Believability of evidence matters for correcting social impressions

Jeremy Cone[a,1], Kathryn Flaharty[b], and Melissa J. Ferguson[c]

[a]Department of Psychology, Williams College, Williamstown, MA 01267; [b]Department of Psychology, Georgetown University, Washington, DC 20001; and [c]Department of Psychology, Cornell University, Ithaca, NY 14850

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved April 1, 2019 (received for review February 27, 2019)

To what extent are we beholden to the information we encounter about others? Are there aspects of cognition that are unduly influenced by gossip or outright disinformation, even when we deem it unlikely to be true? Research has shown that implicit impressions of others are often insensitive to the truth value of the evidence. We examined whether the believability of new, contradictory information about others influenced whether people corrected their implicit and explicit impressions. Contrary to previous work, we found that across seven studies, the perceived believability of new evidence predicted whether people corrected their implicit impressions. Subjective assessments of truth value also uniquely predicted correction beyond other properties of information such as diagnosticity/extremity. This evidence shows that the degree to which someone thinks new information is true influences whether it impacts implicit impressions.

implicit | truth | believability | attitudes | first impressions

A quip often attributed to Mark Twain proposes that "a lie can travel halfway around the world while the truth is still putting on its boots" (1). Ironically, it appears that Mark Twain never said this (although this has not prevented it from traveling halfway around the world). Nonetheless, the idea that we find ourselves awash in a sea of misinformation that can lead us to erroneous conclusions has had a long tradition in Western thought, with evidence of variations of this quote circulating as early as the mid-18th century (2). Still, the digital age presents us with the ability to acquire more information—and perhaps more apocryphal information—than at any time in history (see, e.g., ref. 3). This is perhaps especially true of the (mis)information that we encounter about other people. The things that we learn about others that form the basis of our impressions of them come to us not just through what we directly observe or what we learn from others at the water cooler, but also through status updates, tweets and retweets, third-party information, and weak links in our social networks. It is now, quite literally, easier than ever before for lies about others to travel halfway around the world.

Given the wealth of information available to us, a nontrivial task that we face to successfully exploit it is to accurately decide which particular pieces—and sources—of information deserve our attention, consideration, and, ultimately, acceptance. However, surprisingly, research assessing the processes underlying impression updating in light of new information suggests that we often fall short in this fundamental task. Whereas our explicit impressions (i.e., those that are self-reported and therefore intentional) readily incorporate validity and are highly responsive to the believability of new information, implicit impressions (i.e., those that are measured indirectly and are therefore unintentional) are thought to rely on different underlying processes (4, 5) and appear to be relatively insensitive to such considerations. Accordingly, explicit impressions are highly sensitive to negations (e.g., "all of the information you have just learned is false") but implicit impressions appear to be largely insensitive, exhibiting relatively little, if any, updating (6) (cf. refs. 7 and 8). Similarly, complex object relations that reverse the evaluative connotations of an attitude object [e.g., "sunscreen prevents sunburns" indicates that sunscreen is positive despite often being associated with and occurring in close proximity to something negative (9, 10)] are generally not represented by implicit impressions. Moreover, whereas explicit impressions can reflect information contrary to strong base rates such as when we encounter a male nurse or a female doctor, implicit impressions appear to be less capable of reflecting such counterstereotypical information (11).

This is concerning because implicit impressions have been shown to be uniquely predictive of how we behave (12, 13) (but cf. refs. 14 and 15). For example, previous work has demonstrated that voting behavior is predicted by people's implicit impressions of political candidates (16, 17). Thus, the notion that implicit impressions are insensitive to truth value raises the worrisome possibility that even information that we deem to be false—things we learn from obviously questionable sources or outright disinformation campaigns—might nonetheless become incorporated into our impressions of others and influence how we act toward them.

Our goal in the present investigation is to test whether the perceived truth value of new evidence predicts the extent to which someone uses that evidence to update their implicit impressions. We focus in particular on whether new evidence can undo or modify existing impressions (i.e., updating), rather than initial learning or impression formation. This is important to investigate because impression formation and correction may be governed by distinct processes (18). For example, although implicit impressions can be sensitive to the stated accuracy of information during initial formation (8, 19), they seem to be relatively insensitive to truth value after they have formed (18). We also focus on the role of truth value in revision because it is highly consequential: indeed, the goal of many disinformation campaigns is to change people's minds about someone, such as when rumors are spread to inflict damage on otherwise positive reputations.

Significance

The digital age affords exposure to a staggering amount of information—not all of it true. The extent to which mere exposure to information of uncertain veracity or outright disinformation campaigns shapes our impressions of others, independent of our subjective assessments of its truth value, is thus a key question with important implications, especially because implicit evaluations have been shown to uniquely predict behaviors such as which politician a voter ultimately votes for in an election. This study sheds light on the nature of implicit cognition and the extent to which dissociations between implicit and explicit evaluations can be successfully explained by differential reliance on propositional learning processes.

Author contributions: J.C., K.F., and M.J.F. designed research; J.C. and K.F. performed research; J.C. analyzed data; and J.C., K.F., and M.J.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Unless otherwise indicated, procedures, sample size, exclusion criteria, hypotheses, and data analysis plans were preregistered before data collection on the Open Science Framework (https://osf.io/3h2ww/).

[1]To whom correspondence should be addressed. Email: jdc2@williams.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1903222116/-/DCSupplemental.

