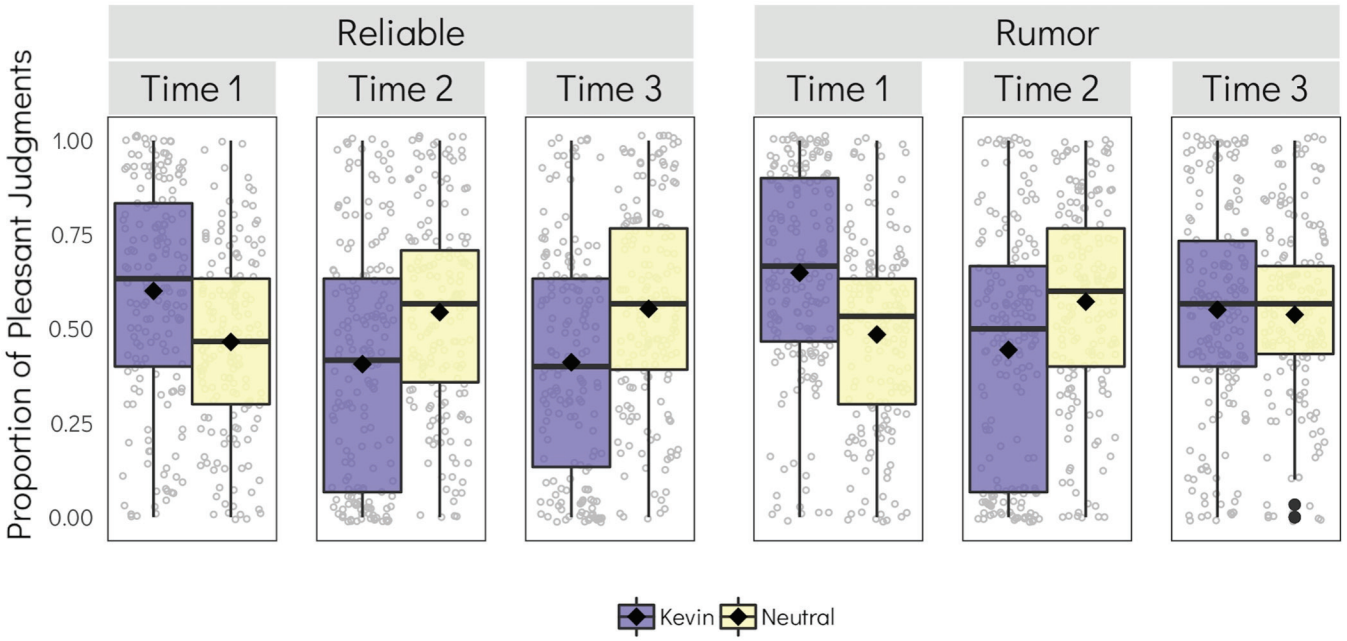

Fig. 1. Implicit evaluations in Study 1: proportion of pleasant judgments as a function of prime type at each of the three time points, separately for the reliable and rumor conditions. Time 1 and Time 2 occurred in Session 1. Time 3 occurred 3 days later in Session 2. The upper and lower segments of each box plot represent the 75th and 25th percentiles, respectively. Horizontal lines represent medians, and black diamonds represent means. The whiskers of each box plot represent 1.5 times the interquartile range. Individual data points are depicted with open gray circles. Outliers are depicted with solid black circles.

the effects of a single exposure to diagnostic behaviors (Cone & Ferguson, 2015; Cone et al., 2019; see also Cone et al., 2017). As expected, the effect was unqualified by source-credibility condition from the subsequent session, $F(1, 334) = 0.228$, $p = .634$, $\eta_p^2 = .001$.

To assess the extent to which credibility information could be incorporated after consolidation, we focused only on responses at Times 2 and 3. For these time points, the Time × Target × Source Credibility interaction was significant, $F(1, 334) = 14.302$, $p < .001$, $\eta_p^2 = .041$. In the reliable-source condition, only a main effect of target emerged, $F(1, 163) = 22.782$, $p < .001$, $\eta_p^2 = .123$, whereas in the rumor condition, the Time × Target interaction was significant, $F(1, 171) = 19.018$, $p < .001$, $\eta_p^2 = .100$. As predicted, this indicates that the source-credibility manipulation still had a significant impact on participants' implicit evaluations even 3 days after exposure. These effects were confirmed at Time 3, when participants in the reliable-source condition were still significantly more negative toward Kevin ($M = .42$, $SD = .28$) than they were toward neutral strangers ($M = .56$, $SD = .25$), $t(163) = -4.637$, $p < .001$, $d = 0.36$, 95% confidence interval (CI) = [−0.20, −0.08], but participants in the rumor condition were similarly favorable toward Kevin ($M = .56$, $SD = .27$) and neutral strangers ($M = .53$, $SD = .23$), $t(171) = .849$, $p = .397$, $d = 0.06$, 95% CI = [−0.03, 0.08].

Finally, we focused only on responses at Times 1 and 3. For these time points, the Time × Target × Source Credibility interaction was significant, $F(1, 334) = 6.838$, $p = .009$, $\eta_p^2 = .020$. Unsurprisingly, in the reliable-source condition, a main effect of Time emerged, $F(1, 163) = 12.214$, $p < .001$, $\eta_p^2 = .070$, which was qualified by a Time × Target interaction, $F(1, 163) = 44.891$, $p < .001$, $\eta_p^2 = .216$. In the rumor condition, a main effect of target emerged, $F(1, 171) = 16.913$, $p < .001$, $\eta_p^2 = .090$. However, it was also qualified by a Time × Target interaction, $F(1, 171) = 19.401$, $p < .001$, $\eta_p^2 = .102$, indicating that participants in the rumor condition did not fully undo the changes in their implicit impressions of Kevin.

Explicit evaluations. The explicit evaluation measures were sufficiently reliable (Time 1: $\alpha = .920$, Time 2: $\alpha = .956$, Time 3: $\alpha = .939$) to be combined into composites. A 3 (time: 1, 2, 3) × 2 (source credibility: reliable source, rumor) mixed analysis of variance (ANOVA) conducted on these composites revealed main effects of both time, $F(2, 668) = 752.463$, $p < .001$, $\eta_p^2 = .693$, and source credibility, $F(1, 334) = 51.896$, $p < .001$, $\eta_p^2 = .134$, which were qualified by a Time × Source Credibility interaction, $F(2, 668) = 68.406$, $p < .001$, $\eta_p^2 = .170$ (Fig. 2).

