IN PRESS AT *COLLABRA: PSYCHOLOGY*

Does Mud Really Stick? No Evidence for Continued Influence of Misinformation on Newly Formed Person Impressions

Amy J. Mickelberg[a], Bradley Walker[a], Ullrich K. H. Ecker[a,b], Piers D. L. Howe[c], Andrew Perfors[c] & Nicolas Fay[a]

[a]School of Psychological Science, University of Western Australia

[b]Public Policy Institute, University of Western Australia

[c]School of Psychological Sciences, University of Melbourne

Address correspondence to: Amy Mickelberg, School of Psychological Science (M304), University of Western Australia, 35 Stirling Hwy, Perth 6009, Australia. Email: amy.mickelberg@research.uwa.edu.au.

Abstract

Despite robust evidence that misinformation continues to influence event-related reasoning after a clear retraction, evidence for the continued influence of misinformation on person impressions is mixed. Across four experiments, we investigated the impact of person-related misinformation and its correction on dynamic (moment-to-moment) impression formation. Participants formed an impression of a protagonist, "John", based on a series of behaviour descriptions, including misinformation that was later retracted. Person impressions were recorded after the presentation of each behaviour description. As predicted, we found a strong effect of information valence on person impressions: negative misinformation had a greater impact on person impressions than positive misinformation (Experiments 1 and 2). Furthermore, in each experiment participants fully discounted the misinformation once retracted, regardless of whether the misinformation was negative or positive. This was true even when the other behaviour descriptions were congruent with (Experiment 2) or causally related to the retracted misinformation (Experiments 3 and 4). To test for generalization, Experiment 4 used a different misinformation statement; it again showed no evidence for the continued influence of retracted misinformation on person impressions. Our findings indicate that person-related misinformation can be effectively discounted following a clear retraction.

Keywords: Misinformation; continued influence effect; information valence; impression formation; person impressions

Does Mud Really Stick? No Evidence for Continued Influence of Misinformation on Newly Formed Person Impressions

In the digital information age, people are exposed to vast amounts of factual and false information. Consider the false allegation that Bill Gates was the architect of the COVID-19 pandemic, manufacturing the virus to profit from a future vaccine (Reuters, 2020), or the false claim that Hillary Clinton was involved in a child paedophile ring at a pizza parlour (O'Rourke, 2016). Of course, false allegations are also often levelled against less-known people both on social media and in offline contexts. This raises the question: Even when discredited, do false allegations continue to influence person impressions? Or, as the saying goes, 'does mud stick'? Despite the threat misinformation poses to a person's reputation, there is a limited understanding of the factors that moderate misinformation's influence on impression formation. The present study examined this issue, testing the extent to which misinformation and subsequent corrections dynamically shape person impressions when these are first formed.

Continued Influence Effect

Misinformation penetrates deep into social networks, spreading further and faster than true information (Juul & Ugander, 2021; Vosoughi et al., 2018). This pernicious impact has inspired a wealth of psychological research into misinformation and its effects. Perhaps most worrying is the finding that misinformation often continues to influence people's judgements, reasoning, and decision making even after it has been retracted or corrected. This phenomenon is known as the continued influence effect (CIE; Johnson & Seifert, 1994; Wilkes & Leatherbarrow, 1988) and it has been demonstrated across a large body of experimental work (for meta-analyses, see Chan et al., 2017; Walter & Murphy, 2018; Walter & Tukachinsky, 2020; for a review, see Ecker et al., 2022).

Research investigating the CIE has focused on event misinformation. In a common CIE paradigm, participants read a narrative text that describes a fictitious event, which contains a critical piece of (typically causal) information (e.g., that negligent storage of flammable materials started a warehouse fire); this misinformation is later retracted (e.g., it is clarified that no flammable materials were present) or not retracted. Although a retraction reduces the number of references to the misinformation participants make in response to inferential-reasoning questions about the event (e.g., "Why was there so much black smoke?"), participants often continue to rely on the retracted misinformation (e.g., Ecker et al., 2010, 2015). Similar effects are observed using variations to this paradigm with shorter statements (see Brydges et al., 2020; Gordon et al., 2017), implied rather than explicitly stated misinformation (e.g., Rich & Zaragoza, 2016; Tay et al., 2022), and real-world misinformation (e.g., the unfounded link between vaccination and autism; see Swire et al., 2021), adding to the robustness of the CIE for event misinformation. These findings align with the broader belief-perseverance literature, which shows that original beliefs are often retained despite disconfirming evidence (Anderson et al., 1980; Asch, 1946; Downey & Christensen, 2006; Green & Donahue, 2011).

One theoretical account of the CIE relates to the way people process event information. When learning about an unfolding event, people build mental models, which are characterized by their temporal sequence and causal structure (Bower & Morrow, 1990; van Oostendorp & Bonebakker, 1999). When misinformation is retracted, this can present a threat to model coherence because the retraction leaves a "gap" in the model. Under these circumstances, the correction may not be integrated into the model, and during reasoning preference may be given to an incorrect but complete mental model rather than a correct but incomplete one (Brydges et al., 2018; Ecker, Lewandowsky, Chang, et al., 2014; Johnson-Laird, 2012; Kendeou et al., 2019; Schul & Mayo, 2014). In support of this theoretical account, it has been found that

misinformation is more resistant to correction if it is central to an event narrative, providing a causal explanation for, or otherwise enhancing comprehension of, the event (Hamby et al., 2020; Johnson & Seifert, 1994; Rich & Zaragoza, 2020; van Oostendorp & Bonebakker, 1999). These findings suggest that the CIE may arise as a consequence of the way people represent event information in memory.

CIE in Impression Formation

In contrast to event-related reasoning, the empirical evidence for a CIE in person-related reasoning (i.e., impression formation) is mixed. On the one hand, there is substantial evidence that discredited misinformation can continue to inform person impressions. Examples include inadmissible evidence affecting jury decisions (see Steblay et al., 2006, for a review), continued stigma towards victims of miscarriages of justice (Brooks & Greenberg, 2021; Clow & Leach, 2015), and discounted rumours swaying voter preferences (Jardina & Traugott, 2019; Weeks & Garrett, 2014). In an experiment investigating misinformation and political figures, Thorson (2016) found that misinformation about a (fictional) candidate accepting donations from a convicted felon led participants to evaluate the candidate more negatively compared to a no-misinformation control condition, even when the misinformation had been corrected (although primarily if the political candidate was affiliated with the participant's non-preferred political party; see also Bullock, 2007). These findings point to the presence of a CIE in impression formation, although this may depend on specific task features (i.e., if an explicit judgement is required) and whether there is motivation to keep the original information active (see Mensink & Rapp, 2011; Rapp & Kendeou, 2007, 2009).

However, other studies have failed to observe a CIE in impression formation (e.g., Cobb et al., 2013; De keersmaecker & Roets, 2017; Ecker & Rodricks, 2020). When investigating the impact of discredited positive misinformation, Cobb et al. (2013) presented participants with a

mock news story which favoured a political candidate. When the story was later corrected, participants reported more *negative* impressions of the candidate compared to those who had not received the positive misinformation; in other words, this study not only found no CIE, but even an overcorrection effect. This demonstrates that corrections can be fully effective in impression formation. Ecker and Rodricks (2020) investigated the effects of discredited negative misinformation on person impressions. Their study featured observations about a fictitious student named "John". Participants received descriptions of behaviours that John had engaged in; these were mainly behaviours of neutral valence (e.g., "John gave a presentation to his class"), but in some conditions a behaviour of negative valence was included ("John slapped his girlfriend") that was subsequently retracted or not retracted. Results showed participants in the retraction condition fully discounted the discredited misinformation in their person impressions. Taken together, this suggests people can fully discredit retracted person-related misinformation.[1]

The mixed findings for a CIE in impression formation may be due, in part, to the different methodological approaches used. In some studies participants were presented with misinformation targeting political candidates (Cobb et al., 2013; Thorson, 2016) while in other studies the misinformation related to people with no political affiliation (De keersmaecker & Roets, 2017; Ecker & Rodricks, 2020). Studies investigating the CIE in impression formation have also differed in the type of allegation and valence of the misinformation (positive and negative; Cobb et al., 2013; Ecker & Rodricks, 2020). These factors may moderate the CIE in impression formation, particularly when the target type (politician vs. private citizen) and type of

[1] In providing evidence for and against the CIE in impression formation, we refer to studies that have measured person impressions explicitly (i.e., direct person judgements). It should be noted that in the impression updating literature, implicit, indirect judgements tend to be more resistant to updating (see Rydell et al., 2007), although even implicit impressions can be influenced by newly provided information to the extent that it is diagnostic (Cone & Ferguson, 2015), believable (Cone et al., 2019) and subject to reinterpretation (Mann & Ferguson, 2015; Okten et al., 2019).

misinformation (positive vs. negative) coheres with pre-existing knowledge (e.g., see Cappella et al., 2015; Walter & Tukachinsky, 2020). As a result, task characteristics may influence the robustness of the CIE in impression formation, and may help explain the mixed results. Thus, it is important to systematically investigate these factors and their influence on the CIE in impression formation.

To better understand the mixed findings, our attention turns to the robustness of the CIE as it relates to person impressions, and the conditions under which it can be detected. Given that impression formation is a dynamic process where people continually update their impressions to integrate new information (Asch, 1946; Kashima & Kerekes, 1994), a CIE might be present at the time of correction (i.e., a primacy effect; see Park, 1986) but then fade as more and more information is received and the impressions continue to be updated. Demonstration of a short-lived CIE could contribute to the mixed evidence if the CIE is generally tested too late to enable detection in an impression-formation task (i.e., at the end of the task; see Cappella et al., 2018; Connor Desai & Reimers, 2019). This possibility is tested here using a dynamic (moment-to-moment) person-impression rating measure that is taken across the entire task.

Another factor that has not been systematically investigated is the valence of the misinformation, and the extent to which it might moderate a CIE in impression formation. While some studies have investigated positively-valenced misinformation and its impact on person impressions (e.g., Cobb et al., 2013), and others have considered the effect of negatively-valenced misinformation on person impressions (e.g., Ecker & Rodricks, 2020; Thorson, 2016), no studies (to the best of our knowledge) have compared both. As a result, we do not know the role that misinformation valence plays in this context.[2] Valence has been found to impact

[2] We note that in the traditional event-related CIE literature, there have been suggestions that positive misinformation may be easier to correct than negative misinformation, but hitherto this notion has not been substantiated (see Chang et al., 2019; Walter & Tukachinsky, 2020).

impression formation in general, with negative information found to be more influential than positive information (i.e., negativity bias; Fiske, 1980; Skowronski & Carlston, 1989; see Rusconi et al., 2020 for a review). It is possible that this negativity bias extends to the misinformation context. To determine the role of misinformation valence, the effect of positively- and negatively-valenced misinformation on person impressions was tested in Experiments 1 and 2.

Another explanation for the inconsistent findings may relate to misinformation coherence, specifically, the extent to which the misinformation coheres with prior information. When information about a person takes the form of unrelated traits or behaviours presented within a randomly ordered list, as is typical of impression-formation research (e.g., Carlston & Smith, 1996; Srull & Wyer, 1989), model coherence may be enhanced by retracting an "outlying" trait description or behaviour that is inconsistent with the initial person model (Fiske & Linville, 1980). For example, in the study by Ecker and Rodricks (2020), John slapping his girlfriend was the only negative behaviour mentioned; when this was retracted, updating of the person model may have been readily and effectively executed because it enhanced overall model coherence (also see Mende-Siedlecki et al., 2013; Srull & Wyer, 1989). In contrast, when the retracted misinformation coheres with other information, removing it may threaten model coherence, thereby promoting continued reliance on the retracted misinformation (e.g., Thorson, 2016). This possibility is consistent with the role that misinformation coherence has been theorised to take in event-related reasoning; here, retracted event-related misinformation which coheres with other information continues to be relied upon because it helps to maintain the causal structure of the mental model of the event (Ecker et al., 2011; Johnson & Seifert, 1994). To test this explanation, coherence-building elements were added in Experiments 2–4.

The Present Study

To investigate the impact of person-related misinformation and its correction on impression formation, we combined the CIE paradigm—which provides participants with misinformation that is or is not subsequently retracted—with a dynamic person-impression rating task. This task records trial-to-trial impression updates, thus measuring how the behaviour descriptions provided to participants—including a misinformation item and its retraction—influence person impressions over time. To manipulate participants' impressions of a target person, we selected behaviour statements from a pool of pre-rated statements (Mickelberg et al., 2022) based on their morality (Brambilla et al., 2021) and believability (Cone et al., 2019), two dimensions known to be important in impression formation. We expected that misinformation would affect person impressions, and that a retraction would reverse this effect, at least partially; by measuring impressions after each piece of information was received, we anticipated that we would be able to detect any CIE even if this effect was short-lived. In addition, if lack of model coherence affected the CIE (Cobb et al., 2013; Ecker & Rodricks, 2020), then adding coherence-building elements to the materials should promote a CIE. Finally, to clarify the role of valence, we compared both positive and negative misinformation.