Published online April 29, 2019.

Our investigation tests the role of believability in updating by building on two previous lines of research. In our first two studies, we extend research showing that implicit impressions are difficult to update (18, 20, 21). In these situations, perhaps any failure to correct initial impressions results from low subjective believability of the new information. If truth value matters for the revision of implicit impressions, then increasing the subjective truth value of new information should result in greater updating. We first test this possibility in Study 1 (preregistered, see *Methods*) by exposing participants to a new paradigm in which it is highly believable to encounter new information about a target that is inconsistent with prior information. In such a case, we would expect that most people will show rapid updating. In Study 2, we measure people's subjective truth value in a paradigm used in previous work, showing that believability predicts whether people correct their impressions.

We then turn to a second line of research showing that implicit impressions can be updated rapidly when the new evidence is extreme in valence (22–24). Our account suggests that a prerequisite for such rapid revision is that the new information must be seen as believable. In these cases, if truth value is an important factor in updating, then reductions in the believability of new information should result in less correction. We test and find support for this possibility in five studies (four preregistered) showing that the influence of even extreme new impression-inconsistent information depends on its believability.

Study 1: Wugs

In Study 1, we tested how implicit impressions were formed and updated in a context in which the changes in the valence of a target were highly believable: a video game in which the game characters switched from positive (helpful) to negative (harmful) between rounds of the game. Consider the situation participants face when playing PAC-MAN: While the ghosts must be avoided in the regular mode of the game, eating a power pellet instantly switches their valence so that they can be chased and eaten for points. In this kind of scenario in which the evaluative nature of characters is instantly changeable, we should expect all participants to show robust and rapid updating of their implicit impressions.

Participants were told that they would be playing a video game consisting of multiple rounds. Participants learned that their character—a purple square—would encounter objects called "wugs," which were triangles with eyes much like the ghosts in the original PAC-MAN. In one round of the game, participants learned that wugs would be helpful to them and if they could "hug a wug"—touch the wugs using their character—they would earn points. In the other round of the game, participants learned that the wugs would be harmful to them and should be avoided. They were told that they should try to avoid getting "mugged by a wug," which meant that the wugs would rob their character of their points.

Participants read the instructions for each of the two rounds of the game in a counterbalanced order. Immediately after each set of instructions, we measured their implicit evaluations of wugs using an affect misattribution procedure (AMP) (60 trials: 30 wug primes, 30 neutral gray squares; ref. 25). The study was thus a 2 (Round: approach, avoid) × 2 (Target: wugs, neutral) × 2 (Order: helpful first, harmful first) design. If implicit impressions are easier done than undone, then we should anticipate that participants will show significant formation at Time 1, but will show little change in their impressions after encountering information inconsistent with their initial exposure. However, our preregistered prediction was that, because the changes in valence are highly believable in this world of video games, participants would show both significant formation and significant revision at each time point (26). An ANOVA indicated that the predicted three-way interaction among Time, Target, and Order emerged, $F(1, 322) = 55.943$, $P < 0.001$, $\eta_p^2 = 0.148$ (Fig. 1, *Left*). Participants who read the approach instructions first were significantly more implicitly positive toward wugs than neutral squares at Time 1, $t(163) = 3.928$, $P < 0.001$, $d = 0.31$ (95% CI of the difference, 0.051 to 0.154), and significantly more negative at Time 2, $t(163) = -5.282$, $P < 0.001$, $d = 0.41$ (95% CI, −0.21 to −0.096).

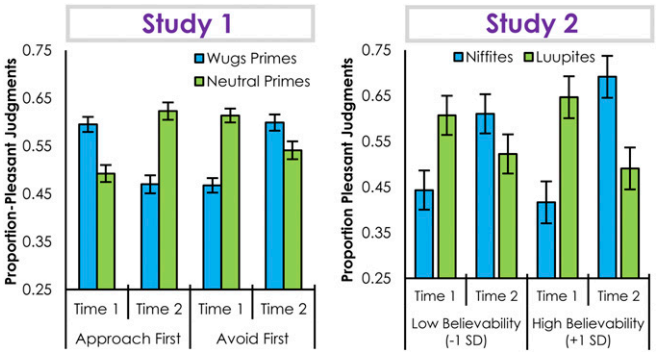

Fig. 1. The results of Studies 1 and 2. (*Left*) Study 1 ($n = 324$): implicit evaluations exhibited a rapid reversal between descriptions of rounds of the game when changes in valence were readily believable, as they are in the context of a video game. (*Right*) Study 2 ($n = 167$): predicted values for a linear mixed-effects model that predicted evaluations of Niffites and Luupites as a function of learning (Time 1), counterlearning (Time 2), and participants' subjective assessments of the believability of the narrative. Values are plotted at believability scores of 4.30 (−1 SD) and 6.33 (+1 SD). Error bars represent SEs.

Similarly, participants who read the avoid instructions first showed significant negativity at Time 1, $t(159) = -6.137$, $P < 0.001$, $d = 0.49$ (95% CI, −0.19 to −0.099), and significant positivity at Time 2, $t(159) = 2.029$, $P = 0.044$, $d = 0.16$ (95% CI, 0.0015 to 0.114).