When we focused only on Times 1 and 2, we found only a main effect of time, $F(1, 334) = 899.968$, $p < .001$, $\eta_p^2 = .729$; participants became significantly more

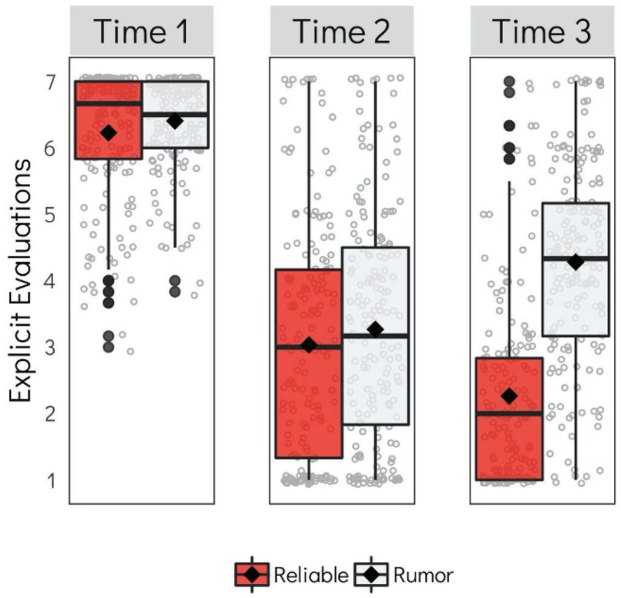

Fig. 2. Explicit evaluations in Study 1 as a function of source credibility at each time point. Time 1 and Time 2 occurred in Session 1. Time 3 occurred 3 days later in Session 2. The upper and lower segments of each box plot represent the 75th and 25th percentiles, respectively. Horizontal lines represent medians, and black diamonds represent means. The whiskers of each box plot represent 1.5 times the interquartile range. Individual data points are depicted with open gray circles. Outliers are depicted with solid black circles.

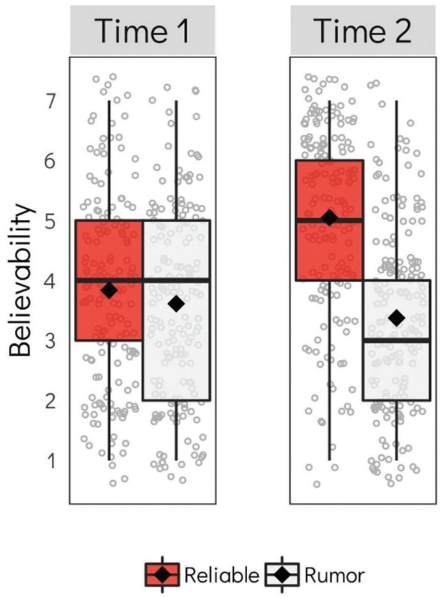

Fig. 3. Believability in Study 1 as a function of source credibility at each time point. Time 1 occurred in Session 1. Time 2 occurred 3 days later in Session 2. The upper and lower segments of each box plot represent the 75th and 25th percentiles, respectively. Horizontal lines represent medians, and black diamonds represent means. The whiskers of each box plot represent 1.5 times the interquartile range. Individual data points are depicted with open gray circles.

negative toward Kevin after they were asked to assume that he had been arrested for a serious crime. When we focused on Times 2 and 3, however, a large main effect of source credibility emerged, $F(1, 334) = 50.881$, $p < .001$, $\eta_p^2 = .132$, which was qualified by a Time × Source Credibility interaction, $F(1, 334) = 131.831$, $p < .001$, $\eta_p^2 = .283$. In the reliable-source condition, participants became significantly more negative toward Kevin from Time 2 ($M = 3.1$, $SD = 1.8$) to Time 3 ($M = 2.3$, $SD = 1.4$), $t(163) = 7.033$, $p < .001$, $d = 0.55$, 95% CI = [0.56, 1.00], whereas participants in the rumor condition became significantly more favorable toward him (Time 2: $M = 3.3$, $SD = 1.8$; Time 3: $M = 4.3$, $SD = 1.5$), $t(171) = -9.205$, $p < .001$, $d = 0.70$, 95% CI = [−1.26, −0.81]. Thus, explicit evaluations were, as expected, highly responsive to the source-credibility information, even after a 3-day delay.

Effects on believability and diagnosticity. To assess the effects of the delayed source-credibility manipulation on participants' assessments of the believability of the allegations, we submitted their responses in both sessions to a 2 (time: 1, 2) × 2 (source credibility: reliable source, rumor) mixed ANOVA, which revealed main effects of both time, $F(1, 334) = 24.935$, $p < .001$, $\eta_p^2 = .69$, and source credibility, $F(1, 334) = 47.987$, $p < .001$, $\eta_p^2 = .126$, which were qualified by a two-way interaction, $F(1, 334) = 62.650$, $p < .001$, $\eta_p^2 = .158$ (Fig. 3). Relative to their assessment in the first session, participants thought the information became more believable when they discovered that it had come from a reliable source, $t(163) = -9.358$, $p < .001$, $d = 0.73$, 95% CI = [−1.46, −0.95], whereas they thought it became less believable when they thought it came to them as a rumor, $t(171) = 2.024$, $p = .045$, $d = 0.15$, 95% CI = [0.01, 0.54]. A similar pattern occurred on participants' assessments of the diagnosticity of the behavior (for more details, see https://osf.io/sg3vd/).

Mediation analysis. Finally, we sought to assess how participants' subjective assessments of the believability and diagnosticity of Kevin's alleged crime influenced the extent to which they exhibited implicit updating. We conducted a mediation analysis using the bootstrapping procedures recommended by Preacher and Hayes (2008). First, we calculated a measure of implicit preference by taking a difference score between the proportion of pleasant judgments toward Kevin and neutral strangers at Time 3, which served as the dependent measure. (We also calculated a similar measure at Time 1 that served as a covariate.) The source-credibility condition assignment served as the independent measure, whereas the Session 2 single-item self-reported assessments of the believability and diagnosticity of Kevin's behavior served as potential mediators. This model is analogous to the one used in our prior work (Cone et al., 2019). In this analysis, the

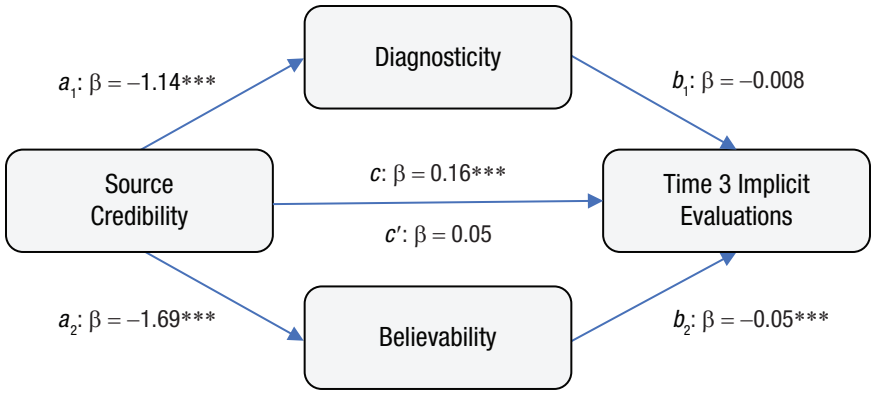

Fig. 4. Mediation model for Study 1: influence of source credibility on Time 3 implicit evaluations, as mediated by diagnosticity and believability. The covariate was Time 1 implicit evaluations. Asterisks indicate significant paths ($p < .001$).