Evidence for a CIE in impression formation was examined over four experiments. In each, participants were presented with a sequence of 27 behaviour statements that related to a fictitious person named "John". The statements described behaviours John had engaged in and included a target misinformation statement that was later retracted or not retracted. Participants were told the information came from several of John's acquaintances. Person impressions were recorded after the presentation of each behaviour statement. In Experiment 1, behaviour statements included a series of neutral filler statements (e.g., "John had ordered his favourite dish from a Chinese restaurant") and a target statement whose valence was either negative

("John had an affair with his best friend's wife"), positive ("John jumped off a boat to save a drowning friend even though this put John's own life at risk") or neutral ("John went to a fancy restaurant but couldn't pronounce the items on the menu"). Statements were presented as stand-alone items in the manner typical of impression-formation research (Cone & Ferguson, 2015; Gregg et al., 2006; Kerpelman & Himmelfarb, 1971). Experiments 2–4 extended Experiment 1 by replacing some of the neutral filler statements with coherence-building elements that we hypothesised would make misinformation more resistant to correction: Experiment 2 included behaviour statements that were congruent with the target behaviour, and Experiments 3 and 4 included behaviour statements that were causally related to the target behaviour. Experiments 1–3 used the same target statements, whereas Experiment 4 included new target materials. To foreshadow, no evidence for a CIE in impression formation was obtained in any of the experiments.

Experiment 1

Experiment 1 presented information about the protagonist "John" in five conditions. The first four conditions varied according to the valence of the target statement (negative or positive) and whether the target statement was later retracted or not retracted. The fifth condition used a neutral target statement and served as a baseline. The experiment thus used a 2 × 2 between-subjects design (negative retraction; positive retraction; negative no-retraction; positive no-retraction), plus a neutral control.

Method

All experiments were approved by the University of Western Australia's Human Research Ethics Office. Participants received an ethics-approved information sheet and provided informed consent (by clicking 'next' to proceed to the study). The experiments were conducted

in accordance with Australia's National Statement on Ethical Conduct in Human Research (NHMRC, 2018).

Participants

A sample of 500 U.S.-based participants (282 male, 213 female, 4 other, 1 preferred not to say) were recruited through the online crowd-sourcing platform Prolific (https://prolific.com); age range was 18–84 years (M = 32.42, SD = 11.31). Participants were randomly assigned to one of the five conditions (n = 100 per condition).

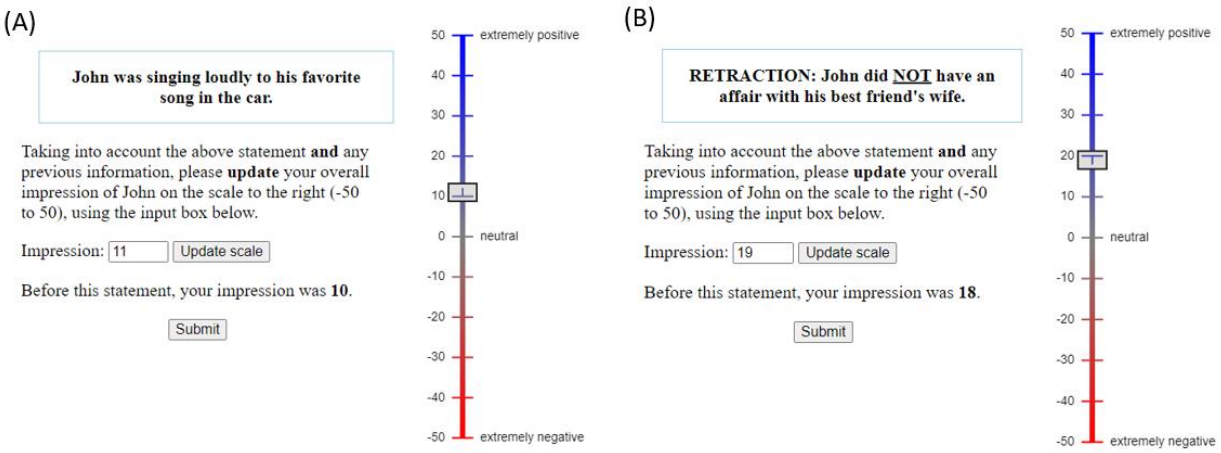
Materials

Impression-Formation Task

Participants viewed 27 behaviour statements concerning a person named John, presented one-by-one. After reading each statement, participants rated their impression of John using an impression rating scale ranging from −50 (*extremely negative*) to 50 (*extremely positive*), with 0 indicating *neutral*. A marker on the scale indicated the participant's current rating (see Figure 1).

Figure 1

Impression-Formation Task and Dynamic Impression Rating Scale



Note. (A) shows a neutral filler statement; (B) shows the retraction of the negative target statement.

The impression-formation task had two phases: (1) presentation of initial information, and (2) presentation of updated information. The first phase encompassed trials 1–17. In this phase, participants were told that the study investigated how people form impressions of others. Participants were then informed they would be given information about a person named “John”. They were told this information had been collected from four of his acquaintances, who provided examples of his behaviour that they had observed over the past few months. A 20 s delay was imposed to allow time to read the instructions before participants could proceed to the first behaviour statement. Over the 17 trials, participants were presented with 16 neutral filler statements, presented in a random order, and a target statement presented at trial 13. The target statement depended on condition, and was either negative, positive, or neutral (see preceding section for the statements used). At the start of the second phase, participants were told that John’s acquaintances were interviewed a second time, to verify the initial information they provided and add new information about John. During the second phase (trials 18–27), ten update statements were provided, in which six new neutral filler statements were presented, two neutral filler statements from the first phase were confirmed (e.g., “CONFIRMED: John was running late, so he drove to work rather than taking the bus”), and two statements from the first phase were retracted. In the retraction and neutral conditions, this included retraction of a neutral filler statement and the target statement (e.g., “RETRACTION: John did NOT have an affair with his best friend’s wife”); in the no-retraction conditions, two filler statements were retracted instead. The presentation order of the updated statements was randomised, except that the target-statement retraction always occurred at trial 19.

The behaviour statements were selected from a corpus of pre-tested stimuli (Mickelberg et al., 2022). Each statement was associated with a morality score ranging from −4 (*very morally bad*) to 4 (*very morally good*), and a believability score ranging from 0 (*not believable*) to 8

(*very believable*). We selected statements with high believability scores (> 6) and morality scores appropriate for neutral filler statements or the neutral target (i.e., close to 0) or strongly valenced (close to ±4) for the positive and negative target statements. To illustrate, the neutral target statement "John went to a fancy restaurant but couldn't pronounce the items on the menu" was associated with morality = 0.06, believability = 6.75. By contrast, the negative target statement was strongly negative ("John had an affair with his best friend's wife"; morality = −3.40, believability = 6.84) and the positive target statement was strongly positive ("John jumped off a boat to save a drowning friend even though this put John's own life at risk"; morality = 3.55, believability = 6.79). See Supplementary Materials for the full list of behaviour statements selected across experiments.

Reysen Likability Scale

The Reysen likability scale (Reysen, 2005) measured the likability of the protagonist John. The scale consists of 11 items (e.g., "This person is likeable") rated on a Likert scale ranging from 1 (*very strongly disagree*) to 7 (*very strongly agree*; see Supplementary Materials for the full scale). The scale was previously used by Ecker and Rodricks (2020) to test for a CIE in impression formation, and has been associated with strong internal consistency reliability (Cronbach's α = .88; Reysen, 2005). In the present experiments, reliability ranged from α = .88 to α = .95. Across experiments, likability scores were moderately positively correlated with impression ratings at trial 27 (r = .53 to r = .69) but not strongly enough to be considered collinear.

Recognition Test

A recognition test was included to measure the extent to which participants attended to and encoded the behaviour statements. Eleven multiple-choice questions, each with three to four response options, tested recognition of the initial and updated behaviour statements (e.g., "What

instrument did John learn to play as a child?"—"The piano", "The drums", "The violin"; see Supplementary Materials for the full list of questions).

Procedure

The experiment was performed online. After reading information about the task and providing informed consent, participants completed the impression-formation task, followed by the likability scale and recognition test, and were then debriefed and thanked for their participation. The median completion time was 8 minutes. Participants were paid £1.00 (approximately US$1.25) upon completion of the study.

Results

The data were analysed using linear regression models. All analyses were performed and all figures were created in R (v4.0.3; R Core Team, 2021). Regression models were estimated using the *lm* function of base R. Figures were created using *ggplot2* (v3.3.5; Wickham, 2016). The data, R script, and supplementary information associated with Experiments 1–4 are available at https://osf.io/ymuap.

Performance on the recognition test was high (M = 88.61%, SD = 13.13%). Only 12 participants scored below 50%; excluding them did not qualitatively affect the results, so analyses of the full-sample data are reported. A manipulation check was conducted to test that target statements at trial 13 affected impressions. Dependent variables for investigating continued influence were impression ratings immediately post-retraction (trial 19) and after a delay (trial 27). The likability scores gave another measure of participants' impressions after a delay. Secondary analyses (for trials 1–12 and 14–18) can be found in Supplementary Materials.

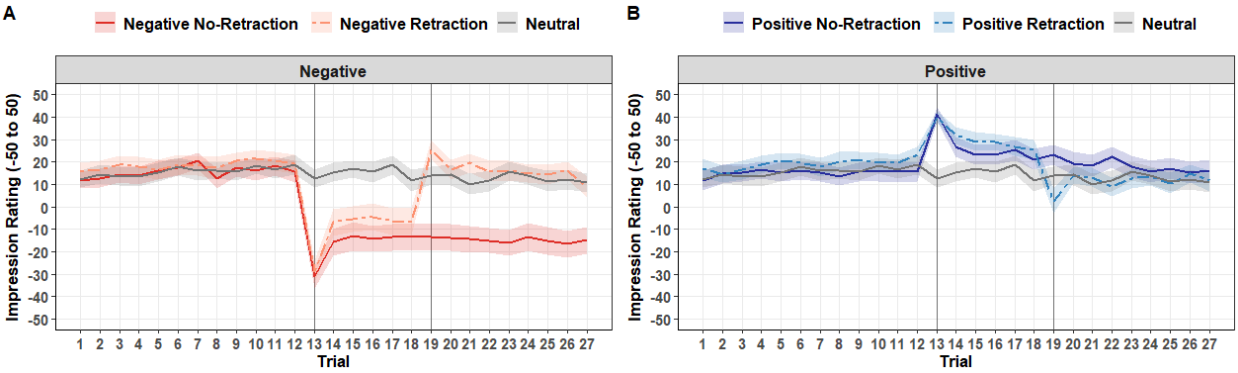
Manipulation Check

As illustrated in Figure 2, participants in all conditions reported a somewhat positive impression of John from the outset, and this gradually improved over trials 1–12 (see

Supplementary Materials). The target statement was then presented at trial 13. To ensure the target-statement manipulation was successful (i.e., target statements affected impressions), we tested the impact of the positive and negative target statements on person impressions at trial 13. First, however, we ascertained that at trial 13, there was no evidence of a statistical difference (based on random noise) between the negative-retraction and negative no-retraction conditions (p = .328), or between the positive-retraction and positive no-retraction conditions (p = .626). For the purpose of comparing to the neutral condition, the conditions were thus collapsed into a combined negative or positive condition according to target-statement valence.

Figure 2

Mean Impression Ratings Across Trials of Experiment 1 for Negative (A), Positive (B), and Neutral Target Statements



Note. The mean impression ratings for the neutral condition are duplicated in panels A and B. The shaded areas are 95% confidence intervals. The vertical line at trial 13 indicates target-statement presentation; the vertical line at trial 19 indicates target-statement retraction (in the retraction conditions).

Impression ratings at trial 13 were entered into a linear regression, with target-statement valence as the predictor (negative; positive; neutral). This showed that the negative conditions returned lower impression ratings than the neutral control (i.e., impressions were more negative when participants had received the negative target statement; $\beta = -42.43$, $SE = 2.58$, $t = -16.43$, $p < .001$), and the positive conditions returned higher impression ratings than the neutral control

(i.e., impressions were more positive when participants had received the positive target statement; β = 26.90, *SE* = 2.58, *t* = 10.42, *p* < .001). The difference in beta-weight magnitude (β = −42.43 compared to β = 26.90) suggests that participants' impressions of John were more strongly affected by the negative target statement than the positive target statement. This was confirmed by comparing the absolute difference in impression ratings from trial 12 to trial 13 (negative: 50.05 > positive: 26.09; *t*(198) = 7.10, *p* < .001).

Between trial 13 (the target statement) and trial 19 (the target-statement retraction, in the retraction conditions), the effect of the negative target statement weakened over trials but impression ratings remained lower than in the neutral condition. The effect of the positive target statement also weakened between trials 13 and 19 but impression ratings remained higher than in the neutral condition. The differences to neutral were statistically significant (see Supplementary Materials).