Thus, in Study 1, participants' implicit impressions were highly sensitive to switches in the evaluative implications of attitude objects (helpful vs. harmful). Moreover, contrary to previous work (18), these changes were symmetrical: Focusing on evaluations in the approach round, participants were equally implicitly positive toward wugs independent of the order in which they read the instructions, $t < 1$. Similarly, in the avoid round, participants were equally implicitly negative toward the wugs, independent of whether they encountered the instructions for this round first or second, $t < 1$.

Study 2: Niffites and Luupites

Outside of video games, the perceived truth value of new information about human actors is often highly variable and in the eye of the beholder. In Study 2, we more directly assessed the role of perceived truth value in rapid updating using a correlational approach in which we measured people's subjective assessments of the believability of the information they learned. Participants read a narrative describing an intergroup conflict between two novel groups of people referred to as the Niffites and Luupites (18). Previous research using this paradigm has suggested that participants can quickly form implicit impressions of the groups. However, several attempts to induce corrections in these impressions resulted in little evidence of updating.

Like previous work, participants were told that the story was based on real events but the names had been changed to ensure that their impressions of the groups were unbiased by prior exposure. They were given a detailed description of how the Niffites were terrible, malicious, and morally bankrupt, while the Luupites were moral, virtuous, and benevolent. Afterward, in a counterinduction, participants were provided with additional narrative details that attempted to convince them that the groups had switched roles: the Niffites collapsed under the weight of their own corruption and realized the errors of their ways, while the Luupites became bitter and angry at their oppression and descended into acts of terrorism against the Niffites.

In between the induction and the counterinduction, participants completed an AMP that incorporated the names of Niffites and Luupites as primes (60 trials: 30 Niffites, 30 Luupites). At the end of the experiment, we asked participants how believable they found each part of the narrative (six items; e.g., "To what extent could this happen in real life?") as well as how much they endorsed the idea that groups of people can change character over time (three items; "In general, to what extent do you think

that groups of people are likely to change their character over time?"). A factor analysis indicated that these two subscales formed distinct factors, and they were thus analyzed separately.

Using a linear mixed-effects model (Table 1), we tested how participants' assessments of the believability of the information influenced both formation and revision of implicit evaluations. When the first factor (believability) was included in the model, the three-way interaction among Time, Target, and Believability was significant (Fig. 1, *Right*). This interaction was not significant for the group malleability factor.

Studies 3 to 6: Rumors

In Studies 3 to 6, we sought to complement the results of the first two studies by using situations in which previous research has showed that implicit impressions are capable of rapid revision in single-trial learning (22–24) and by demonstrating that casting doubt on the validity of impression-inconsistent information results in a reduction in implicit impression correction. Additionally, to complement the correlational results from Study 2 with an experimental approach, Studies 3 to 6 manipulated the subjective truth value of impression-inconsistent information.

These studies used an impression-formation paradigm (20, 21, 27) in which participants learned about the behaviors of a novel individual named Kevin. On each trial [$n = 50$ (except Study 6, which included 20; see *Methods*)], a picture of Kevin's face (randomly assigned from one of six college-aged white males) was shown above a behavioral statement. Participants indicated on each trial whether they thought the behavior was characteristic or uncharacteristic of Kevin and received immediate feedback. On a 100% positive reinforcement schedule, all positive behaviors (e.g., "Kevin donates his time at the soup kitchen"; $n = 25$) were indicated to be characteristic of Kevin and all negative behaviors (e.g., "Kevin had someone else take a math final for him"; $n = 25$) were indicated to be uncharacteristic.

After developing a uniformly positive impression of Kevin, participants learned their 51st piece of information about him. However, this piece of information was selected to be highly inconsistent with the previous impression they had formed and to be extreme in valence (see ref. 22): "Kevin was arrested a few years back for domestic abuse of his ex-wife" (Study 3) or "Kevin was arrested a few years back for child molestation of his young niece" (Studies 4 to 6). We manipulated the believability of this new information by varying the credibility of the source from which it was learned. Some participants learned that the information was acquired by discovering police reports that unequivocally established his guilt (reliable source condition). Others were asked to imagine that they had acquired the information from a coworker who had reason to spread a negative rumor about Kevin: Molly shared the details of the alleged crime shortly after one of her friends had broken up with him. Because previous work has suggested that information that is described as a rumor is deemed to be less reliable (28), we anticipated that participants would consider the diagnostic Time 2 behavior as less believable in this condition.

Like the previous studies, we measured participants' implicit evaluations of Kevin immediately before and after learning this information using an AMP (60 trials: 30 trials Kevin, 30 trials neutral strangers). At the end of Studies 3 and 4, we measured participants' assessments of both the diagnosticity of the Time 2 behavior (22) as well as how much they believed it to be true (e.g., "How likely do you think it is that Kevin actually engaged in this behavior?"). This allowed us to evaluate a preregistered multiple mediation model to assess each of their relative contributions to rapid implicit impression updating.