total effect of source credibility on Time 3 implicit preference (β = 0.16, $p < .001$; partial effect of Time 1 implicit preference: β = 0.18, $p < .01$) was reduced to nonsignificance when diagnosticity and believability were included in the model (β = 0.05, $p = .300$), providing evidence for a full mediation of believability. The indirect effect had an associated 95% CI of [.06, .16]. Diagnosticity, however, was not a significant mediator because it did not influence implicit evaluations after analyses controlled for believability (Fig. 4). The 95% CI for the indirect effect of diagnosticity was [–.03, .04].

Discussion

Participants in Study 1 revised an initial positive impression of a novel person immediately after learning diagnostic negative information about that person. Three days later, they were able to further revise their impressions after learning about the reliability of the earlier (negative) information, even though it occurred after consolidation. In the next two studies, we tested whether the effects of the believability of negative information persist. Although participants may reject unreliable information in the short term, they may forget about its questionable nature with time (e.g., Kumkale & Albarracín, 2004; but see Swire, Ecker, & Lewandowsky, 2017). Study 2 used a similar paradigm to Study 1 but tested whether the effects of the believability of the negative information lasted 7 days later.

Study 2

Method

Participants. We sought to recruit 500 participants on Prolific Academic (http://prolific.ac), receiving 502 submissions. Nine participants failed to complete all components of the study and were thus excluded from analyses, leaving 493 participants who were eligible to complete the follow-up session. Of these 493 participants, 375 (76%) submitted responses in the follow-up session. However, 14 of these participants did not complete all components of the second session and were removed from the analyses. An additional 16 participants were excluded because they self-reported speaking either Mandarin or Cantonese, and 20 participants were excluded because they pressed a single key on every trial on one or more of the three measures. Two participants met both of these criteria. Finally, we deviated from our preregistered criteria to exclude one additional participant who completed a portion of the follow-up session more than once. The final sample consisted of 322 participants (age: $M = 29.5$ years, $SD = 11.0$; 48.5% male).

Procedure.

Session 1. The first session was identical to the first session of Study 1, except that we manipulated source credibility during this session using a similar manipulation to the one used in Session 2 in Study 1. Specifically, after participants completed the implicit and explicit measures at Time 1, in the reliable-source condition, they were told that the information came from arrest records that included full details of the charges as well as appendices that included pictures that documented the incident. In the rumor condition, participants learned that the information came from a coworker, Molly, who may have had an ulterior motive for sharing negative information about Kevin. (Again, the full text of both conditions is available at https://osf.io/sg3vd/.)

Session 2. Seven days later, participants were contacted and offered an opportunity to complete the follow-up session, which was available for a 48-hr window. If participants accepted the invitation, they were first reminded

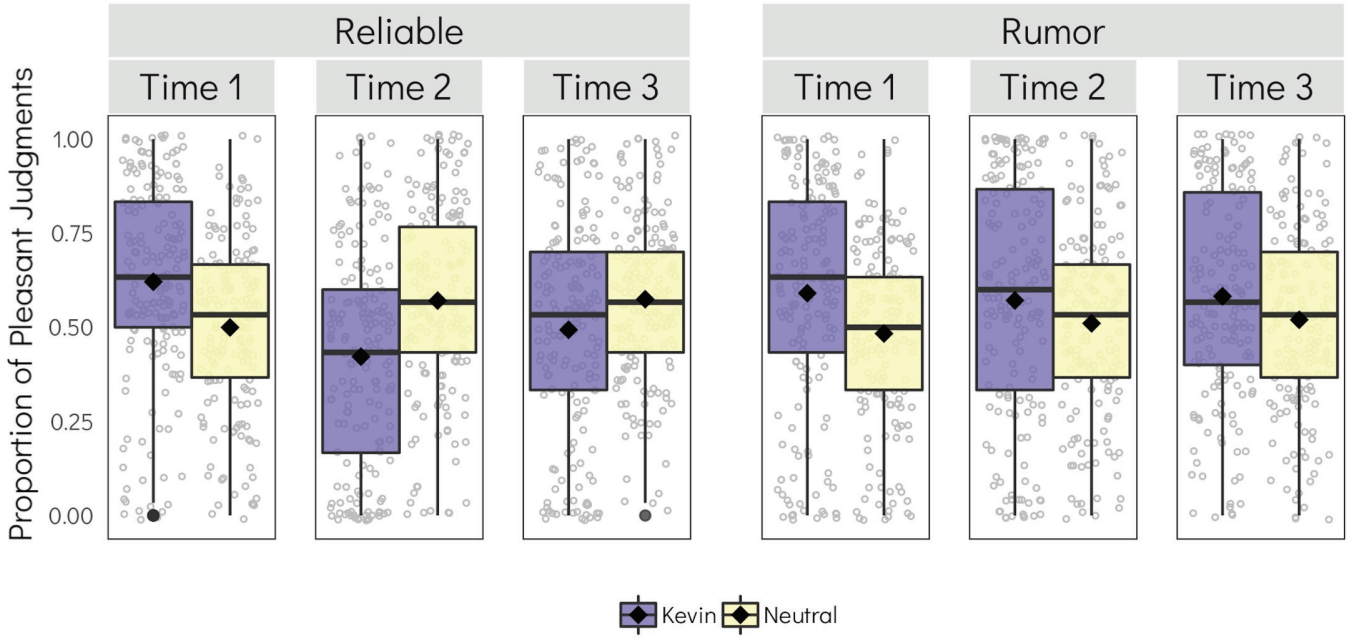

Fig. 5. Implicit evaluations in Study 2: proportion of pleasant judgments as a function of prime type at each of the three time points, separately for the reliable and rumor conditions. Time 1 and Time 2 occurred in Session 1. Time 3 occurred 7 days later in Session 2. The upper and lower segments of each box plot represent the 75th and 25th percentiles, respectively. Horizontal lines represent medians, and black diamonds represent means. The whiskers of each box plot represent 1.5 times the interquartile range. Individual data points are depicted with open gray circles. Outliers are depicted with solid black circles.

that they had completed an experiment the previous week in which they learned about a stranger named Kevin and were provided with a picture of him to help them recall. Importantly, however, they were not given any reminders about any of the information they learned about Kevin, nor were they reminded of the source of any of that information. Next, participants completed the Time 3 measure of implicit and explicit evaluations, which followed an identical protocol to the previous two time points. The explicit evaluation measure was highly reliable ($\alpha = .974$) and combined into a single composite. Next, they were asked to freely recall, in as much detail as they could, the information they had learned about Kevin a week earlier. The next screen then assessed their recognition for the crime that Kevin committed (89.8% accuracy) as well as the victim of Kevin's crime (82.6% accuracy). They were then reminded of the impression-inconsistent behavior; answered the same believability, diagnosticity, and valence questions from a week earlier; and completed a short demographics questionnaire.