Tests for Continued Influence

Evidence for a CIE would be observed if the target statements continued to influence person impressions after being retracted. We tested for a CIE at two time-points: immediately following the target-statement retraction (trial 19) and after a delay (trial 27).

Immediate Test

The trial-19 impression ratings were entered into linear regressions with condition as a predictor (retraction; no-retraction; neutral), with separate regression models for the negative and positive conditions. For negative misinformation, the model was significant, *F*(2, 297) = 66.37, *p* < .001, R^2 = .31. In the negative no-retraction condition, impression ratings were lower than in the neutral control condition, β = −26.78, *SE* = 3.60, *t* = −7.43, *p* < .001, indicating that the effect of the negative target statement persisted at trial 19. When the negative target statement was retracted in the negative-retraction condition, impression ratings were higher than in the negative

no-retraction condition, β = 38.13, *SE* = 3.69, *t* = 10.34, *p* < .001, and also higher than in the neutral control condition, β = 11.35, *SE* = 2.84, *t* = 4.00, *p* < .001. This indicates the retraction of the negative target statement was fully effective; in fact, participants overcorrected at immediate test.

For positive misinformation, the model was also significant, $F(2, 297) = 18.68$, $p < .001$, $R^2 = .11$. In the positive no-retraction condition, impression ratings were higher than in the neutral control condition, β = 8.57, *SE* = 2.95, *t* = 2.91, *p* = .004, indicating that the effect of the positive target statement persisted at trial 19. Note that the effect of the positive information on person impressions was not as strong as the negative information on this trial (β = 8.57 compared to β = −26.78). This was confirmed by comparing the absolute difference in impression ratings between trials 12 and 19 for the negative and positive no-retraction conditions (negative: 38.48 > positive: 19.27; $t(198) = 5.89$, $p < .001$). When the positive target statement was retracted in the positive-retraction condition, impression ratings were lower than in the positive no-retraction condition, β = −20.30, *SE* = 3.60, *t* = −5.64, *p* < .001, and also lower than in the neutral control condition, β = −11.73, *SE* = 3.42, *t* = −3.43, *p* < .001. This indicates the retraction of the positive target statement was fully effective, and that again participants overcorrected at immediate test.

Delayed Test

Next, the same set of analyses was conducted on the trial-27 impression ratings. For negative misinformation, the model was significant, $F(2, 297) = 30.82$, $p < .001$, $R^2 = .17$. In the negative no-retraction condition, impression ratings were lower than in the neutral control condition, β = −24.76, *SE* = 3.67, *t* = −6.74, *p* < .001, indicating that the effect of the negative target statement persisted at trial 27. For the negative-retraction condition, impression ratings were higher than those in the negative no-retraction condition, β = 23.30, *SE* = 3.81, *t* = 6.13,

$p < .001$, but were comparable to control, $\beta = -1.43$, $SE = 3.10$, $t = -0.46$, $p = .646$. This indicates that while the retraction of the negative target statement was still effective after a delay, there was no longer evidence of an overcorrection.

For positive misinformation, the model was not significant, $F(2, 297) = 1.41$, $p = .245$, $R^2 = .01$, indicating there was no statistical evidence of a difference between conditions. Thus, there was no evidence for a CIE (the positive-retraction and neutral conditions did not differ); however, no strong conclusions can be drawn from this because the effect of the positive misinformation appeared to have "worn off" regardless of the retraction (i.e., the positive no-retraction and neutral conditions did not differ).

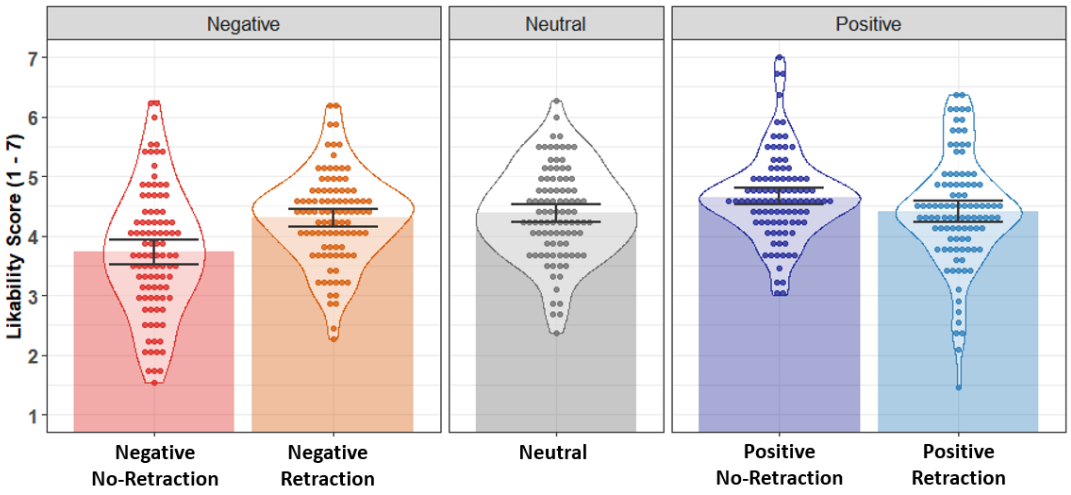
Likability Scores

Scores on the likability scale were consistent with the results of the impression-formation task (see Figure 3). Likability scores were entered into a linear regression with condition as the predictor (retraction; no-retraction; neutral), and with separate models for the negative and positive conditions. For negative misinformation, the model was significant, $F(2, 297) = 18.15$, $p < .001$, $R^2 = .11$. Likability was lower in the negative no-retraction condition compared to the neutral control condition, $\beta = -0.71$, $SE = 0.13$, $t = -5.31$, $p < .001$, and the negative-retraction condition, $\beta = -0.60$, $SE = 0.13$, $t = -4.60$, $p < .001$. There was no evidence of a statistical difference between the negative-retraction and neutral conditions, $\beta = -0.10$, $SE = 0.12$, $t = -0.91$, $p = .364$. This again indicates the retraction of negative misinformation was fully effective (i.e., no CIE).

For positive misinformation, the model was not significant, $F(2, 297) = 2.30$, $p = .102$, $R^2 = .02$, indicating there was no statistical evidence of a difference between conditions. This indicates there was neither an impact of the misinformation, nor a CIE.

Figure 3

Mean Likability Scores Across Conditions in Experiment 1



Note. Bars indicate condition means; error bars are 95% bootstrapped confidence intervals; data points indicate jittered participant mean scores; violins provide distributional information.

Discussion

Experiment 1 tested for a CIE in impression formation and whether it was moderated by misinformation valence (i.e., negative vs. positive). Results showed that, first of all, negative misinformation had a strong and immediate negative impact on person impressions, and (when not retracted) the impact endured after a delay. Positive misinformation had a weaker immediate impact, and the impact did not endure after a delay. Once retracted, negative and positive misinformation was fully discounted, leaving no evidence of a continued impact on person impressions. For negative misinformation, an overcorrection was observed immediately following the retraction, such that impression ratings became more positive than those in the neutral control condition. Person impressions then returned to the neutral baseline after a delay. A temporary overcorrection was also observed immediately following the retraction of the positive misinformation (as in Cobb et al., 2013). The immediate overcorrection of person

impressions demonstrates the effectiveness of a retraction in correcting person-related misinformation.

Prior to a retraction, our findings show that negative misinformation is given greater weight than positive misinformation. This suggests that person impressions can be influenced more by a single negative statement than by a single positive statement (despite the statements being close in extremity according to pre-ratings; see Mickelberg et al., 2022). This demonstrates an information valence effect, adding to the robustness of previous findings of a negativity bias in impression formation (Rozin & Royzman, 2001; Skowronski & Carlston, 1989; Rusconi et al., 2020; see also Mensink & Rapp, 2011).

Our finding that people can fully discount retracted misinformation when making person impression judgements supports some previous studies (Cobb et al., 2013; Ecker & Rodricks, 2020) but contrasts with CIE research in general, which typically finds that there is still continued reliance on corrected misinformation (Chan et al., 2017; Ecker et al., 2022; Walter & Tukachinsky, 2020). A possible explanation for this is that the misinformation used in Experiment 1 may not have adequately cohered with the other person information provided, and this may have made it easier for participants to fully discount the misinformation. In other words, the misinformation used in Experiment 1 did not provide a causal explanation for an event (as is typical in CIE studies), nor did it serve to strengthen mental-model coherence due to it lacking consistency with the other behavioural information presented. Thus it could be that the retraction did not present a threat to model coherence, and may not have created a "gap" in the participants' mental model of the protagonist (Ecker et al., 2011). Consequently, there was no reason to rely on the misinformation post-retraction to maintain model coherence, explaining the absence of a CIE.

To test this explanation, Experiment 2 introduced coherence-building elements in the form of behaviour statements that were congruent with (i.e., conceptually related to) the target statements, to promote the integration of the misinformation with the other protagonist information. We hypothesised that improved congruency between the general information and misinformation about John would give rise to a CIE.

Experiment 2

Experiment 2 included behaviour statements that are congruent with the target statement. For instance, in the conditions with the negative target statement "John had an affair with his best friend's wife", the congruent statements included "John cheated in a card game" (also suggesting a poor moral character) and "John and his best friend had a fight" (consistent with John having an affair with the friend's wife). We reasoned that including the congruent statements would promote the integration of the target statement within participants' mental model of John.

Method

Experiment 2 used the same design as Experiment 1, but included six behaviour statements that were congruent with the target statement. There were four experimental conditions as in Experiment 1 (negative retraction; positive retraction; negative no-retraction; positive no-retraction), but Experiment 2 had two neutral conditions. One (the negative-congruent neutral condition) had a neutral target statement and the same congruent information as the negative conditions, to serve as the baseline for the negative-retraction and negative no-retraction conditions. The other (the positive-congruent neutral condition) had a neutral target statement and the same congruent information as the positive conditions, to serve as the baseline for the positive-retraction and positive no-retraction conditions.

Participants

A convenience sample of 600 first-year undergraduate students from the University of Western Australia (438 females, 155 males, 3 other, 4 preferred not to say) aged 16–62 years (M = 20.13, SD = 4.88) were recruited. Participants were randomly assigned to one of the six conditions (n = 100 per condition).

Materials

Impression-Formation Task

The task was identical to that used in Experiment 1, but six of the neutral filler statements presented in trials 8, 9, 10, 11, 15, and 17 were replaced with behaviour statements that were either valence-congruent or meaning-congruent with the target statement.

Valence-Congruent Statements. Two behaviour statements that were valence-congruent with the target statement were presented prior to the target statement (at trials 8 and 10). Valence-congruent statements were selected from the same corpus of behaviour stimuli (Mickelberg et al., 2022). The valence-congruent statements were matched according to the valence of the target statement (positive or negative) but were selected to be milder in valence (i.e., morality close to ±2) and similarly high in believability (> 6). Two negative behaviour statements were selected for the negative conditions ("John cheated in a card game while playing with a group of his friends"; morality = −1.91, believability = 6.71; and "John found a wallet with $50 in it, took the money out and left the wallet on the floor"; morality = −2.35; believability = 6.81). Two positive behaviour statements were used for the positive conditions ("John shaved his head when he found out his partner had cancer and required radiation therapy;" morality = 2.46, believability = 7.23; and "John volunteers at a dog refuge, walking the dogs and cleaning their kennels once a week"; morality = 2.67, believability = 7.04).

Meaning-Congruent Statements. Four meaning-congruent statements were generated by the authors, designed to align with the (negative and positive) target statements but to be neutral in morality. The two meaning-congruent statements that aligned with the negative target statement (that John had an affair with his best friend's wife) were "John and his best friend had a fight" and "John went out for a coffee with his best friend's wife". The two meaning-congruent statements that aligned with the positive target statement (that John saved a friend from drowning) were "John attended first aid training at the local sailing club" and "John went to a Surf Life Saving awards dinner". Two statements (one positive, one negative) were introduced prior to the presentation of the target statement (at trials 9 and 11) and two (one positive, one negative) were presented after the target statement (at trials 15 and 17). It was anticipated that these statements would be interpreted as neutral fillers by participants who did not receive the related target statements, so for simplicity we gave the same four meaning-congruent statements in every condition (expecting only two to be actually meaning-congruent in a given condition, or none in the neutral conditions). The presentation order of the four statements was counterbalanced across participants.

Procedure

The procedure was identical to Experiment 1. Median completion time was 10 minutes. Participants received partial course credit upon completion of the study.