All four experiments followed this 2 (Time: 1, 2) × 2 (Target: Kevin, neutral strangers) × 2 (Source: reliable, rumor) mixed design (except Study 6, which included only a single implicit measure; see below). Study 3 sought to establish the effect using a moderately negative behavior (i.e., "Kevin was arrested for domestic abuse"). In an ANOVA, a significant three-way interaction among Time, Target, and Source emerged, $F(1, 388) = 7.685$, $P = 0.006$, $\eta_p^2 = 0.019$ (Fig. 2). All participants formed significantly positive impressions of Kevin at Time 1 [reliable: $t(197) = 6.531$, $P < 0.001$, $d = 0.46$ (95% CI, 0.10 to 0.19); rumor: $t(191) = 7.347$, $P < 0.001$, $d = 0.53$ (95% CI, 0.12 to 0.21)]. After learning the diagnostic information, participants' implicit evaluations in the reliable source condition significantly changed over time [evidenced by a two-way interaction between Time and Target, $F(1, 197) = 30.151$, $P < 0.001$, $\eta_p^2 = 0.133$, becoming implicitly neutral toward Kevin relative to neutral strangers, $t(197) = -1.126$, $P = 0.261$, $d = 0.08$ (95% CI, −0.08 to 0.02)]. In contrast, those in the rumor condition exhibited significant change [again, as evidenced by a significant two-way interaction between Time and Target, $F(1, 191) = 7.029$, $P = 0.009$, $\eta_p^2 = 0.035$, but were still nonetheless significantly implicitly more positive toward Kevin relative to neutral strangers, $t(191) = 4.031$, $P < 0.001$, $d = 0.29$ (95% CI, 0.05 to 0.15)].

In Study 3, we evaluated a multiple mediation model (29) to provide evidence for the mechanism(s) behind rapid implicit change. We created a measure of implicit preference for Kevin at each time point by creating a difference score from the proportion-pleasant judgments for Kevin and neutral strangers. Time 2 implicit preference served as the dependent measure and Time 1 implicit preference served as a covariate (22). Source credibility condition served as the independent variable, and both believability and diagnosticity (single-item measures; see above) were entered as simultaneous mediators. The results of the bootstrapped analysis are reported in Table 2. Like our previous work, in Study 3, the extent to which participants saw Time 2 behavior as diagnostic was a significant mediator of the effect of condition on the extent of impression updating (but not Study 4; see Table 2). However, importantly, believability also emerged as a significant mediator, even after controlling for perceptions of diagnosticity.

Study 3 established the importance of believability of information for the extent of implicit revision. However, because there was not a significant reversal of implicit responses at Time 2 (cf. ref. 22), it is unclear whether believability matters in the face of more extreme (and thus diagnostic) behaviors. Study 4 was therefore identical to Study 3, except that it was preregistered and made use of a more extreme negative behavior: "Kevin was arrested for child molestation." Like Study 3, a significant three-way interaction among Time, Target, and Source emerged, $F(1, 366) = 25.719$, $P < 0.001$, $\eta_p^2 = 0.066$, indicating that the believability of the information influenced the extent of implicit revision (Fig. 2).

All participants exhibited significant formation at Time 1. However, contrary to Study 3, in the reliable source condition, participants now exhibited a significant reversal in their implicit responses toward Kevin relative to neutral strangers, $t(187) = -7.350$, $P < 0.001$, $d = 0.54$ (95% CI, −0.28 to −0.16). In the rumor condition, whereas

Table 1. Linear mixed-effects model in Study 2

Effect	Model 1	Model 2
Time	−0.01 (0.009)	−0.01 (0.009)
Target	0.02 (0.009)**	0.02 (0.009)**
Time × Target	0.09 (0.009)***	0.09 (0.009)***
Believability (B)	−0.001 (0.008)	
Time × B	0.01 (0.009)	
Target × B	0.01 (0.010)	
Time × Target × B	0.02 (0.009)*	
Group malleability (GM)		0.02 (0.008)
Time × GM		0.01 (0.009)
Target × GM		−0.01 (0.009)
Time × Target × GM		0.01 (0.009)
Constant	0.56 (0.045)***	0.45 (0.039)***
Sample size	164	164
Adjusted R^2	0.077	0.059

The outcome variable in all models is implicit evaluations, and all models include a random effect of subject. Values in parentheses are standard errors.
*$P < 0.05$.
**$P < 0.01$.
***$P < 0.001$.

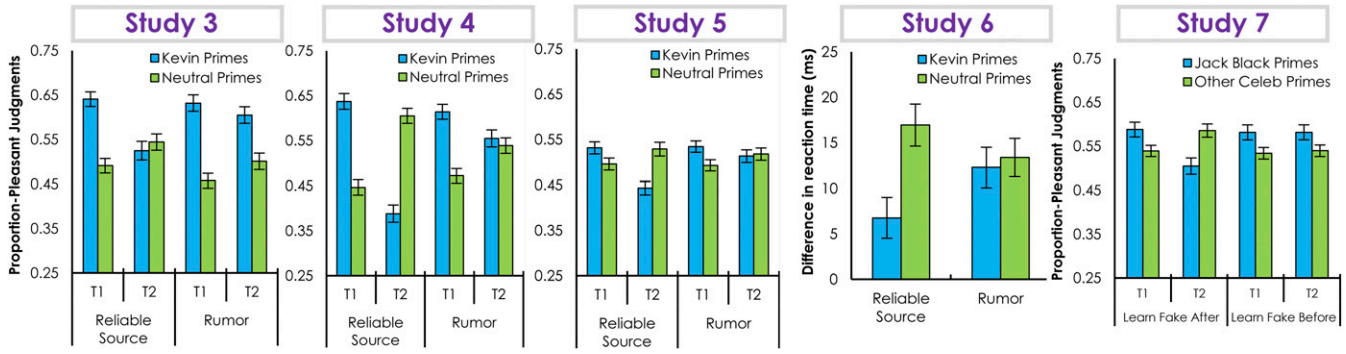

Fig. 2. The results of Studies 3 to 7. Study 3 (n = 390): an initial demonstration that made use of a moderately diagnostic Time 2 behavior. Study 4 (n = 368): a preregistered replication that used a more extreme Time 2 behavior. Study 5 (n = 455): a modified version of the study that induced a 120-s delay before participants encountered the source credibility manipulation. Study 6 (n = 705): a preregistered replication that used an evaluative priming task instead of an AMP. Study 7 (n = 457): a preregistered adaptation using actual misinformation and a real-world target (Jack Black). Error bars represent SEs. Celeb, celebrities.