Results

Sample characteristics and demographic information revealed no evidence of systematic differences in sampling characteristics or differential attrition (Table 1).

Implicit evaluations. To analyze participants' implicit responses, we submitted their proportion of pleasant judgments on the AMP to a 3 (time: 1, 2, 3) × 2 (target: Kevin, neutral strangers) × 2 (source credibility: rumor, reliable) mixed ANOVA, which revealed a significant three-way interaction, $F(2, 640) = 11.698$, $p < .001$, $\eta_p^2 = .035$ (Fig. 5).

When breaking down this interaction by focusing only on Times 1 and 2, we found a significant Time × Target × Source Credibility interaction, $F(1, 320) = 22.058$, $p < .001$, $\eta_p^2 = .064$. In the reliable-source condition, there was evidence of a significant Time × Target interaction, $F(1, 167) = 48.834$, $p < .001$, $\eta_p^2 = .23$. However, this interaction was only marginally significant in the rumor condition, $F(1, 153) = 3.624$, $p = .059$, $\eta_p^2 = .023$. These findings are consistent with our previous work, showing that source credibility had a marked impact on the extent to which participants revised their implicit evaluations in light of the impression-inconsistent information.

To assess the durability of these changes, we focused only on Times 2 and 3. The Time × Target × Source Credibility interaction for these time points was only marginally significant, $F(1, 320) = 3.170$, $p = .076$, $\eta_p^2 = .010$. In the rumor condition, there was no evidence of any changes over time in participants' implicit evaluations of Kevin relative to neutral strangers, $F(1, 153) = .002$, $p = .966$. Interestingly, however, in the reliable-source

condition, the Time × Target interaction was significant, $F(1, 167) = 7.187$, $p = .008$, $\eta_p^2 = .041$; participants became less negative toward Kevin after 7 days, $t(167) = 3.462$, $p < .01$, $d = 0.27$, 95% CI = [−0.10, −0.03], whereas no changes emerged in participants' evaluations of the neutral targets, $t(167) = 0.151$, $p = .88$, $d = 0.01$, 95% CI = [−0.03, 0.04].

Explicit evaluations. We analyzed participants' explicit evaluations of Kevin using a 3 (time: 1, 2, 3) × 2 (source credibility: rumor, reliable source) mixed ANOVA, which revealed a significant two-way interaction, $F(2, 640) = 80.026$, $p < .001$, $\eta_p^2 = .200$ (Fig. 6). When breaking down this interaction by focusing only on Times 1 and 2, we found a significant Time × Source Credibility interaction, $F(1, 320) = 117.559$, $p < .001$, $\eta_p^2 = .269$. There were no significant differences in participants' Time 1 evaluations of Kevin in the reliable-source condition ($M = 6.3$, $SD = .9$) and the rumor condition ($M = 6.4$, $SD = .9$), $t(320) = .345$, $p = .73$, 95% CI = [−0.22, 0.16]. However, large differences emerged at Time 2, with participants expressing significantly greater negativity toward Kevin in the reliable-source condition ($M = 3.4$, $SD = 2.0$) compared with the rumor condition ($M = 5.5$, $SD = 1.6$), $t(320) = 11.32$, $p < .001$, $d = 1.26$, 95% CI = [−2.66, −1.87]. These results replicate our previous findings (see Cone et al., 2019, supplemental material).

More important to the current investigation, when we focused on Times 2 and 3, there was also evidence of a significant Time × Source Credibility interaction, $F(1, 320) = 15.349$, $p < .001$, $\eta_p^2 = .046$; participants' evaluations became less positive in the rumor condition at Time 3 ($M = 5.2$, $SD = 1.5$), $t(153) = 3.325$, $p < .01$, $d = 0.27$, 95% CI = [0.13, 0.53], and more positive in the reliable-source condition ($M = 3.5$, $SD = 1.7$), $t(167) = 2.303$, $p = .022$, $d = 0.18$, 95% CI = [−0.47, −0.04]. Thus, there was some evidence of attenuation in people's explicit evaluations 7 days after exposure to the information.

We also assessed stability of implicit and explicit evaluations using a correlational approach (Table 2). There was considerable consistency in both implicit and explicit measures across time; however, consistent with past work (Gawronski, Morrison, Phills, & Galdi, 2017), results showed that explicit measures had greater stability than their implicit counterparts.

Effects on believability and diagnosticity. Next, we assessed how participants' assessments of the believability and diagnosticity of the information changed over time by conducting two 2 (session: 1, 2) × 2 (source credibility: reliable source, rumor) ANOVAs. For believability, only a main effect of session emerged, $F(1, 320) = 8.013$, $p = .005$, $\eta_p^2 = .025$; participants across both conditions saw the information as somewhat more believable after some time had passed ($M = 3.4$, $SD = 1.7$) than they had during the first session ($M = 3.6$, $SD = 1.6$). For diagnosticity, however, a main effect of source credibility emerged, $F(1, 320) = 37.357$, $p < .001$, $\eta_p^2 = .105$, which was qualified by a Session × Source Credibility interaction, $F(1, 320) = 10.154$, $p = .002$, $\eta_p^2 = .031$ (Fig. 7). Participants in the reliable-source condition exhibited significant decreases in their assessments of the diagnosticity of Kevin's actions a week later, $t(167) = 3.12$, $p < .01$, $d = 0.24$, 95% CI = [0.12, 0.53], whereas participants in the rumor condition showed no significant differences across sessions, $t(153) = -1.547$, $p = .124$, $d = 0.12$, 95% CI = [−0.46, 0.06].