Results

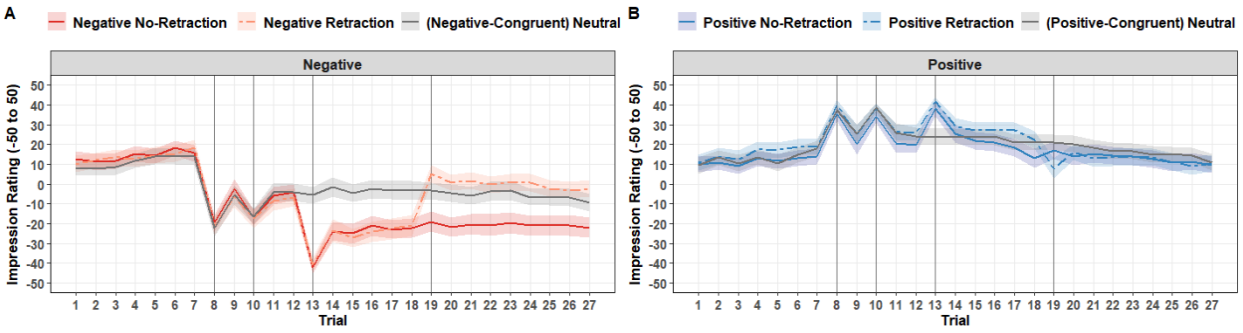
Performance on the recognition test was high (M = 93.73%, SD = 8.99%). Only 4 participants scored below 50%; excluding them did not qualitatively affect the results so analyses of the full-sample data are reported.

Manipulation Check

As illustrated in Figure 4, prior to the presentation of valence-congruent information at trial 8, participants in all conditions reported a positive impression of John from the outset, and this gradually improved over trials 1 to 7 (see Supplementary Materials). In the negative conditions the valence-congruent behaviour statements at trials 8 and 10 caused the impression ratings to become more negative over trials 8 to 12. By contrast, in the positive conditions the valence-congruent statements at trials 8 and 10 caused the impression ratings to become more positive over trials 8 to 12 (see Supplementary Materials).

Figure 4

Mean Impression Ratings Across Trials of Experiment 2 for Negative (A), Positive (B), and Neutral Target Statements



Note. The shaded areas are 95% confidence intervals. The vertical lines at trials 8 and 10 indicate valence-congruent-statement presentations. The vertical line at trial 13 indicates target-statement presentation and the vertical line at trial 19 indicates target-statement retraction (in the retraction conditions). The meaning-congruent statements were presented at trials 9, 11, 15, and 17.

The target statement was then presented at trial 13. To ensure the target-statement manipulation was successful, we tested the impact of the positive and negative target statements on person impressions at trial 13. Again, we first checked that at trial 13, there was no evidence of a statistical difference between the negative-retraction and negative no-retraction conditions (p = .406) or the positive-retraction and positive no-retraction conditions (p = .052); thus, the

conditions were collapsed into a combined negative or positive condition according to target-statement valence, for the purpose of comparing to the (negative-congruent and positive-congruent) neutral conditions.

Impression ratings at trial 13 were entered into a linear regression, with target-statement valence as a predictor and separate models for the negative conditions (negative; negative-congruent neutral) and positive conditions (positive; positive-congruent neutral). This showed that the negative conditions returned lower impression ratings than the (negative-congruent) neutral condition (i.e., person impressions were more negative when participants had received the negative target statement; $\beta = -35.72$, $SE = 2.20$, $t = -16.24$, $p < .001$), and the positive conditions returned higher impression ratings than the (positive-congruent) neutral condition (i.e., ratings became more positive when participants had received the positive target statement; $\beta = 15.90$, $SE = 2.10$, $t = 7.58$, $p < .001$). The difference in beta-weight magnitude ($\beta = -35.72$ vs. $\beta = 15.90$), suggests that participants' impressions of John were more strongly affected by the negative than the positive target statement. This was confirmed by comparing the absolute difference in impression ratings from trial 12 to trial 13 (negative: 38.77 > positive: 19.89; $t(198) = 6.82$, $p < .001$).

Between trial 13 (the target statement) and trial 19 (the target-statement retraction, in the retraction conditions), the effect of the negative target statement weakened over trials, but ratings remained lower than in the (negative-congruent) neutral condition (the difference to the negative-congruent neutral condition was statistically significant). There was no effect of the positive target statement over trials 13 and 19 (the difference to the positive-congruent neutral condition was not significant; see Supplementary Materials).

Tests for Continued Influence

The CIE was again tested at two time-points: immediately following the target-statement retraction (trial 19) and after a delay (trial 27).

Immediate Test

The trial-19 impression ratings were entered into linear regressions with condition as a predictor (retraction, no-retraction, neutral), with separate models run for the negative and positive conditions. For negative misinformation, the model was significant, $F(2, 297) = 29.30$, $p < .001$, $R^2 = .16$. In the negative no-retraction condition, impression ratings were lower than in the (negative-congruent) neutral condition, $\beta = -15.82$, $SE = 3.33$, $t = -4.76$, $p < .001$, indicating that the effect of the negative target statement persisted at trial 19. When the negative target statement was retracted in the negative-retraction condition, impression ratings were higher than in the negative no-retraction condition, $\beta = 24.93$, $SE = 3.44$, $t = 7.25$, $p < .001$, and also higher than in the (negative-congruent) neutral condition, $\beta = 9.11$, $SE = 3.12$, $t = 2.92$, $p = .004$. This indicates the retraction of the negative target statement was fully effective; in fact, participants overcorrected at the immediate test.

For positive misinformation, the model was significant, $F(2, 297) = 8.47$, $p < .001$, $R^2 = .05$. However, there was no evidence of a statistical difference between the positive no-retraction and (positive-congruent) neutral conditions, $\beta = -4.06$, $SE = 3.03$, $t = -1.34$, $p = .182$, indicating that the effect of the positive target statement was not observable at trial 19. Note that the effect of the positive information again was not as strong as the negative information by this trial ($\beta = -15.82$ vs. $\beta = -4.06$), indicating that the effect of the negative target statement was more persistent than that of the positive target statement. This was confirmed by comparing the absolute difference in impression ratings between trials 12 and 19 for the negative and positive no-retraction conditions (negative: 23.21 > positive: 12.59; $t(198) = 4.68$, $p < .001$). When the

positive target statement was retracted in the positive-retraction condition, impression ratings were lower than in the positive no-retraction condition, $\beta = -8.95$, $SE = 3.39$, $t = -2.64$, $p = .009$, and also lower than in the (positive-congruent) neutral condition, $\beta = -13.01$, $SE = 3.27$, $t = -3.98$, $p < .001$. This means there was no evidence of a CIE for positive misinformation; however, no strong conclusions can be drawn from this because the effect of the positive misinformation had "worn off" regardless of retraction (i.e., the positive no-retraction and [positive-congruent] neutral conditions did not differ).

Delayed Test

Next, the same set of analyses was conducted on the trial-27 impression ratings. For negative misinformation, the model was significant, $F(2, 297) = 19.93$, $p < .001$, $R^2 = .12$. In the negative no-retraction condition, impression ratings were lower than in the (negative-congruent) neutral condition, $\beta = -12.98$, $SE = 3.35$, $t = -3.88$, $p < .001$, indicating that the effect of the negative target statement persisted at trial 27. For the negative-retraction condition, impression ratings were higher than those in the negative no-retraction condition, $\beta = 19.95$, $SE = 3.23$, $t = 6.18$, $p < .001$, and also those in the (negative-congruent) neutral condition, $\beta = 6.97$, $SE = 3.03$, $t = 2.30$, $p = .023$. This indicates the retraction of the negative misinformation was effective, and there was still evidence of an overcorrection after a delay.

For positive misinformation, the model was not significant, $F(2, 297) = 0.12$, $p = .889$, $R^2 < .01$, indicating there was no statistical evidence of a difference between conditions. Thus, there was no evidence for a CIE (the positive-retraction and [positive-congruent] neutral conditions did not differ); however, as with the immediate test, no strong conclusions can be drawn from this because the effect of the positive misinformation had "worn off" regardless of retraction.

Likability Scores

Likability scores were entered into a linear regression with condition as a predictor (retraction, no-retraction, neutral), with separate models for the negative and positive conditions. For negative misinformation, the model was significant, $F(2, 297) = 21.27$, $p < .001$, $R^2 = .13$ (see Figure 5). Likability was lower in the negative no-retraction condition compared to the (negative-congruent) neutral condition, $\beta = -0.51$, $SE = 0.11$, $t = -4.84$, $p < .001$, and the negative-retraction condition, $\beta = -0.59$, $SE = 0.09$, $t = -6.46$, $p < .001$. There was no evidence of a statistical difference between the negative-retraction and (negative-congruent) neutral conditions, $\beta = 0.08$, $SE = 0.10$, $t = 0.82$, $p = .415$. This again indicates the retraction of negative misinformation was fully effective (i.e., no CIE).

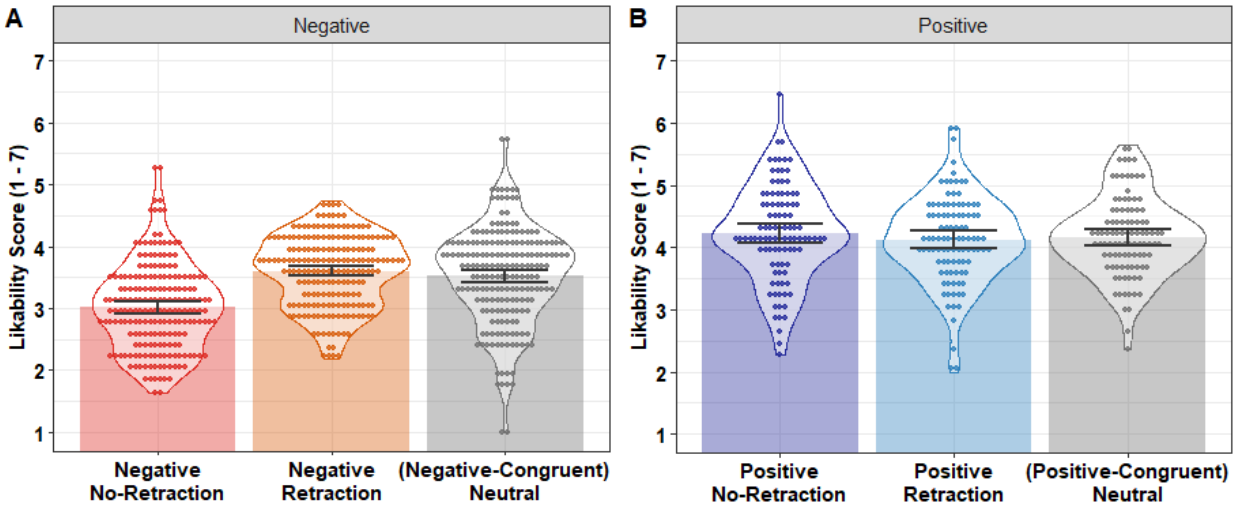
For positive misinformation, there was no statistical evidence of a difference between the conditions, $F(2, 297) = 0.25$, $p = .780$, $R^2 < .01$. This indicates there was neither an impact of the misinformation, nor a CIE.

Discussion

The Experiment 2 finding of a stronger and more enduring impact of negative misinformation compared to positive misinformation replicated the negativity bias shown in Experiment 1. Also in line with Experiment 1, Experiment 2 again demonstrated that people can fully discount discredited negative or positive misinformation. For negative misinformation, the lack of CIE was attributable to a fully effective retraction; in fact, there was again an immediate overcorrection such that impression ratings became more positive compared to the neutral-condition baseline. Unlike Experiment 1, this overcorrection was maintained after a delay. By contrast, the effect of positive misinformation wore off over time even without a retraction. This replicates a pattern shown in impression ratings and likability scores at the end of Experiment 1, but extended also to impression ratings at the time of the retraction.

Figure 5

Mean Likability Scores Across Conditions in Experiment 2



Note. Bars indicate condition means; error bars are 95% bootstrapped confidence intervals; data points indicate jittered participant mean scores; violins provide distributional information.

Thus, contrary to our prediction, the results of Experiment 2 showed that adding information that is congruent with the misinformation did not promote a CIE in impression formation (immediately or after a delay). We predicted that including congruent information would foster integration of the misinformation into participants' mental model of the protagonist, making the misinformation more resistant to correction (Ecker et al., 2010; Ecker, Lewandowsky, Fenton, et al., 2014). This prediction was not supported: we found no evidence that people continued to rely on misinformation after it was retracted.

It is possible that the congruent information (e.g., John cheated in a card game) was not sufficiently related to the misinformation (e.g., John having an affair) to affect its integration into the mental model of the protagonist. Another possibility is that the relation between the misinformation and the supporting information needs to be more causal, rather than simply congruent (Connor Desai & Reimers, 2022; Hamby et al., 2020; Johnson & Seifert, 1994). Experiment 3 therefore included causally-related information, with a view to making the

misinformation more resistant to correction. If the misinformation provides a causal explanation for another piece of information provided, people may be motivated to hold onto the misinformation despite a retraction due to its functional role in their mental model; thus, a CIE should be more likely to emerge with the addition of causally-related information.

Finally, as mentioned earlier, the CIE is typically detected using inferential-reasoning questions rather than impression ratings (Brydges et al., 2018; Johnson & Seifert, 1994; Lewandowsky et al., 2012; Wilkes & Leatherbarrow, 1988). It could be that the impression-rating questions used in Experiments 1 and 2 are not ideally suited to measuring continued influence in the person-impression paradigm. In light of this, an additional inferential-reasoning measure was introduced in Experiment 3 to determine whether a CIE may be observed in inferential judgements even in the absence of a CIE in impression ratings.