participants exhibited significant changes in their implicit evaluations in response to the new information, $F(1, 179) = 17.873$, $P < 0.001$, $\eta_p^2 = 0.091$, they exhibited less change than the reliable source condition, becoming implicitly neutral at Time 2 toward Kevin relative to neutral strangers, $t(180) = 0.565$, $P = 0.573$, $d = 0.04$ (95% CI, −0.04 to 0.07).

Some research has found that implicit impressions become less responsive to validity information after a delay. Whereas negations may be successfully incorporated at the time of information acquisition, they apparently become more rigid after even just a 2.5-min delay before learning whether information is true or false (7) (but cf. ref. 8). Thus, in Study 5, we tested how the delayed discovery of the reliability of information influenced implicit impression correction. The study design was preregistered and similar to Study 4, except that all participants initially learned that Kevin had been arrested for child molestation. Next, they were required to wait for 150 s before having an opportunity to continue with the study, and only after continuing did they then learn whether the information came from a reliable source or was an unsubstantiated rumor. They then proceeded to the Time 2 implicit evaluation measure as in the previous studies (but using a slightly modified protocol; see *Methods*). Contrary to previous research (7), we find that the extent of revision in this condition is indistinguishable from the previous results in which the credibility of the information was considered at the time of information acquisition. Like the previous studies, the three-way interaction emerged, $F(1, 453) = 5.996$, $P = 0.015$, $\eta_p^2 = 0.013$. While participants in the reliable source condition exhibited a significant reversal at Time 2, $t(205) = -5.198$, $P < 0.001$, $d = 0.36$ (95% CI, −0.12 to −0.05), participants in the rumor condition were less responsive to the Time 2 information, showing no implicit preference between Kevin and neutral strangers, $t(248) = -0.310$, $P = 0.756$, $d = 0.02$ (95% CI, −0.03 to 0.02).

Some research suggests that outcomes on the AMP can sometimes diverge from other implicit measures. In particular, the AMP has been empirically challenged on the extent to which it successfully captures implicit vs. intentional responding (30–32). [In light of these concerns, we also conducted a study (reported in *SI Appendix, Supplementary Materials*) that modified the protocol of the AMP in ways that have been shown to produce better psychometric properties. This study fully replicated the other studies.] To test whether our findings were unique to the AMP, in Study 6, we conducted a preregistered replication of Study 4 using an evaluative priming task (EPT) (33). This reaction-time–based measure assesses whether a prime (i.e., attitude object) facilitates versus interferes with responding to unrelated positive or negative information. We included only one implicit measure, which occurred after exposure to the impression-inconsistent rumor information. Thus, the study was a 2 (Target: Kevin, neutral stranger) × 2 (Source: reliable, rumor) design. We tested our predictions using a linear mixed-effects model that included random effects of both subject and target word (see *SI Appendix, Supplementary Materials* for details). Consistent with the previous findings, the Source × Prime × Target interaction was significant in this model (Fig. 2 shows differences in reaction times for positive and negative targets. Higher numbers on this measure indicate greater implicit positivity). Thus, our findings emerged on two widely used implicit measures and were not unique to the AMP.

Study 7: Misinformation About a Well-Known Target

The previous studies consistently support the role of believability as a key determinant of correcting implicit impressions in response

Table 2. Bootstrapped multiple mediation analyses for Studies 3 and 4

Mediator	Effect of independent variable on mediator (a)	Effect of mediator on dependent variable (b)	Indirect effect of mediator (ab)	Total indirect effect (ab)	Direct effect (c′)	Total effect (c)	Mediation type
Study 3							
Believability	−1.43***	−0.0256†	0.0365	0.084*	0.0291	0.11**	Full
Diagnosticity	−1.45***	−0.0331*	0.0478*				
Study 4							
Believability	−1.99***	−0.031*	0.061*	0.085*	0.15**	0.23***	Partial
Diagnosticity	−1.77***	−0.013	0.024				

All analyses included condition as the independent variable, believability and diagnosticity as simultaneous mediators, Time 2 implicit evaluations as the predictor variable, and Time 1 implicit evaluations as a covariate.
*$P < 0.05$.
**$P < 0.01$.
***$P < 0.001$.
†$P = 0.052$.

to diagnostic information. However, all these studies used novel and hypothetical targets. If believability influenced the updating of implicit impressions only of novel and recently encountered targets, but not of more familiar targets about whom we have much greater and more diverse evaluative histories, then the applicability of this work would be limited. Thus, our goal in our final preregistered study was to extend these findings into the realm of real-world exposure to disinformation about a well-known target. On the basis of pretesting (*SI Appendix*, *Supplementary Materials*), we selected as our target of misinformation a well-known and well-liked celebrity about whom participants had largely positive implicit evaluations: Jack Black. We also used (mis)information that was similar in theme and content to actual news stories focused on revelations concerning well-known celebrities' bad behavior toward women.