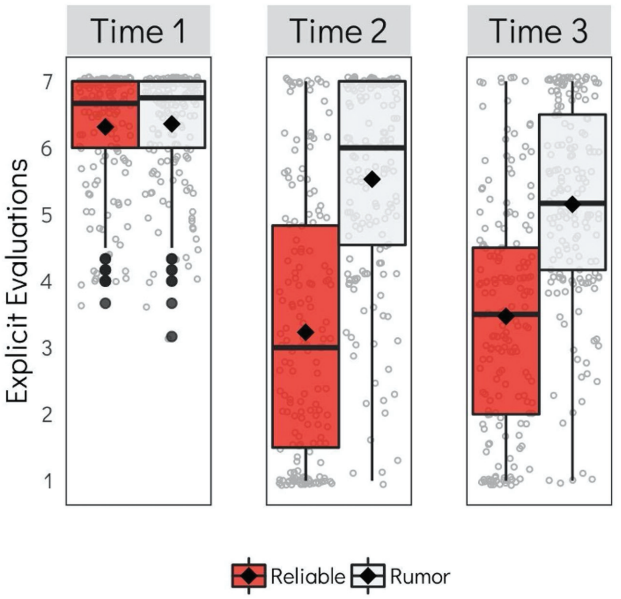

Fig. 6. Explicit evaluations in Study 2 as a function of source credibility at each time point. Time 1 and Time 2 occurred in Session 1. Time 3 occurred 7 days later in Session 2. The upper and lower segments of each box plot represent the 75th and 25th percentiles, respectively. Horizontal lines represent medians, and black diamonds represent means. The whiskers of each box plot represent 1.5 times the interquartile range. Individual data points are depicted with open gray circles. Outliers are depicted with solid black circles.

Mediation analysis. Finally, we sought to assess how participants' subjective assessments of the believability and diagnosticity of Kevin's alleged crime influenced the extent to which they exhibited (durable) implicit updating using a similar mediation model to Study 1, except that we used the Session 1 measures of believability and diagnosticity as mediators instead of those from Session 2. In this analysis, the total effect of source credibility on Time 3 implicit preference ($\beta = 0.21$, $p < .001$; partial effect of Time 1 implicit preference: $\beta = 0.41$, $p < .001$) was reduced when diagnosticity and believability were included in the model ($\beta = 0.15$, $p = .001$), providing evidence for a partial mediation of believability. The indirect effect had an associated 95% CI of [.02, .12]. Diagnosticity,

Table 2. Grand Means, Standard Deviations, and Stability of Implicit and Explicit Measures in Study 2

	Time 2		Time 3		Stability	
Condition and measure	*M*	*SD*	*M*	*SD*	*r*	*p*
Rumor and reliable source						
Implicit	−.05	.40	−.01	.40	.60	< .001
Explicit	4.3	2.1	4.3	1.8	.77	< .001
Rumor only						
Implicit	.06	.40	.06	.40	.57	< .001
Explicit	5.5	1.6	5.2	1.5	.68	< .001
Reliable source only						
Implicit	−.15	.37	−.08	.37	.60	< .001
Explicit	3.3	2.0	3.5	1.7	.71	< .001

Note: Implicit evaluations are represented by a difference score between Kevin primes and neutral-stranger primes. Explicit evaluations are the six-item composites at each time point.

however, did not serve as a significant mediator because it did not influence implicit evaluations after analyses controlled for believability (Fig. 8). The 95% CI for the indirect effect of diagnosticity was [–.03, .04].

Discussion

These results demonstrate that the believability of information had durable effects for 7 days, showing that source information about even highly negative information can have a durable influence on implicit impressions. However, so far, we have tested the durability of updated implicit impressions of *novel* targets, about whom participants have limited information and perhaps limited interest. In Study 3, we tested whether the *lack* of implicit updating in response to false information about a well-known target was also similarly durable.

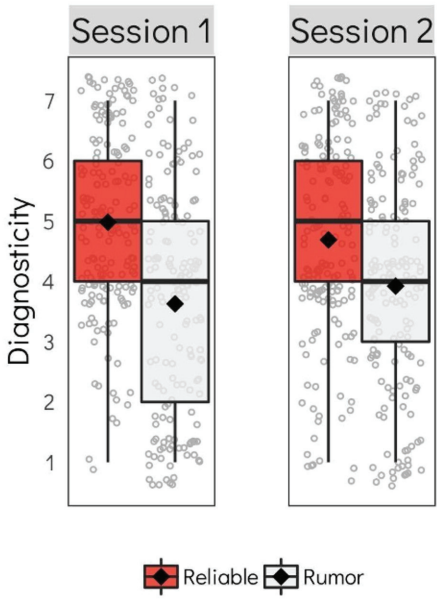

Fig. 7. Diagnosticity assessments in Study 2 as a function of source credibility at each session. Session 2 occurred 7 days after Session 1. The upper and lower segments of each box plot represent the 75th and 25th percentiles, respectively. Horizontal lines represent medians, and black diamonds represent means. The whiskers of each box plot represent 1.5 times the interquartile range. Individual data points are depicted with open gray circles.

Study 3

Method

Participants. The 487 participants who completed all components of a previous study (described below) were eligible to participate in the follow-up experiment. Of these 487 participants, 341 (69.8%) accepted the invitation and submitted a completion code. However, one of these participants did not complete all components of the follow-up session and was excluded from the analyses. An additional seven participants self-reported speaking Mandarin or Cantonese, and 21 pressed a single key on one or more of the three measures of implicit evaluations. Finally, we deviated from our preregistered criteria to exclude 23 participants who partially completed the follow-up session more than once. This resulted in a final total sample of 289 participants (age: M = 29.4 years, SD = 10.2; 59.2% male).

To further test whether there were any lasting effects of exposure to misinformation relative to people who had never seen it, we also recruited a separate set of 200 control participants who had not completed the previous experiment. Two participants failed to complete all components of the experiment, six spoke Mandarin or Cantonese, and five pressed a single key on

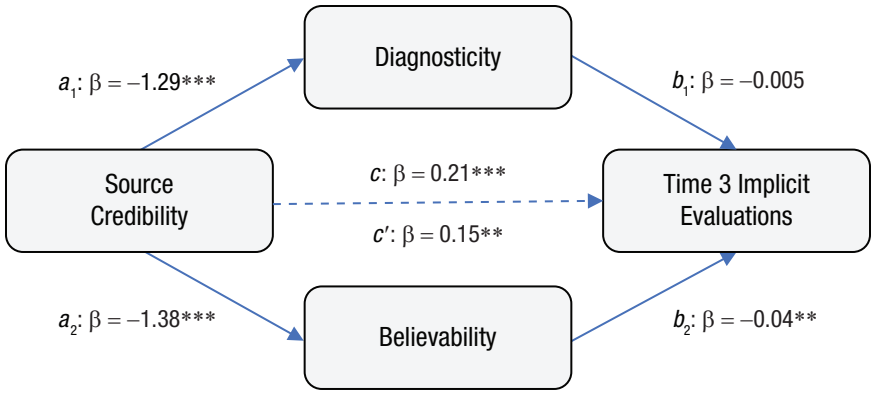

Fig. 8. Mediation model in Study 2: influence of source credibility on Time 3 implicit evaluations, as mediated by diagnosticity and believability. The covariate was Time 1 implicit evaluations. Asterisks indicate significant paths (**p < .01, ***p < .001).

the AMP, leaving a final total sample of 187 participants (age: M = 30.7 years, SD = 11.3; 56.1% male).