Experiment 3

Experiment 3 tested whether inclusion of a statement that is causally related to the target misinformation statement would promote a CIE in impression formation. Given the effects of positive misinformation tended to dissipate in Experiments 1 and 2 even without a retraction, Experiment 3 focused exclusively on negative misinformation. Experiment 3 included a statement ("John was spotted in a hotel lobby with his best friend's wife") that had a pre-tested causal relation to the target statement ("John had an affair with his best friend's wife"). We reasoned that this statement (the "causally-related statement" from here on) would promote stronger integration of the target statement within participants' mental model of the protagonist (given it would be unlikely for the event in the causally-related statement to occur without the target statement being true), making the misinformation more resistant to retraction. In other words, retracting the target statement would leave the information in the causally-related statement unexplained, thus creating a motivation to dismiss the retraction and retain the

misinformation in the mental model. Experiment 3 used the same impression-formation task as Experiment 1 (i.e., without the congruent information from Experiment 2), but with the addition of the causally-related statement. In addition, Experiment 3 included an indirect measure of misinformation reliance, using inferential-reasoning questions (for a similar approach, see Brydges et al., 2018; Ecker et al., 2017), allowing us to investigate whether inferential reasoning about the target person is subject to a CIE even if impression formation per se is not.

Method

Participants

To pilot-test candidate causally-related statements, we recruited 100 U.S.-based participants via Prolific. Two participants were excluded due to uniform responding, leaving a sample of N = 98 (67 female, 29 male, 1 other, 1 preferred not to say) aged 18–57 (M = 27.74, SD = 8.14). For Experiment 3, a separate sample of 300 U.S.-based participants (206 females, 81 males, 12 other, 1 preferred not to say) aged 18–69 (M = 29.91, SD = 10.56) was recruited via Prolific. Participants were randomly assigned to one of three conditions (n = 100 per condition). Prolific users who had participated in Experiment 1 were not invited to participate.

Materials

Causally-Related Statement Selection

Pilot testing was conducted to select a statement that people would interpret as being causally related to the target statement. Participants were told to assume—in line with the misinformation retraction—that John was NOT having an affair, and given this information were asked to rate the conditional likelihood of ten scenarios (e.g., “John went to the drycleaner to get a lipstick stain removed from his collar”), using a 5-point Likert scale from 1 (*very unlikely*) to 5 (*very likely*). The statement “John was spotted with his best friend’s wife in a hotel lobby” (M = 1.82, SD = 0.98; see Supplementary Materials) was selected on the basis that it was rated to

be unlikely (under the assumption there was no affair) but without being overly indicative of an affair (compared to, e.g., the "wife announced she was pregnant with John's baby"). The pilot test took approximately 2 minutes, and participants were paid £0.30 (approximately US$0.35) for their participation.

Inference Questions

In Experiment 3, seven inference questions (two open-ended questions and five Likert rating scales) were designed to measure reliance on the misinformation during inferential reasoning. The open-ended questions (presented as the first and last items) asked participants for information about John ("If you could tell someone about one specific thing John has done, what would it be?"; "Describe briefly in one sentence what kind of relationship John has with his best friend's wife"). The rating scales asked participants to rate their endorsement of John across five statements (e.g., "John is an honorable man"; see Supplementary Materials) using an 11-point Likert scale from 0 (*strongly disagree*) to 10 (*strongly agree*).

Inference-Score Calculation. Responses to the open-ended inference questions were coded by two scorers blind to the experimental conditions, following a standardised guide (see Supplementary Materials). All scoring discrepancies were resolved through discussion. Any unambiguous reference to the target statement was scored 1 (e.g., "John was having an affair with his best friend's wife"). References to the misinformation suggesting an ambiguous level of endorsement were scored 0.5 (e.g., "John may have been having an affair?"). Responses were scored 0 where the misinformation was not mentioned or was specifically discredited (e.g., "It was suggested that John was having an affair but that was not the case"). To put the Likert scale inference questions on the same scale as the open-ended questions, responses were divided by 10 (e.g., Brydges et al., 2018; Ecker et al., 2010); scales that were positively worded were reverse-

scored. All responses were then summed to create an overall inference score, which ranged from 0 to 7, with higher scores indicating stronger endorsement of the misinformation.

Procedure

The impression-formation task in Experiment 3 matched that in Experiment 1, but with the causally-related statement replacing a neutral filler statement at trial 15 (shortly after the target statement was presented). Following the impression-formation task, participants completed the likability scale and inference questions, and the recognition test.[3] Participants were then debriefed and thanked for their participation. Median completion time was 10 minutes, and participants were paid £1.50 (approximately US$2.05).

Results

Performance on the recognition test was high (M = 92.67%, SD = 7.11%). All participants scored at least 50%.

Manipulation Checks

As illustrated in Figure 6, participants in all conditions reported a positive impression of John from the outset, and this gradually improved over trials 1–12 (see Supplementary Materials). The target statement was then presented at trial 13. To ensure the target-statement manipulation was successful, we tested the impact of the negative target statement on person impressions at trial 13. Again, at trial 13, there was no evidence of a statistical difference between the negative-retraction and negative no-retraction conditions (p = .604), and thus for the purpose of comparing to the neutral condition, conditions were collapsed into a combined negative condition. Impression ratings at trial 13 were entered into a linear regression with

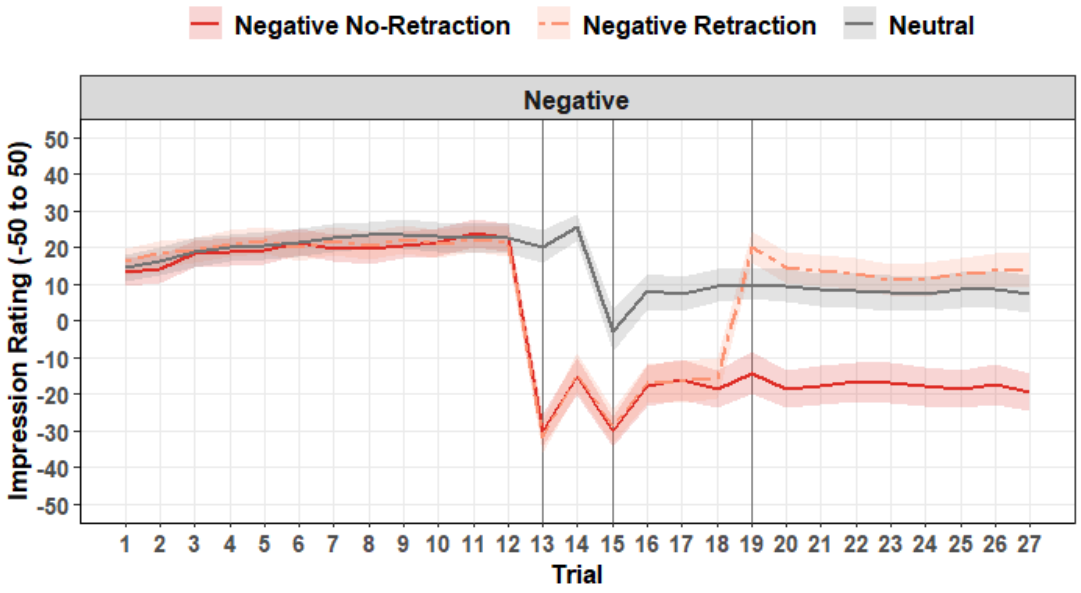
[3] An additional impression rating was added at the very end of the procedure to determine if there were any carry-over effects following the inference questions. This was not the case. For the interested reader, the results of the final impression rating can be found in the Supplementary Materials.

target-statement valence as the predictor (negative; neutral). This showed that the negative conditions returned lower impression ratings than the neutral control (i.e., impressions were more negative when participants had received the negative target statement; $\beta = -51.04$, $SE = 2.77$, $t = -18.42$, $p < .001$).

Between trial 13 (the target statement) and trial 19 (the target-statement retraction, in the retraction condition), the effect of the negative target statement weakened but ratings remained lower than the neutral-condition baseline (the difference to neutral was statistically significant; see Supplementary Materials).

Figure 6

Mean Impression Ratings Across Trials of Experiment 3 for Negative and Neutral Target Statements



Note. The shaded areas are 95% confidence intervals. The vertical line at trial 13 indicates target-statement presentation, the vertical line at trial 15 indicates causally-related statement presentation, and the vertical line at trial 19 indicates target-statement retraction (in the retraction condition).

Tests for Continued Influence

The CIE was tested at two time-points: immediately following the target-statement retraction (trial 19) and after a delay (trial 27).

Immediate Test

The trial-19 impression ratings were entered into a linear regression with condition as a predictor (negative retraction; negative no-retraction; neutral). The model was significant, $F(2, 297) = 54.69$, $p < .001$, $R^2 = .27$. In the negative no-retraction condition, impression ratings were lower than in the neutral control condition, $\beta = -24.32$, $SE = 3.56$, $t = -6.83$, $p < .001$, indicating that the effect of the negative target statement persisted at trial 19. When the negative target statement was retracted in the negative-retraction condition, impression ratings were higher than in the negative no-retraction condition, $\beta = 34.59$, $SE = 3.61$, $t = 9.58$, $p < .001$, and also higher than in the neutral control condition, $\beta = 10.27$, $SE = 2.98$, $t = 3.45$, $p < .001$. This indicates the retraction of the negative target statement was fully effective; in fact, participants again overcorrected at immediate test.

Delayed Test

Next, the same set of analyses was conducted on the trial-27 impression ratings. The model was significant, $F(2, 297) = 53.26$, $p < .001$, $R^2 = .26$. In the negative no-retraction condition, impression ratings were lower than in the neutral control condition, $\beta = -26.79$, $SE = 3.59$, $t = -7.47$, $p < .001$, indicating that the effect of the negative target statement persisted at trial 27. For the negative-retraction condition, impression ratings were higher than those in the negative no-retraction condition, $\beta = 33.39$, $SE = 3.37$, $t = 9.90$, $p < .001$, and also higher than in the neutral control condition, $\beta = 6.60$, $SE = 3.31$, $t = 1.99$, $p = .048$. This indicates that the retraction of the negative target statement was still fully effective after a delay, and again there was evidence of an overcorrection.

Likability Scores

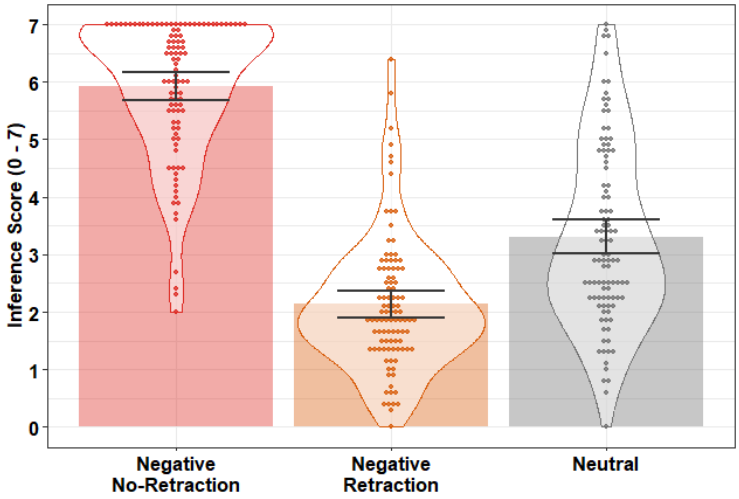
Likability scores were entered into a linear regression with condition as a predictor (negative retraction; negative no-retraction; neutral). The model was significant, $F(2, 297) = 20.40$, $p < .001$, $R^2 = .12$. Likability was lower in the negative no-retraction condition compared to the neutral control condition, $\beta = -0.62$, $SE = 0.12$, $t = -5.16$, $p < .001$, and the negative-retraction condition, $\beta = -0.66$, $SE = 0.12$, $t = -5.62$, $p < .001$. There was no evidence of a statistical difference between the negative-retraction and neutral conditions, $\beta = 0.05$, $SE = 0.11$, $t = 0.42$, $p = .673$, indicating the retraction was fully effective.

Inference Scores

Inference scores supported the observations made based on the impression ratings and likability scores (see Figure 7). Inference scores were entered into a linear regression with condition as a predictor (negative retraction; negative no-retraction; neutral). The model was significant, $F(2, 297) = 210.10$, $p < .001$, $R^2 = .59$. Inference scores were higher (i.e., participants made more references to the negative misinformation) for the negative no-retraction condition compared to the neutral condition, $\beta = 2.62$, $SE = 0.20$, $t = 13.05$, $p < .001$, and the negative-retraction condition, $\beta = 3.79$, $SE = 0.17$, $t = 22.46$, $p < .001$. In the negative-retraction condition, inference scores were lower than in the neutral condition, $\beta = -1.23$, $SE = 0.20$, $t = -5.96$, $p < .001$. This indicated that the retraction was fully effective and there was no evidence of a CIE in inferential reasoning; in fact, participants overcorrected.