All participants were exposed to a story suggesting that Jack Black had been charged with multiple accounts of domestic abuse (doctored to look like a real news site). We measured participants' implicit evaluations of Jack Black using an AMP (50 trials: 25 Jack Black, 25 other well-known male celebrities) completed before and after exposure to the story. All participants learned, over the course of the study, that the story was fake. However, our primary manipulation was the timing of this information: either immediately before the second AMP or immediately after. In this way, all participants were exposed to an instance of disinformation, but half of our subjects completed the second AMP with no additional information about its veracity, and half completed it with the awareness that it was fake. Previous work suggests that this knowledge should have little influence on implicit evaluations (7). The design was a 2 (Time: 1, 2) × 2 (Target: Jack Black, other celebrities) × 2 (Timing of Learning Story Was Fake: before Time 2 AMP, after Time 2 AMP) design.

Consistent with the previous studies, the three-way interaction emerged, $F(1, 455) = 19.232$, $P < 0.001$, $\eta_p^2 = 0.041$ (Fig. 2). Whereas a significant two-way interaction emerged when participants thought the story was real, $F(1, 227) = 31.115$, $P < 0.001$, $\eta_p^2 = 0.121$, when they knew the story was fake, there was only a main effect of Target, $F(1, 288) = 4.889$, $P = 0.028$, $\eta_p^2 = 0.021$, that was unqualified by Time, $F(1, 288) = 0.119$, $P = 0.73$. A single exposure to fake news can therefore result in instantaneous implicit revision, even about a well-known and well-liked target (see also ref. 34). However, simply indicating that the story is false is sufficient to completely eliminate any effects of exposure to misinformation, even at the implicit level (cf. ref. 7). [Reconducting all of these analyses among only the subset of participants who reported recognizing Jack Black (79.6%) did not change any of our conclusions; see *SI Appendix*, *Supplementary Materials* for more details].

General Discussion

In an environment increasingly populated by information of dubious veracity and outright disinformation, it would be troubling if our impressions of others were unduly influenced by information we reject as unreliable, raising questions about the functionality of implicit cognition as well as the ways in which explicit beliefs interact with implicit mental representations. However, across seven studies (five preregistered) using four different paradigms including both correlational and experimental designs, we consistently show that the believability of information influences the extent to which that information leads to rapid changes in people's implicit impressions of others. Even when the information that participants encountered was highly diagnostic—a factor shown in prior work to be a key determinant of rapid implicit revision (22, 23)—believability was not just a predictor but a unique predictor of revision, suggesting that even when information is extreme, it must still possess some quality of truth for it to be fully incorporated into impressions of others.

Consistent with prior research on attitude formation (19), these results emphasize the importance of factors beyond the content of the information in determining its effects on implicit impressions. Whereas previous work has often assumed that the impact of new information is largely a property of its valence assessed in isolation (20, 21, 26), our findings underscore the need to understand not just what we learn, but also whom we learn it from, the context in which it is encountered, and our metacognitions about the quality of the information and its source. In this way, the very same new information about an attitude object may or may not lead to substantive correction depending on the perceived credibility of the source—perceptions that are likely to differ markedly among people. Perhaps, then, the philosopher Origgi (35) is correct when she suggests that we live not in an "age of information" but rather in an "age of reputation" in which the most impactful information is that which is shared by the most reputable among us—not just for deliberate evaluations, but also for relatively more unintentional ones.

Perhaps the most striking aspect of our work is that we find an effect of manipulations of believability—even after a time delay and even after participants have expressed their initial impression—on both an implicit and an explicit measure. Even after people have rehearsed and expressed their initial impression, once they confront new evidence that is believable, they can correct it, even at the implicit level. This extends earlier work that suggested that information can be negated only if we know that it is false at the time we learn it (7) or very soon after if there is extensive rehearsal of the information (8) but that such negations are less impactful if we only later discover that the information is false. This earlier work suggested (7) that once an association has formed through the associative processes that are argued to underlie implicit cognition, it cannot be easily corrected thereafter.

However, we find equivalent updating in the immediate and time-delay conditions. Even after acquiring negative information about Kevin and only later learning it lacked credibility, participants corrected their implicit impressions. Earlier work may have failed to find evidence for updating only because the new evidence was generally low in believability. For example, Study 2 shows using a prior paradigm that believability significantly affected updating. This challenges the idea that there are different processes underlying implicit formation vs. revision of implicit impressions (18) as well as the contention that implicit and explicit learning and cognition operate via different processes (7, 36). Indeed, these findings reveal a larger role for propositional processes in implicit cognition than most theories currently grant (37).

These results suggest that disinformation may have a more difficult time taking root, even implicitly, than current theories might suggest. However, if believability is an influential factor in the acceptance of new information, then why is it the case that misinformation seems to make it halfway around the world while the truth is still putting on its boots? Our findings suggest that if a person deems information to be less believable, it will be more likely to be cast aside and ignored. However, the factors that influence whether and when people deem information less believable still remains an open question, and it is likely that motivated reasoning plays a role in these subjective assessments. Whether a source is deemed credible or not is likely influenced by whether the views shared by that source are congenial to one's worldview and preexisting beliefs (e.g., whether left-leaning or right-leaning political sources are seen as objective sources or instruments of biased, partisan viewpoints; ref. 38). In the current studies, participants had no reason to question the credibility of the information that we provided; thus, motivated reasoning had less opportunity to influence subjective assessments. There are likely many situations in which the information that we learn is similarly unassailable, as when we directly observe someone committing wrongdoing or when someone is unequivocally exonerated on the basis of reliable scientific testing. Thus, a key determinant of how people respond to misinformation is the extent to which they can be convinced of the accuracy and credibility of a source (or, in the other direction, its bias and unreliability).