Procedure. The first session has been previously reported elsewhere (Cone et al., 2019, Study 7) and involved a procedure in which participants were exposed to fake news about a celebrity (described in more detail below). We conducted a 2-month follow-up to assess whether there were any lingering effects of this exposure.

Session 1. Participants were first shown an image of the actor Jack Black (selected on the basis of a pretest; for more details, see Cone et al., 2019) and asked whether they recognized him. They were also asked to identify him by name in a free-response question. Next, participants completed an AMP that followed the same protocol as in Study 1 except that it assessed their implicit evaluations of Jack Black (25 trials) relative to five other well-known male celebrities (25 trials). Participants did not complete an explicit evaluation measure in the first session.

After participants completed the implicit measure, we exposed them to actual misinformation about Jack Black. Specifically, participants were told that they would be reading a recent news story about Jack Black and that they would be asked questions about it later in the experiment. On the next screen, they were shown a screenshot of what appeared to be a story from an unfamiliar news source that suggested that police records had been uncovered that indicated that Black had been arrested on domestic abuse charges against his wife and stepdaughter.

Afterward, participants were notified that the story was fake and that it was an example of a false story that had been doctored to look as though it came from a reputable news source. The key question of interest at this first time point concerned whether the knowledge that the story was fake was enough to undo its effects on people's implicit evaluations of Jack Black. We assessed this through the primary manipulation, which concerned at what point participants learned that the story was fake—either before or after they completed the Time 2 implicit and explicit evaluation measures. This meant that half of the participants completed the Time 2 evaluation measures with the knowledge that the story was fake, and half completed it unaware that it had been fabricated. After completing the Time 2 implicit and explicit measures, participants filled out a short demographics questionnaire similar to the one in the previous study.

Session 2 (60-day follow-up). Sixty days after the completion of the first session, participants were invited to participate in a follow-up experiment. If participants accepted the invitation, they were first shown the same image of Jack Black as the one they saw at the first time point and were asked to indicate whether they recognized him and to identify him by name. Next, they completed a third measure of implicit and explicit evaluations using the same protocol as the first two measures.

Following the completion of the AMP and explicit evaluation items,[3] as in the first two studies, participants responded to several items for exploratory analyses that assessed their memory for the misinformation they saw 2 months earlier. First, they were asked to recall in as much detail as possible what they were told about Jack Black in the first session. Then, on the subsequent screen, they were asked to recognize the topic of the information they learned on a forced-choice item that included four possibilities (84.4% accuracy) as well as a forced-choice item that assessed their recognition memory for the alleged victims of Black's crimes (85.5% accuracy). They also completed an item that assessed

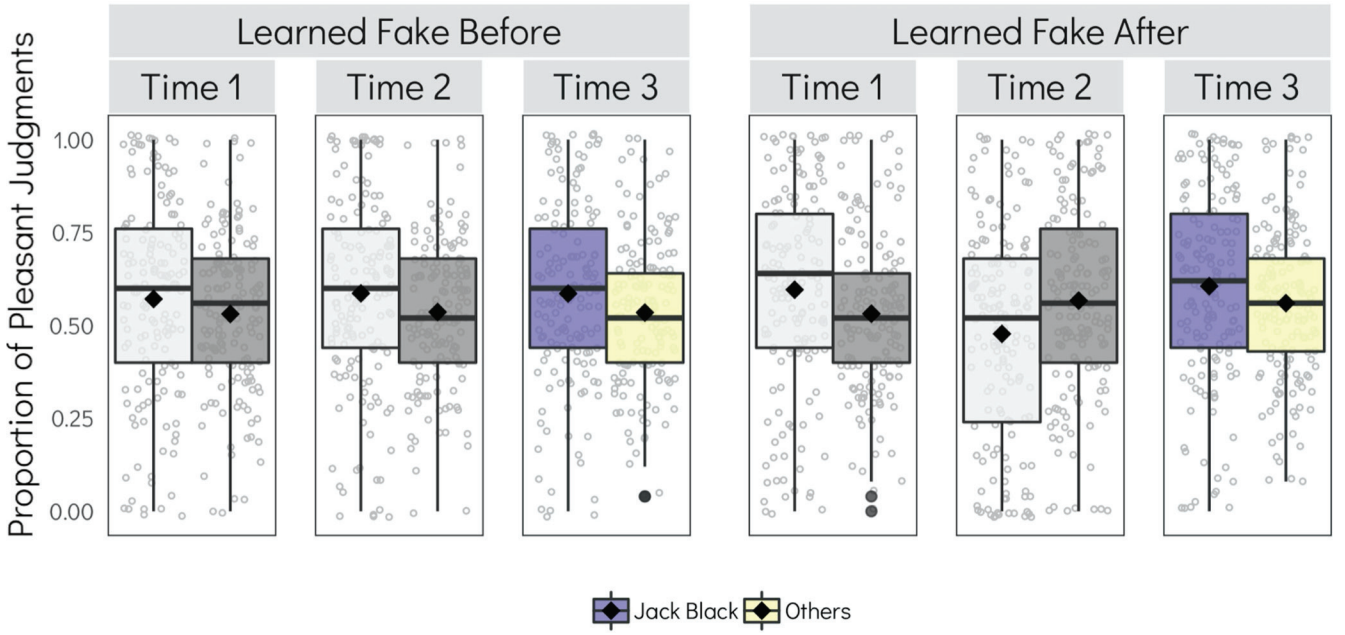

Fig. 9. Implicit evaluations in Study 3: proportion of pleasant judgments as a function of prime type at each of the three time points, separately for the learned-fake-before and learned-fake-after conditions. Time 1 and Time 2 occurred in Session 1, and these data have been reported elsewhere (gray bars; Cone, Flaharty, & Ferguson, 2019). Time 3 occurred 2 months later in Session 2 (colored bars). The upper and lower segments of each box plot represent the 75th and 25th percentiles, respectively. Horizontal lines represent medians, and black diamonds represent means. The whiskers of each box plot represent 1.5 times the interquartile range. Individual data points are depicted with open gray circles. Outliers are depicted with solid black circles.

whether they remembered that the information they learned was false (82.0% accuracy) and were asked whether they looked up anything about Jack Black in the intervening 60 days since they completed the first study (response options were "yes," "no," and "unsure"; 10.7% reported that they had). Finally, they completed a short demographics questionnaire similar to the one in the previous study.

Time 3 controls. The procedure for control participants was identical to that in the Session 2 follow-up, except that we removed all of the exploratory memory items.

Results

As in the first two studies, there was no evidence of systematic differences between the Session 1 and Session 2 samples, nor was there any evidence of differential attrition (Table 1).