Figure 7

Mean Inference Scores Across Conditions in Experiment 3



Note. Bars indicate condition means; error bars are 95% bootstrapped confidence intervals; data points indicate jittered participant mean scores; violins provide distributional information.

Discussion

In line with Experiments 1 and 2, Experiment 3 showed that non-retracted negative information has an enduring impact on person impressions. Experiment 3 further corroborated the finding that people can fully discount discredited misinformation in impression formation. Results again showed that there was an overcorrection immediately following the retraction of the negative misinformation and after a delay, such that impressions were more positive than in the neutral condition. In addition, Experiment 3 extended these findings to inferential reasoning, suggesting that retracted misinformation is not relied upon to make subsequent inferences in the context of person impressions.

Thus, despite our predictions, the introduction of an inferential-reasoning measure did not allow us to detect a CIE in person-related reasoning. This is in contrast to the robust findings for the CIE in the context of events, where retracted misinformation reliably continues to influence inferential reasoning (Chan et al., 2017; Ecker et al., 2022; Walter & Tukachinsky,

2020). These findings were obtained despite the inclusion of a causally-related statement designed to strengthen person-model coherence and make the misinformation more resistant to retraction.

Experiment 3 also showed that participants who received the causally-related statement but not the negative target statement (i.e., those in the neutral condition) formed a more negative impression of John than those who received both but subsequently had the negative target statement retracted (negative retraction condition). The inference questions showed a similar pattern, with more references to the misinformation in the neutral compared to the negative-retraction condition, despite the fact that the misinformation was not presented in the neutral condition. This suggests that participants inferred the misinformation even if it was not explicitly mentioned, based on the innuendo provided by the causally-related statement (i.e., participants learning only that John had been spotted in a hotel lobby with his friend's wife inferred they were having an affair); it also demonstrates that participants perceived a relation between the causally-related statement and the target statement, in line with pre-testing. In the retraction condition, the effect of the causally-related statement on participants' impressions and inferences was counteracted by the target statement's retraction, leading to a more neutral response. Other CIE literature has shown that misinformation can still affect people's judgements when it is not explicitly mentioned but merely hinted at (Ecker, Lewandowsky, Chang, et al., 2014; Rich & Zaragoza, 2016; but see Tay et al., 2022).

The experiments reported so far used the same misinformation statement, which may limit the generalisability of our results. While single statements are typical in continued-influence and impression-formation research (Brydges et al., 2020; Cone & Ferguson, 2015; Gordon et al., 2017; Rydell et al., 2007; Skowronski, 2002; Skowronski & Carlston, 1989; Srull & Wyer, 1989), it is possible that varying the misinformation statement could return different

results (i.e., presence of a CIE). To test this, we ran a final experiment that used a different target statement and a different causally-related statement.

Experiment 4

Experiment 4 tested whether the inclusion of a different target statement and a different causally-related statement would promote a CIE in impression formation. This was done to guard against the possibility that the null effects observed in Experiments 1–3 might be restricted to the materials used. Experiment 4 used the same impression-formation task as Experiment 3.

Method

Participants

For Experiment 4, a sample of 300 U.S.-based participants (183 females, 110 males, 5 other, 2 preferred not to say) aged 18–74 (M = 37.78, SD = 13.86) was recruited via Prolific. Participants were randomly assigned to one of three conditions (n = 100 per condition). Prolific users who had participated in Experiment 1 or Experiment 3 were not invited to participate.

Materials

Impression-Formation Task

The task was identical to that used in Experiment 3, but with a new target statement and a new causally-related statement.

Target statement. The new target statement was selected from the corpus of pre-tested behavioural stimuli (Mickelberg et al., 2022) and met similar criteria to the original target statement (morality value close to −4, believability > 5). The target statement selected was “John kicked his pet dog hard in the head when it didn’t come when called” (morality = −3.31, believability = 5.82).

Causally-related statement. The causally-related statement was generated by the authors. The causally-related statement needed to (i) be causally related to the target statement, and (ii)

be perceived as neutral and not imply the misinformation when presented in the absence of the target statement (i.e., in the neutral condition). The causally-related statement selected was "John accompanied his wife to the vet to have his dog treated for a head injury".

Inference Questions

As per Experiment 3, seven inference questions (two open-ended questions and five Likert rating scales) were designed to measure reliance on the misinformation during inferential reasoning. The questions were updated to reflect the new misinformation statement: open-ended questions (presented as the first and last items) asked participants for information about John ("If you could tell someone about one specific thing John has done, what would it be?"; "Describe briefly in one sentence how you think John's pet dog could have received a head injury"). The rating scales asked participants to rate their endorsement of John across five statements (e.g., "John is a good person"; see Supplementary Materials) using an 11-point Likert scale from 0 (*strongly disagree*) to 10 (*strongly agree*).

Inference-Score Calculation. The inference-score calculation matched Experiment 3, with an updated scoring guide to reflect the new target statement (see Supplementary Materials).

Procedure

The procedure was identical to Experiment 3. Median completion time was 10 minutes, and participants were paid £1.50 (approximately US$2.05).

Results

Performance on the recognition test was high (M = 91.30%, SD = 8.59%). One participant was excluded due to technical difficulties preventing them from completing the task, leaving a final sample of N = 299. One participant scored below 50% in the recognition test; excluding them did not qualitatively affect the results so they were included in the analysis.

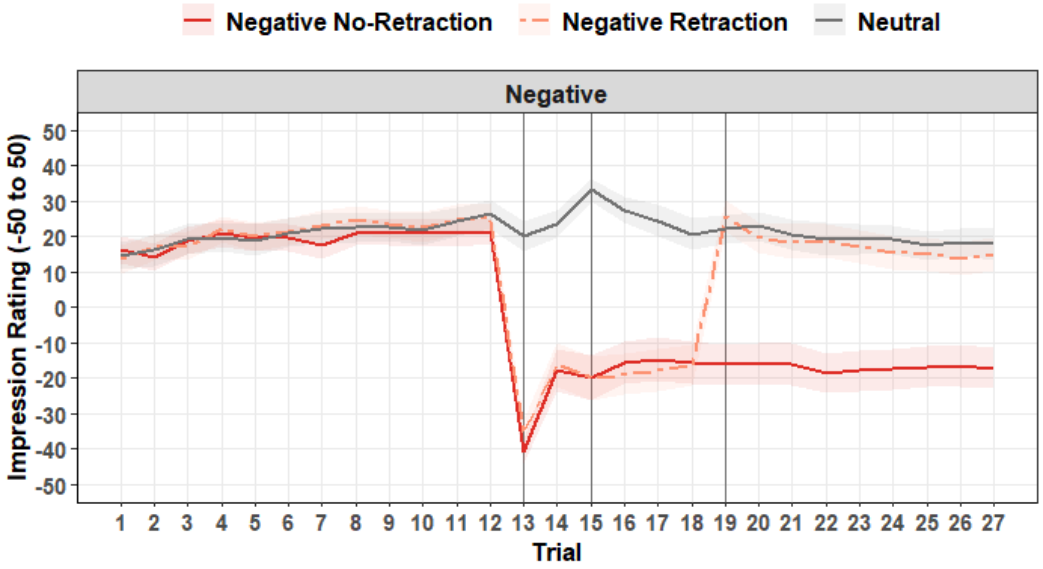
Manipulation Checks

As illustrated in Figure 8, participants in all conditions reported a positive impression of John from the outset, and this gradually improved over trials 1–12 (see Supplementary Materials). The target statement was then presented at trial 13. To ensure the target-statement manipulation was successful, we tested the impact of the negative target statement on person impressions at trial 13. At trial 13, there was no evidence of a statistical difference between the negative-retraction and negative no-retraction conditions ($p = .076$), and thus for the purpose of comparing to the neutral condition, these were collapsed into a combined negative condition. Impression ratings at trial 13 were entered into a linear regression with target-statement valence as the predictor (negative, neutral). This showed that the negative conditions returned lower impression ratings than the neutral control (i.e., impressions were more negative when participants had received the negative target statement; $\beta = -58.38$, $SE = 2.44$, $t = -23.84$, $p < .001$).

Between trial 13 (the target statement) and trial 19 (the target-statement retraction, in the retraction condition), the effect of the negative target statement weakened but ratings remained lower than the neutral-condition baseline (the difference to neutral was statistically significant; see Supplementary Materials).

Figure 8

Mean Impression Ratings Across Trials of Experiment 4 for Negative and Neutral Target Statements



Note. The shaded areas are 95% confidence intervals. The vertical line at trial 13 indicates target-statement presentation, the vertical line at trial 15 indicates causally-related statement presentation, and the vertical line at trial 19 indicates target-statement retraction (in the retraction condition).

Tests for Continued Influence

The CIE was tested at two time-points: immediately following the target-statement retraction (trial 19) and after a delay (trial 27).

Immediate Test

The trial-19 impression ratings were entered into a linear regression with condition as a predictor (negative retraction; negative no-retraction; neutral). The model was significant, $F(2, 296) = 96.77$, $p < .001$, $R^2 = .39$. In the negative no-retraction condition, impression ratings were lower than in the neutral control condition, $\beta = -38.38$, $SE = 3.46$, $t = -11.09$, $p < .001$, indicating that the effect of the negative target statement persisted at trial 19. When the negative

target statement was retracted in the negative-retraction condition, impression ratings were higher than in the negative no-retraction condition, β = 42.15, *SE* = 3.62, *t* = 11.64, *p* < .001, and did not differ from the neutral control condition, β = 3.77, *SE* = 2.94, *t* = 1.28, *p* = .202. This indicates the retraction of the negative target statement was fully effective.

Delayed Test

Next, the same set of analyses was conducted on the trial-27 impression ratings. The model was significant, $F(2, 296) = 59.88$, $p < .001$, $R^2 = .29$. In the negative no-retraction condition, impression ratings were lower than in the neutral control condition, β = −34.71, *SE* = 3.67, *t* = −9.46, *p* < .001, indicating that the effect of the negative target statement persisted at trial 27. For the negative-retraction condition, impression ratings were higher than those in the negative no-retraction condition, β = 31.88, *SE* = 3.73, *t* = 8.55, *p* < .001, and did not differ from the neutral control condition, β = −2.83, *SE* = 3.15, *t* = −0.90, *p* = .370. This indicates that the retraction of the negative target statement was still fully effective after a delay.

Likability Scores

Likability scores were entered into a linear regression with condition as a predictor (negative retraction; negative no-retraction; neutral). The model was significant, $F(2, 296) = 81.88$, $p < .001$, $R^2 = .36$. Likability was lower in the negative no-retraction condition compared to the neutral control condition, β = −1.62, *SE* = 0.15, *t* = −10.98, *p* < .001, and the negative-retraction condition, β = −1.51, *SE* = 0.15, *t* = −10.14, *p* < .001. There was no evidence of a statistical difference between the negative-retraction and neutral conditions, β = −0.12, *SE* = 0.13, *t* = −0.91, *p* = .362, indicating that the retraction was fully effective.

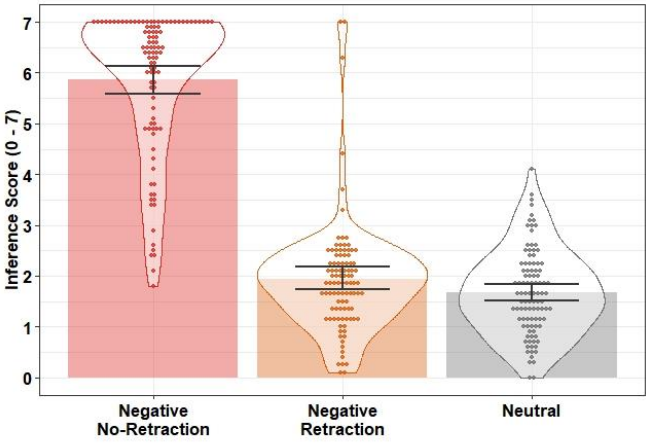
Inference Scores

Inference scores supported the observations made based on the impression ratings and likability scores (see Figure 9). Inference scores were entered into a linear regression with

condition as a predictor (negative retraction, negative no-retraction, neutral). The model was significant, $F(2, 296) = 410.50$, $p < .001$, $R^2 = .74$. Inference scores were higher for the negative no-retraction condition compared to the neutral condition, $\beta = 4.19$, $SE = 0.17$, $t = 25.40$, $p < .001$, and the negative-retraction condition, $\beta = 3.93$, $SE = 0.18$, $t = 21.53$, $p < .001$. There was no evidence of a statistical difference between negative-retraction and neutral conditions, $\beta = 0.26$, $SE = 0.14$, $t = 1.82$, $p = .071$. This indicates that the retraction was fully effective and there was no evidence of a CIE in inferential reasoning.

Figure 9

Mean Inference Scores Across Conditions in Experiment 4



Note. Bars indicate condition means; error bars are 95% bootstrapped confidence intervals; data points indicate jittered participant mean scores; violins provide distributional information.