In addition to these practical implications, this work also joins other recent findings that have challenged contemporary assumptions about the nature of implicit cognition and the characteristics that govern implicit learning, updating, and correction. Our research suggests that lies can indeed travel halfway around the world, but they are perhaps more likely to take hold among those who see truth in those lies, even at the implicit level.

Methods

All studies were reviewed and approved by either the Cornell University or the Williams College IRB, and all participants gave their consent to participate.

Preregistration Documents. The procedures, sample size, exclusion criteria, hypotheses, and data analysis plans for Studies 1 and 4–7 were preregistered prior to data collection on the Open Science Framework (https://osf.io/3h2ww/.)

AMP. The procedure for the AMP was identical across all studies (except Study 5, which used a modified protocol; see below). On each trial, participants saw a prime (75 ms), followed by an interstimulus interval (ISI; 125 ms), a Chinese pictograph (100 ms), and, finally, a white noise pattern until the participant provided a response (25). They pressed the "d" key to indicate that the Chinese pictograph was less pleasant than average or the "k" key to indicate that the pictograph was more pleasant than average. The measure was the proportion of times they selected more pleasant than average for each prime type.

Study 1: Wugs. Participants were 360 Amazon's Mechanical Turk (MTurk) workers [60.2% male, mean age (M_{age}) = 36.3 y]. Twenty-three participants submitted a completion code but did not complete all components of the study and were excluded from analyses. The final sample was $n = 324$. Game instructions are available in *SI Appendix, Supplementary Materials*.

Study 2: Niffites and Luupites. Participants were 167 MTurk workers (40% male, M_{age} = 33.8 y), three exclusions for all one key or speaking Chinese. Niffites were always initially indicated to be negative and Luupites as positive. The narrative and counternarrative, as well as AMP stimuli and factor analysis of the believability scale, are available in *SI Appendix, Supplementary Materials*. The measure was a difference score between Niffites and Luupites primes.

Study 3. Participants were $n = 405$ MTurk workers (52.6% male, M_{age} = 37.3 y). Eleven participants pressed only one key on the AMP and four indicated that they spoke Chinese. The final sample was $n = 390$. Full text of the source credibility manipulation is available in *SI Appendix, Supplementary Materials*.

Study 4. Participants were $n = 400$ MTurk workers (48.7% male, M_{age} = 35.1 y). Six submitted completion codes but did not complete all components of the study. Twenty-one pressed only one key on the AMP. Five indicated that they spoke Chinese. The final sample was $n = 368$.

Study 5. Participants were $n = 500$ Prolific Academic (PA) participants (40.4% male, M_{age} = 34 y). Twenty did not complete all components of the study. Twenty-five pressed only one key. The final sample was $n = 455$. The study was identical to Study 4, except that participants were first told that Kevin had been discovered to have been charged with child molestation and then were asked to wait 2.5 min before continuing to the next task in the study. A timer was presented during the imposed delay. After this, participants were provided with the source credibility information, and they proceeded to the second AMP. This study made use of a modified AMP protocol in which participants judged colorful paintings instead of Chinese pictographs and a more explicit warning about the influence of the primes was included in the instructions (see ref. 32).

Study 6. Participants were $n = 800$ PA participants (40.4% male, M_{age} = 34 y). Due to a server failure, the EPT responses from 67 participants were lost. Twenty-eight participants were excluded for excessive errors (>2.5 SDs above the mean; M = 8.79, SD = 13.99). The final sample was $n = 705$. Each trial of the EPT consisted of a cue (+; 500 ms), an ISI (500 ms), a prime [Kevin or a (single) neutral target; 200 ms], an ISI (20 ms), and a positive or negative target word. Participants' task was to categorize the target word as positive or negative as quickly as possible. Mean reaction times (RTs) for positive targets (40 trials, 20 per prime) were subtracted from mean RTs for negative targets (40 trials) for each prime type. The study was otherwise identical to Study 4, except that the learning paradigm had 20 trials instead of 50. More details of the procedure and trial exclusion criteria are available in *SI Appendix, Supplementary Materials*.

Study 7. Participants were $n = 500$ PA participants (56.5% male, M_{age} = 28 y). Thirteen did not complete all components of the study. Ten indicated that they spoke Chinese. Twenty pressed only one key. The final sample was $n = 457$. Materials are available in *SI Appendix, Supplementary Materials*.