Implicit evaluations. We submitted participants' proportion of pleasant judgments on the AMP to a 3 (time: 1, 2, 3) × 2 (target: Jack Black, other celebrities) × 2 (condition: learned fake before, learned fake after) mixed ANOVA. (Control participants were excluded from this analysis.) This revealed a significant three-way interaction, $F(2, 574) = 8.946$, $p < .001$, $\eta_p^2 = .030$ (Fig. 9). To evaluate any changes that may have occurred in participants' implicit responses over the 2-month period between sessions, we focused on Times 2 and 3, conducting a 2 (time: 2, 3) × 2 (target: Jack Black, other celebrities) × 2 (condition: learned fake before, learned fake after) mixed ANOVA, which revealed the predicted three-way interaction, $F(1, 287) = 8.979$, $p = .003$, $\eta_p^2 = .030$. At Time 2, when only a subset of our participants had learned that the information was false, there was a significant Target × Condition interaction, $F(1, 287) = 9.920$, $p = .002$, $\eta_p^2 = .033$. However, at Time 3, only a main effect of target emerged, $F(1, 287) = 7.006$, $p = .009$, $\eta_p^2 = .024$, indicating that the effects of misinformation had been fully undone 2 months later, and the conditions were indistinguishable from one another.

To further assess whether participants exhibited any effects of their exposure to the fake-news story, we compared previous study participants' implicit evaluations in Session 2 with the implicit evaluations of control participants who had never had any exposure to misinformation about Jack Black. Specifically, we conducted a 2 (target: Jack Black, other celebrities) × 3 (condition: learned fake before, learned fake after, controls) mixed ANOVA on study participants' Time 3 AMP

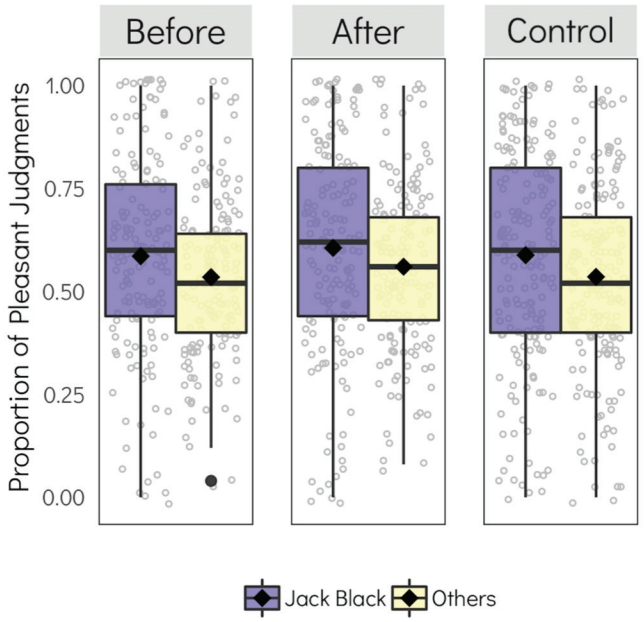

Fig. 10. Implicit evaluations for Session 2 (Time 3) in Study 3: proportion of pleasant judgments as a function of prime type, separately for the learned-fake-before, learned-fake-after, and control conditions. The upper and lower segments of each box plot represent the 75th and 25th percentiles, respectively. Horizontal lines represent medians, and black diamonds represent means. The whiskers of each box plot represent 1.5 times the interquartile range. Individual data points are depicted with open gray circles. The solid black circle is an outlier.

responses (Fig. 10). This analysis revealed only a main effect of target, $F(1, 488) = 13.512$, $p < .001$, $\eta_p^2 = .027$; participants were more implicitly positive toward Jack Black relative to other celebrities. The interaction failed to reach significance, $F < 1$, indicating that follow-up study participants' implicit responses were indistinguishable from controls' responses.

Explicit evaluations. The original study did not include measures of explicit evaluations of Jack Black. However, we conducted a one-way ANOVA to assess whether condition (learned fake before, learned fake after, control) had any effect on the six-item composite ($\alpha = .903$) in the second session relative to control participants. This analysis revealed that there were no significant differences in explicit evaluations among control participants ($M = 5.1$, $SD = 1.0$) versus those who learned that the news story was fake before ($M = 5.0$, $SD = .94$) or after ($M = 5.0$, $SD = 1.0$) the Time 2 evaluation measures, $F(2, 488) = 0.780$, $p = .459$.

Discussion

Study 3 extended our reasoning from novel targets to well-known, familiar ones and is notable for what it does not show—specifically, there was no evidence of a sleeper effect (Kumkale & Albarracín, 2004), nor was there evidence of a continued-influence effect (Johnson & Seifert, 1994). Thus, there was stability in the lack of changes in participants' implicit evaluations in response to unreliable information, despite the extreme negativity of that information.

General Discussion

Little research has focused on implicit-impression updating beyond single-session interventions in which evaluations are measured immediately afterward. The current investigation joins only a small percentage of studies (6.7%; Forscher et al., 2019)[4] that have tested durability of interventions over a period of days, weeks, or months. These findings show that implicit evaluations can be updated in light of new information that casts doubt on the validity of prior learning and that the changes that occur in response to diagnostic and believable information are strikingly robust.

Our results are inconsistent with those of recent work (Lai et al., 2016) showing that many types of interventions reduce implicit intergroup bias temporarily but disappear over longer periods of time (cf. Vuletich & Payne, 2019). Contrary to the findings from interventions used by Lai and colleagues, such as the presentation of a vivid counterstereotypical scenario or providing extensive conditioning of counterstereotypical associations, the new diagnostic evidence in our studies presumably changed participants' beliefs about the target's disposition—although we should be clear that the mediation evidence provided in Studies 1 and 2 is not causal evidence (Judd & Kenny, 2010), and so this proposed mechanism demands further experimental testing. The resulting updating, therefore, persisted across time. Nonetheless, there are a number of other differences between our approach and the nine interventions used by Lai and colleagues. Can these differences explain our conflicting results? For example, our work focused on evaluations of individuals, whereas Lai and colleagues' interventions focused on groups. Even if individual group members engage in diagnostic behaviors such as the ones used in the current work, this does not necessarily generalize to diagnosticity beliefs about the entire group; there are likely a number of factors that influence these more general beliefs, including stereotyping processes such as subtyping (Kunda & Oleson, 1995). It is likely that these more general diagnosticity beliefs—and not beliefs about particular group members—are necessary to elicit durable changes in implicit impressions about groups.

Relatedly, Lai and colleagues (2016) focused on well-known groups. Is stability more likely for novel targets

such as the ones used in the current work? Diagnostic and believable evidence does lead to rapid revision even for well-established targets (i.e., well-known celebrities, Cone et al., 2019; or historical figures, Van Dessel, Ye, & De Houwer, 2019), and we showed in the present Study 3 that updating can be durable for a well-known target. Although more research is needed to examine updating for familiar versus novel targets, one important difference between the two is that it should be more difficult (but not impossible) to create diagnostic and believable evidence that contradicts considerable versus minimal prior learning.