Equivalence Analysis

In line with Experiments 1–3, Experiment 4 showed that people can fully discount discredited misinformation in impression formation (the only difference across experiments was that there was some evidence for overcorrection in the immediate test of Experiments 1–3 and the delayed test in Experiments 2 and 3). The same pattern of results was observed for the likability scores in Experiments 1–4. However, this conclusion regarding the absence of

evidence for a CIE in impression formation is based on null results, and it therefore needs to be made with caution.

To corroborate the absence of an effect, an equivalence analysis (Counsell & Cribbie, 2015; Lakens, 2017) was conducted on the non-significant linear regression results between the negative-retraction condition and neutral condition across Experiments 1–4, using impression ratings and likability scores. Equivalence analysis allows for direct inferences to be made about the absence of an effect or the presence of a negligibly small effect. We used the Anderson-Hauck procedure (Anderson & Hauck, 1983) in the *reg.equiv* function in R (Alter & Counsell, 2021). The Anderson-Hauck procedure has been shown to have greater statistical power than the two one-sided tests procedure (TOST; Schuirmann, 1987) when comparing regression coefficients at smaller sample sizes (e.g., Alter & Counsell, 2021; Counsell & Cribbie, 2015). In line with Campbell (2023), we chose standardized regression coefficients of ±0.1 (i.e., a small effect size according to Cohen, 1988), and anything that falls within the +0.1 to −0.1 range was determined to be a negligible effect. Across all experiments, the results indicated that there was insufficient evidence for negligible effects (i.e., no consistent evidence in favour of the null) across all the measures (impression ratings and likability scores), meaning that the true population effect could be larger/smaller than ±0.1 (see Supplementary Materials).[4]

Discussion

Experiment 4 used a new set of materials and showed much the same pattern of results as Experiments 1–3. Non-retracted negative misinformation had an enduring impact on person

[4]Because the equivalence bounds were conducted following the main analysis, the equivalence analysis was rerun using a medium effect size (±0.3). Equivalence tests at the ±0.3 level indicated a negligible effect (evidence in favour of the null) for the likability scores (Experiments 1–3) but not for the dynamic impression ratings (Experiments 1 and 4) or likability scores (Experiment 4).

impressions, and a clear retraction eliminated the influence of misinformation on impression formation (see also De keersmaecker & Roets, 2017; Ecker & Rodricks, 2020). Impression ratings did not differ from the neutral control condition immediately following the retraction of the negative misinformation, or after a delay. Contrary to Experiments 1–3, there was no evidence of an overcorrection in Experiment 4. As per Experiment 3, no CIE was detected in person-related inferential reasoning.

General Discussion

The present study investigated the impact of person-related misinformation and its correction on person impressions. As predicted, negative misinformation had a greater impact on person impressions compared to positive misinformation, and the impact was longer-lived (when not retracted), representing a negativity bias in impression formation (Rusconi et al., 2020; Ybarra, 2001). Critically, across four experiments, we found no statistical evidence for the continued influence of retracted misinformation on person impressions. This was true for negative misinformation and for positive misinformation.

The study set out to clarify mixed findings in the existing literature regarding the CIE in impression formation, by testing three explanations. Firstly, we suggested that a short-lived CIE may have gone undetected in previous studies, as they only tested for it at the end of the task. We therefore introduced a dynamic impression-rating measure, but did not detect a CIE, neither at the time the misinformation was retracted, nor at the end of the task. Secondly, as previous studies had used a mixture of negative and positive misinformation, we suggested that misinformation valence may moderate the CIE in impression formation. However, we found no CIE with either negative or positive misinformation. Finally, we suggested there may have been differences in misinformation coherence in previous studies, and this could have affected

whether or not a CIE was observed. However, we found that the introduction of coherence-building elements did not promote a CIE in impression formation.

Our findings are consistent with Ecker and Rodricks (2020) and Cobb (2013), suggesting that people can fully discount retracted person-related misinformation. Moreover, the present results are in line with the broader impression-formation literature whereby person judgements can be rapidly updated in light of new information (Rydell et al., 2007; Skowronski & Carlston, 1992)—this was shown by the success of the retractions, but also by the fact that a single piece of misinformation impacted person impressions.

Our results contrast with the robust evidence for the CIE in event-related reasoning (Chan et al., 2017; Walter & Tukachinsky, 2020). One way to interpret this result is that corrections of event-related and person-related misinformation may be processed differently. While the retraction of causal event-related misinformation can threaten mental-model coherence, promoting a CIE, mental-model coherence may be less important (or more robust) in impression formation. As a result, mental models may be easier to update when person-related misinformation is retracted, and therefore misinformation can be corrected more effectively. If there are fundamental differences between how events and people are represented in mental models (e.g., more emphasis on temporal and causal relations for events; Bower & Morrow, 1990; van Oostendorp & Bonebakker, 1999), this may explain why even the addition of coherence-building elements did not lead to a CIE in impression ratings (Experiments 2–4) and inferential reasoning (Experiments 3 and 4). To confidently establish if there are fundamental differences between events and people, it would be ideal to compare event and person-related misinformation within a single study, with target type (event or person) as an independent variable. However, designing a study in which events can be equivalently compared to people is a non-trivial challenge.

We found no evidence of a CIE across three different pieces of misinformation (one positive and two negative), indicating that the result is somewhat generalisable. However, a CIE may occur for other pieces of misinformation. Going forward, it will be important to test new statements that vary in content (e.g., morality vs. competence), while being careful not to introduce confounds (e.g., evaluative extremity) that have been shown to moderate person judgements (Rusconi et al., 2017; see Brambilla et al., 2021, for a review).

It should also be noted that equivalence tests did not return evidence in favour of the null hypothesis. Our null findings therefore highlight the need for additional research, rather than serving as strong evidence against the existence of continued influence in the context of person impressions. Specifically, we hope that the present paper will contribute to the bigger picture by drawing attention to our null findings (thus avoiding the 'file-drawer problem,' which has contributed to limited reproducibility of, and eroded credibility in, psychological findings; Ferguson & Heene, 2012; Open Science Collaboration, 2015) and allowing interested researchers to establish better-calibrated estimates of the 'true' effect size of the CIE in the context of person impressions, thus ultimately advancing CIE theory.

Limitations and Future Research Directions

Despite the consistent results returned across four experiments, there are several limitations of the current research. First, the present set of experiments used only a single fictional person ("John"). Although a previous study on continued influence in person impressions found equivalent (null) effects regardless of the fictional protagonist's name and implied cultural background (i.e., "John" vs. "Vladimir"; Ecker & Rodricks, 2020), future research could use a range of different names, including female or gender-nonspecific names. Moreover, more research is needed on how misinformation influences impressions of familiar versus unfamiliar people. For example, a CIE may be more likely to emerge if participants

already hold a negative attitude towards a person based on previous encounters, existing knowledge, or incompatible worldviews (Thorson, 2016; see also Bullock, 2007). Given the documentation of clear cases of misinformation impacting person judgements in other contexts (Brooks & Greenberg, 2021; Clow et al., 2012; Jardina & Traugott, 2019; Steblay et al., 2006; Weeks & Garrett, 2014), a long-term research goal will be to identify the boundary conditions under which person misinformation does versus does not continue to influence person impressions after a retraction.

Second, when the misinformation was retracted in the present study, it was done using an unambiguous retraction (i.e., "John did NOT have an affair with his best friend's wife") and there was no reason to question the retraction's credibility. As believability of new evidence is a predictor of impression revision (Cone et al., 2019), the unambiguous nature of the retraction may have contributed to the elimination of the retracted misinformation's influence. That said, previous studies that found a CIE with event misinformation also used unambiguous retractions (e.g., Ecker et al., 2010; Johnson & Seifert, 1994), and apart from the clear wording, no specific measures were taken in the present study to boost the credibility of the retraction or its source; thus, this line of reasoning may not provide a satisfactory explanation for the absence of an effect. However, it is possible that a CIE in impression formation might arise if the correction is more tentative (MacFarlane et al., 2021) or the correction source is less credible (e.g., if a retraction is provided by the accused person themselves; also see Connor Desai et al., 2020; Ecker & Antonio, 2021), which is something that future studies might examine.

Finally, the use of a dynamic impression measure may have led to greater retraction efficacy than would otherwise be expected. Research has shown that memory for event information is a predictor of susceptibility to the CIE (Sanderson et al., 2021). It could be that eliciting trial-to-trial impression updates improved participants' memory of the behaviour

statements, allowing a higher-fidelity mental model to be created. As a result, it may be easier to integrate the retraction and update the mental model, reducing participants' reliance on the retracted misinformation (Ecker et al., 2017; Kendeou et al., 2014, 2019). A complementary explanation is based on evidence that eliciting repeated judgements (similar to the dynamic impression-rating measure) can lead people to give greater weight to more recent information when forming person impressions (i.e., recency effect, see Kashima & Kerekes, 1994, for a review). This could lead participants to prioritise the retraction over the initial misinformation when making their post-retraction impression ratings, reducing reliance on the misinformation. While a dynamic measure was necessary to assess the time course of information updating, future research should consider whether paradigm selection (i.e., continuous vs. single judgements) affects results. The use of dynamic and static impression measures could also provide insight into how misinformation shapes person impressions across different real-world information environments (e.g., social media vs. traditional media), where the frequency of updating may vary.

Conclusion

The continued reliance on person-related misinformation can have serious implications for those who are the target of false allegations, as well as for broader society. The present study found that when people form an impression of a fictional person, negative misinformation has a greater impact than positive misinformation. However, and regardless of valence, when misinformation was unequivocally retracted, participants fully discounted the retracted misinformation, even when other behaviour descriptions were congruent or causally related to the retracted misinformation. Thus, for the scenarios considered, once person-related misinformation was retracted, person impressions could be fully updated, suggesting that mud does not always stick.

Author Contributions

AM, BW, NF and UE contributed to conception and design. BW developed the experiment. BW and AM collected and stored the data. AM analysed the data. AM wrote the manuscript, with input from all authors. All authors revised the article and contributed to interpretation of data. All authors approved the final submitted version of the manuscript.

Competing Interests

The authors declared that they had no conflicts of interest with respect to authorship or the publication of this article.

Data Accessibility Statement

All the participant data can be found on the project page at OSF: https://osf.io/ymuap

Funding

Research reported in this publication was supported by a Postgraduate Research Scholarship from the Defence Science and Technology Group of the Department of Defence and an Australian Government Research Training Program Scholarship to the first author (AM), an Australian Research Council grant FT190100708 to the third author (UE), and an Office of National Intelligence and Australian Research Council grant NI210100224 to the last author (NF).

References

Alter, U., & Counsell, A. (2021). *Determining negligible associations in regression*. PsyArXiv. https://doi.org/10.31234/osf.io/ugc9e

Anderson, C., Lepper, M., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, *39*, 1037–1049. https://doi.org/10.1037/h0077720

Anderson, S., & Hauck, W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics - Theory and Methods*, *12*(23), 2663–2692. https://doi.org/10.1080/03610928308828634

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, *41*(3), 258–290.

Bower, G. H., & Morrow, D. G. (1990). Mental models in narrative comprehension. *Science*, *247*(4938), 44–48. https://doi.org/10.1126/science.2403694

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 64, pp. 187–262). Academic Press. https://doi.org/10.1016/bs.aesp.2021.03.001

Brooks, S. K., & Greenberg, N. (2021). Psychological impact of being wrongfully accused of criminal offences: A systematic literature review. *Medicine, Science and the Law*, *61*(1), 44–54. https://doi.org/10.1177/0025802420949069

Brydges, C. R., Gignac, G. E., & Ecker, U. K. H. (2018). Working memory capacity, short-term memory capacity, and the continued influence effect: A latent-variable analysis. *Intelligence*, *69*, 117–122. https://doi.org/10.1016/j.intell.2018.03.009

Brydges, C. R., Gordon, A., & Ecker, U. K. H. (2020). Electrophysiological correlates of the continued influence effect of misinformation: An exploratory study. *Journal of Cognitive Psychology*, *32*(8), 771–784. https://doi.org/10.1080/20445911.2020.1849226

Bullock, J. G. (2007). Experiments on partisanship and public opinion: Party cues, false beliefs, and Bayesian updating [Ph.D., Stanford University]. In *ProQuest Dissertations and Theses* (304810798). ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/experiments-on-partisanship-public-opinion-party/docview/304810798/se-2?accountid=14681

Campbell, H. (2023). *Equivalence testing for linear regression*. PsyArXiv. https://doi.org/10.48550/arXiv.2004.01757

Cappella, J. N., Maloney, E., Ophir, Y., & Brennan, E. (2015). Interventions to correct misinformation about tobacco products. *Tobacco Regulatory Science*, *1*(2), 186–197. https://doi.org/10.18001/TRS.1.2.8

Cappella, J. N., Ophir, Y., & Sutton, J. (2018). The importance of measuring knowledge in the age of misinformation and challenges in the tobacco domain. In *Misinformation and mass audiences* (pp. 51–70). University of Texas Press. https://doi.org/10.7560/314555-005

Carlston, D. E., & Smith, E. R. (1996). Principles of mental representation. In *Social psychology: Handbook of basic principles* (pp. 184–210). The Guilford Press.