1. Chokshi N (April 26, 2017) That wasn't Mark Twain: How a misquotation is born. *NY Times*. Available at https://www.nytimes.com/2017/04/26/books/famous-misquotations.html. Accessed September 17, 2018.
2. O'Toole G (2014) A lie can travel halfway around the world while the truth is putting on its shoes. Available at https://quoteinvestigator.com/2014/07/13/truth/. Accessed September 17, 2018.
3. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. *Science* 359:1146–1151.
4. McConnell AR, Rydell RJ (2014) The systems of evaluation model. *Dual-Process Theories of the Social Mind*, eds Sherman JW, Gawronski B, Trope Y (Guilford, New York), pp 204–217.
5. Gawronski B, Bodenhausen GV (2006) Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychol Bull* 132:692–731.
6. Deutsch R, Gawronski B, Strack F (2006) At the boundaries of automaticity: Negation as reflective operation. *J Pers Soc Psychol* 91:385–405.
7. Peters KR, Gawronski B (2011) Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Pers Soc Psychol Bull* 37: 557–569.
8. Moran T, Bar-Anan Y, Nosek BA (2017) The effect of the validity of co-occurrence on automatic and deliberate evaluations. *Eur J Soc Psychol* 47:708–723.
9. Moran T, Bar-Anan Y (2013) The effect of object-valence relations on automatic evaluation. *Cogn Emotion* 27:743–752.
10. Hu X, Gawronski B, Balas R (2017) Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations. *Pers Soc Psychol Bull* 43:17–32.
11. Cao J, Banaji MR (2016) The base rate principle and the fairness principle in social judgment. *Proc Natl Acad Sci USA* 113:7475–7480.
12. Cameron CD, Brown-Iannuzzi JL, Payne BK (2012) Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Pers Soc Psychol Rev* 16:330–350.
13. Kurdi B, et al. (December 13, 2018) Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *Am Psychol*, 10.1037/amp0000364.
14. Forscher PS, et al. (2019) A meta-analysis of procedures to change implicit measures. Open Science Framework. osf.io/awz2p.
15. Lai CK, et al. (2016) Reducing implicit racial preferences: II. Intervention effectiveness across time. *J Exp Psychol Gen* 145:1001–1016.
16. Arcuri L, et al. (2008) Predicting the vote: Implicit attitudes as predictors of the future behavior of decided and undecided voters. *Polit Psychol* 29:369–387.
17. Lundberg KB, Payne BK (2014) Decisions among the undecided: Implicit attitudes predict future voting behavior of undecided voters. *PLoS One* 9:e85680.
18. Gregg AP, Seibt B, Banaji MR (2006) Easier done than undone: Asymmetry in the malleability of implicit preferences. *J Pers Soc Psychol* 90:1–20.
19. Smith CT, De Houwer J, Nosek BA (2013) Consider the source: Persuasion of implicit evaluations is moderated by source credibility. *Pers Soc Psychol Bull* 39:193–205.
20. Rydell RJ, McConnell AR (2006) Understanding implicit and explicit attitude change: A systems of reasoning analysis. *J Pers Soc Psychol* 91:995–1008.
21. Rydell RJ, et al. (2007) Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *Eur J Soc Psychol* 37:867–878.
22. Cone J, Ferguson MJ (2015) He did what? The role of diagnosticity in revising implicit evaluations. *J Pers Soc Psychol* 108:37–57.
23. Cone J, Mann TC, Ferguson MJ (2017) Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. *Advances in Experimental Social Psychology*, ed Olson JM (Elsevier, Amsterdam), Vol 56, pp 131–199.
24. Mann TC, Ferguson MJ (2015) Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *J Pers Soc Psychol* 108:823–849.
25. Payne BK, Cheng CM, Govorun O, Stewart BD (2005) An inkblot for attitudes: Affect misattribution as implicit measurement. *J Pers Soc Psychol* 89:277–293.
26. Cone J, Flaharty K, Ferguson MJ (2019) Data from "Believability of evidence matters for correcting social impressions." Open Science Framework. Available at https://osf.io/3h2ww/. Deposited February 21, 2019.
27. Rydell RJ, McConnell AR, Mackie DM, Strain LM (2006) Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychol Sci* 17:954–958.
28. Kamins MA, Folkes VS, Perner L (1997) Consumer responses to rumors: Good news, bad news. *J Consum Psychol* 6:165–187.
29. Preacher KJ, Hayes AF (2008) Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods* 40:879–891.
30. Bar-Anan Y, Nosek BA (2012) Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Pers Soc Psychol Bull* 38:1194–1208.
31. Payne BK, et al. (2013) Intention invention and the affect misattribution procedure: Reply to Bar-Anan and Nosek (2012). *Pers Soc Psychol Bull* 39:375–386.
32. Mann TC, Cone J, Heggeseth B, Ferguson MJ (2019) Updating implicit impressions: New evidence on intentionality and the affect misattribution procedure. *J Pers Soc Psychol* 116:349–374.
33. Fazio RH, Sanbonmatsu DM, Powell MC, Kardes FR (1986) On the automatic activation of attitudes. *J Pers Soc Psychol* 50:229–238.
34. Van Dessel P, Ye Y, De Houwer J (2018) Changing deep-rooted implicit evaluation in the blink of an eye: Negative verbal information shifts automatic liking of Gandhi. *Soc Psychol Personal Sci* 10:266–273.
35. Origgi G (2017) *Reputation: What It Is and Why It Matters* (Princeton Univ Press, Princeton).
36. Gawronski B, Brannon SM, Bodenhausen GV (2016) The associative-propositional duality in the representation, formation, and expression of attitudes. *Reflective and Impulsive Determinants of Human Behavior*, eds Deutsch R, Gawronski B, Hofmann W (Routledge, New York), pp 103–118.
37. Houwer JD (2014) A propositional model of implicit evaluation. *Soc Personal Psychol Compass* 8:342–353.
38. Kunda Z (1990) The case for motivated reasoning. *Psychol Bull* 108:480–498.