Across our studies, implicit evaluations exhibited a great deal of stability. However, an important conceptual issue concerns what counts as stability. A lack of changes in the overall means observed on an evaluative measure could belie instability in individuals' implicit evaluations because increases in positivity for one individual could be canceled out by decreases in positivity for another. Investigations that have made use of an alternative measure of stability—correlations in evaluative responses over time—have found that implicit evaluations appear to be less stable than their explicit counterparts (Gawronski et al., 2017). As Table 2 shows, our findings also reveal that although implicit evaluations were quite stable in aggregate, they were still nonetheless somewhat less stable than explicit evaluations.

One notable exception in the patterns of stability that we observed was in the reliable-source condition in Study 2, which showed an attenuation in implicit negativity at follow-up. When and why such attenuation occurs are important topics for future research. However, notably, these changes were mirrored by attenuations in people's self-reported beliefs of the diagnosticity of the behavior, which is a strong predictor of implicit updating, thus potentially explaining the changes.

But it is nonetheless striking that attenuation was the exception rather than the rule. Indeed, in nearly every other case, there was scarcely any difference between implicit evaluations measured immediately after exposure and those assessed after additional time had passed. Taken together, these studies suggest that previous claims about the nature of implicit-impression updating on the basis of single-session paradigms (see Cone et al., 2017) can be successfully generalized to larger time horizons in ways that are predictable and durable. This not only provides an important theoretical test of major tenets of current models but also points to practical implications of how social media, gossip, and fake news can affect us at the implicit level.

Transparency

Action Editor: Jamin Halberstadt
Editor: D. Stephen Lindsay

Author Contributions

All the authors developed the study concept together and contributed to the study designs. J. Cone programmed the studies, collected the data, and analyzed the results. J. Cone wrote the initial draft of the manuscript, and K. Flaharty and M. J. Ferguson provided critical revisions. All the authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

Deidentified data and materials for Studies 1 to 3 have been made publicly available via OSF and can be accessed at https://osf.io/sg3vd/. Stimuli that include identifiable human faces are not posted; however, they were sourced from widely available face databases. The design and analysis plans for all three studies were preregistered at https://osf.io/sg3vd/. Changes to the preregistration are reported in the text. This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Jeremy Cone https://orcid.org/0000-0001-5968-6711

Notes

1. The counterbalanced image did not exhibit any main effects or interactions in either impression-formation study (Study 1: *ps* > .128; Study 2: *ps* > .124) and is thus not discussed further.
2. In all three studies, the final sample had similar characteristics to the full sample in Session 1, and there was no evidence of differential attrition across conditions.
3. Because of an oversight, we failed to mention in the preregistration document that we collected an explicit evaluation measure in the method description. All of the analyses we report for the explicit measure are exploratory.
4. This study was a meta-analysis of interventions designed to reduce implicit prejudice toward preexisting groups.

References

Brannon, S. M., & Gawronski, B. (2017). A second chance for first impressions? Exploring the context-(in)dependent updating of implicit evaluations. *Social Psychological and Personality Science*, *8*, 275–283.

Bright-Paul, A., & Jarrold, C. (2009). A temporal discriminability account of children's eyewitness suggestibility. *Developmental Science*, *12*, 647–661.

Cao, J., & Banaji, M. R. (2016). The base rate principle and the fairness principle in social judgment. *Proceedings of the National Academy of Sciences, USA*, *113*, 7475–7480.

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

Cone, J., Flaharty, K., & Ferguson, M. J. (2019). Believability of evidence matters for correcting social impressions. *Proceedings of the National Academy of Sciences, USA*, *116*, 9802–9807.

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. In J. M. Olson (Ed.), *Advances in experimental social psychology* (Vol. 56, pp. 131–199). San Diego, CA: Academic Press.

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

Dudai, Y., Karni, A., & Born, J. (2015). The consolidation and transformation of memory. *Neuron*, *88*, 20–32.

Ecker, U. K., Lewandowsky, S., Cheung, C. S., & Maybery, M. T. (2015). He did it! She did it! No, she did not! Multiple causal explanations and the continued influence of misinformation. *Journal of Memory and Language*, *85*, 101–115.

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

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, *117*, 522–559.

Fourakis, E., Heggeseth, B., & Cone, J. (2020). *Explaining why: The role of causal attribution in implicit impression formation and revision*. Manuscript submitted for publication.

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, *43*, 300–312.

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2018). Contextualized attitude change. In J. M. Olson (Ed.), *Advances in experimental social psychology* (Vol. 57, pp. 1–52). San Diego, CA: Academic Press.

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.

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.

Guillory, J. J., & Geraci, L. (2013). Correcting erroneous inferences in memory: The role of source credibility. *Journal of Applied Research in Memory and Cognition*, *2*, 201–209.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1420–1436.

Judd, C. M., & Kenny, D. A. (2010). Data analysis in social psychology: Recent and recurring issues. *Handbook of Social Psychology*, *1*, 115–139.

Kerpelman, J. P., & Himmelfarb, S. (1971). Partial reinforcement effects in attitude acquisition and counterconditioning. *Journal of Personality and Social Psychology*, *19*, 301–305.

Kumkale, G. T., & Albarracín, D. (2004). The sleeper effect in persuasion: A meta-analytic review. *Psychological Bulletin*, *130*, 143–172.

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology*, *68*, 565–579.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., . . . Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*, 1001–1016.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago Face Database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*, 1122–1135.

Mann, T. C., Cone, J., Heggeseth, B., & Ferguson, M. J. (2019). Updating implicit impressions: New evidence on intentionality and the affect misattribution procedure. *Journal of Personality and Social Psychology*, *116*, 349–374.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

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

McGaugh, J. L. (2000). Memory: A century of consolidation. *Science*, *287*, 248–251.

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.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*, 233–248.

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.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891.

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

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.

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.

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

Swire, B., Ecker, U. K., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1948–1961.

Van Dessel, P., Ye, Y., & De Houwer, J. (2019). Changing deep-rooted implicit evaluation in the blink of an eye: Negative verbal information shifts automatic liking of Gandhi. *Social Psychological and Personality Science*, *10*, 266–273.

Vuletich, H. A., & Payne, B. K. (2019). Stability and change in implicit bias. *Psychological Science*, *30*, 854–862.

Zárate, M. A., & Enge, L. R. (2013). The role of memory consolidation during sleep in social perception and stereotyping. In B. Derks, D. Scheepers, & N. Ellemers (Eds.), *Neuroscience of prejudice and intergroup relations* (pp. 130–145). New York, NY: Psychology Press.