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Chang, E. P., Ecker, U. K. H., & Page, A. C. (2019). Not wallowing in misery – retractions of negative misinformation are effective in depressive rumination. *Cognition and Emotion*, *33*(5), 991–1005. https://doi.org/10.1080/02699931.2018.1533808

Clow, K., & Leach, A.-M. (2015). After innocence: Perceptions of individuals who have been wrongfully convicted. *Legal and Criminological Psychology*, *20*, 147–164. https://doi.org/10.1111/lcrp.12018

Clow, K., Ricciardelli, R., & Cain, T. L. (2012). Stigma-by-association: Prejudicial effects of the prison experience for offenders and exonerees. In *The psychology of prejudice: Interdisciplinary perspectives on contemporary issues* (pp. 127–154).

Cobb, M. D., Nyhan, B., & Reifler, J. (2013). Beliefs don't always persevere: How political figures are punished when positive information about them is discredited. *Political Psychology*, *34*(3), 307–326. https://doi.org/10.1111/j.1467-9221.2012.00935.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37–57. https://doi.org/10.1037/pspa0000014

Cone, J., Flaharty, K., & Ferguson, M. J. (2019). Believability of evidence matters for correcting social impressions. *Proceedings of the National Academy of Sciences*, *116*(20), 9802–9807.

Connor Desai, S., Pilditch, T. D., & Madsen, J. K. (2020). The rational continued influence of misinformation. *Cognition*, *205*, 104453. https://doi.org/10.1016/j.cognition.2020.104453

Connor Desai, S., & Reimers, S. (2019). Comparing the use of open and closed questions for web-based measures of the continued-influence effect. *Behavior Research Methods*, *51*(3), 1426–1440. https://doi.org/10.3758/s13428-018-1066-z

Connor Desai, S., & Reimers, S. (2022). Does explaining the origins of misinformation improve the effectiveness of a given correction? *Memory & Cognition*, 1–15. https://doi.org/10.3758/s13421-022-01354-7

Counsell, A., & Cribbie, R. A. (2015). Equivalence tests for comparing correlation and regression coefficients. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 292–309. https://doi.org/10.1111/bmsp.12045

De keersmaecker, J., & Roets, A. (2017). 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence*, *65*, 107–110. https://doi.org/10.1016/j.intell.2017.10.005

Downey, J. L., & Christensen, L. (2006). Belief persistence in impression formation. *North American Journal of Psychology*, *8*(3), 479–487.

Ecker, U. K. H., & Antonio, L. M. (2021). Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition*, *49*, 631–644. https://doi.org/10.3758/s13421-020-01129-y

Ecker, U. K. H., Hogan, J. L., & Lewandowsky, S. (2017). Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition*, *6*(2), 185–192. https://doi.org/10.1037/h0101809

Ecker, U. K. H., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied*, *20*(4), 323–335. https://doi.org/10.1037/xap0000028

Ecker, U. K. H., Lewandowsky, S., Cheung, C. S. C., & 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. https://doi.org/10.1016/j.jml.2015.09.002

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Ecker, U. K. H., Lewandowsky, S., Fenton, O., & Martin, K. (2014). Do people keep believing because they want to? Preexisting attitudes and the continued influence of misinformation. *Memory & Cognition*, *42*(2), 292–304. https://doi.org/10.3758/s13421-013-0358-x

Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, *18*(3), 570–578. https://doi.org/10.3758/s13423-011-0065-1

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*(8), 1087–1100. https://doi.org/10.3758/MC.38.8.1087

Ecker, U. K. H., & Rodricks, A. E. (2020). Do false allegations persist? Retracted misinformation does not continue to influence explicit person impressions. *Journal of Applied Research in Memory and Cognition*, *9*(4), 587–601. https://doi.org/10.1016/j.jarmac.2020.08.003

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

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. https://doi.org/10.1037/0022-3514.38.6.889

Fiske, S. T., & Linville, P. W. (1980). What does the schema concept buy us? *Personality and Social Psychology Bulletin*, *6*(4), 543–557. https://doi.org/10.1177/014616728064006

Gordon, A., Brooks, J. C. W., Quadflieg, S., Ecker, U. K. H., & Lewandowsky, S. (2017). Exploring the neural substrates of misinformation processing. *Neuropsychologia*, *106*, 216–224. https://doi.org/10.1016/j.neuropsychologia.2017.10.003

Green, M. C., & Donahue, J. K. (2011). Persistence of belief change in the face of deception: The effect of factual stories revealed to be false. *Media Psychology*, *14*(3), 312–331. https://doi.org/10.1080/15213269.2011.598050

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

Hamby, A., Ecker, U., & Brinberg, D. (2020). How stories in memory perpetuate the continued influence of false information. *Journal of Consumer Psychology*, *30*(2), 240–259. https://doi.org/10.1002/jcpy.1135

Jardina, A., & Traugott, M. (2019). The genesis of the birther rumor: Partisanship, racial attitudes, and political knowledge. *Journal of Race, Ethnicity, and Politics*, *4*(1), 60–80. https://doi.org/10.1017/rep.2018.25

Johnson, H., & Seifert, C. (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. https://doi.org/10.1037/0278-7393.20.6.1420

Johnson-Laird, P. N. (2012). Inference with mental models. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 134–154). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0009

Juul, J. L., & Ugander, J. (2021). Comparing information diffusion mechanisms by matching on cascade size. *Proceedings of the National Academy of Sciences*, *118*(46), e2100786118. https://doi.org/10.1073/pnas.2100786118

Kashima, Y., & Kerekes, A. R. Z. (1994). A distributed memory model of averaging phenomena in person impression formation. *Journal of Experimental Social Psychology*, *30*(5), 407–455. https://doi.org/10.1006/jesp.1994.1021

Kendeou, P., Butterfuss, R., Kim, J., & Van Boekel, M. (2019). Knowledge revision through the lenses of the three-pronged approach. *Memory & Cognition*, *47*(1), 33–46. https://doi.org/10.3758/s13421-018-0848-y

Kendeou, P., van den Broek, P., Helder, A., & Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research & Practice*, *29*(1), 10–16. https://doi.org/10.1111/ldrp.12025

Kerpelman, J. P., & Himmelfarb, S. (1971). Partial reinforcement effects in attitude acquisition and counterconditioning. *Journal of Personality and Social Psychology*, *19*, 301–305. https://doi.org/10.1037/h0031447

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

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

MacFarlane, D., Tay, L. Q., Hurlstone, M. J., & Ecker, U. K. H. (2021). Refuting spurious COVID-19 treatment claims reduces demand and misinformation sharing. *Journal of*

Applied Research in Memory and Cognition, *10*(2), 248–258. https://doi.org/10.1016/j.jarmac.2020.12.005

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

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, *8*(6), 623–631. https://doi.org/10.1093/scan/nss040

Mensink, M. C., & Rapp, D. N. (2011). Evil geniuses: Inferences derived from evidence and preferences. *Memory & Cognition*, *39*(6), 1103–1116. https://doi.org/10.3758/s13421-011-0081-4

Mickelberg, A., Walker, B., Ecker, U. K. H., Howe, P., Perfors, A., & Fay, N. (2022). Impression formation stimuli: A corpus of behavior statements rated on morality, competence, informativeness, and believability. *PLOS ONE*, *17*(6). https://doi.org/10.1371/journal.pone.0269393

NHMRC. (2018, updated). *National statement on ethical conduct in human research (2007)*. https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018#toc__725

Okten, I., Schneid, E. D., & Moskowitz, G. B. (2019). On the updating of spontaneous impressions. *Journal of Personality and Social Psychology*, *117*(1), 1–25. https://doi.org/10.1037/pspa0000156

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

O'Rourke, C. (2016). *Four years later, there's still no evidence to support Pizzagate theory*. Politifact. https://www.politifact.com/factchecks/2020/oct/07/facebook-posts/four-years-later-theres-still-no-evidence-support-/

Park, B. (1986). A method for studying the development of impressions of real people. *Journal of Personality and Social Psychology, 51*, 907–917. https://doi.org/10.1037/0022-3514.51.5.907

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rapp, D. N., & Kendeou, P. (2007). Revising what readers know: Updating text representations during narrative comprehension. *Memory & Cognition, 35*(8), 2019–2032. https://doi.org/10.3758/BF03192934

Rapp, D. N., & Kendeou, P. (2009). Noticing and revising discrepancies as texts unfold. *Discourse Processes, 46*(1), 1–24. https://doi.org/10.1080/01638530802629141

Reuters. (2020). *Fact check: Bill Gates is not responsible for COVID-19*. Reuters. https://www.reuters.com/article/uk-factcheck-gates-idUSKBN2613CK

Reysen, S. (2005). Construction of a new scale: The Reysen Likability Scale. *Social Behavior and Personality: An International Journal*, *33*(2), 201–208. https://doi.org/10.2224/sbp.2005.33.2.201

Rich, P. R., & Zaragoza, M. S. (2016). The continued influence of implied and explicitly stated misinformation in news reports. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(1), 62–74. https://doi.org/10.1037/xlm0000155

Rich, P. R., & Zaragoza, M. S. (2020). Correcting misinformation in news stories: An investigation of correction timing and correction durability. *Journal of Applied Research in Memory and Cognition*, S2211368120300280. https://doi.org/10.1016/j.jarmac.2020.04.001

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Rusconi, P., Sacchi, S., Brambilla, M., Capellini, R., & Cherubini, P. (2020). Being honest and acting consistently: Boundary conditions of the negativity effect in the attribution of morality. *Social Cognition*, *38*(2), 146–178. https://doi.org/10.1521/soco.2020.38.2.146

Rusconi, P., Sacchi, S., Capellini, R., Brambilla, M., & Cherubini, P. (2017). You are fair, but I expect you to also behave unfairly: Positive asymmetry in trait-behavior relations for moderate morality information. *PLOS ONE*, *12*(7), e0180686. https://doi.org/10.1371/journal.pone.0180686

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*(5), 867–878. https://doi.org/10.1002/ejsp.393

Sanderson, J., Gignac, G., & Ecker, U. (2021). Working memory capacity, removal efficiency and event specific memory as predictors of misinformation reliance. *Journal of Cognitive Psychology*, 1–15. https://doi.org/10.1080/20445911.2021.1931243

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. https://doi.org/10.1007/BF01068419

Schul, Y., & Mayo, R. (2014). Discounting information: When false information is preserved and when it is not. In D. N. Rapp & J. L. G. Braasch (Eds.), *Processing inaccurate information* (pp. 203–222). The MIT Press; JSTOR. https://doi.org/10.2307/j.ctt9qf9b7.14

Skowronski, J. J. (2002). Honesty and intelligence judgments of individuals and groups: The effects of entity-related behavior diagnosticity and implicit theories. *Social Cognition*, *20*, 136–169. https://doi.org/10.1521/soco.20.2.136.20993

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Skowronski, J. J., & Carlston, D. E. (1992). Caught in the act: When impressions based on highly diagnostic behaviours are resistant to contradiction. *European Journal of Social Psychology*, *22*(5), 435–452. https://doi.org/10.1002/ejsp.2420220503

Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review*, *96*(1), 58–83. https://doi.org/10.1037/0033-295X.96.1.58

Steblay, N., Hosch, H. M., Culhane, S. E., & McWethy, A. (2006). The impact on juror verdicts of judicial instruction to disregard inadmissible evidence: A meta-analysis. *Law and Human Behavior*, *30*(4), 469–492. https://doi.org/10.1007/s10979-006-9039-7

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2021). Processing political misinformation: Comprehending the Trump phenomenon. *Royal Society Open Science*, *4*(3), Article 3. https://doi.org/10.1098/rsos.160802

Tay, L. Q., Hurlstone, M. J., Kurz, T., & Ecker, U. K. H. (2022). A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *British Journal of Psychology*, *113*(3), 591–607. https://doi.org/10.1111/bjop.12551

Thorson, E. A. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*(3), 460–480. https://doi.org/10.1080/10584609.2015.1102187

van Oostendorp, H., & Bonebakker, C. (1999). Difficulties in updating mental representations during reading news reports. In *The construction of mental representations during reading* (pp. 319–339). Lawrence Erlbaum Associates Publishers.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423–441. https://doi.org/10.1080/03637751.2018.1467564

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*, *47*(2), 155–177. https://doi.org/10.1177/0093650219854600

Weeks, B. E., & Garrett, R. K. (2014). Electoral consequences of political rumors: Motivated reasoning, candidate rumors, and vote choice during the 2008 U.S. presidential election. *International Journal of Public Opinion Research*, *26*(4), 401–422. https://doi.org/10.1093/ijpor/edu005

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2-book.org/

Wilkes, A., & Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *Quarterly Journal of Experimental Psychology*, *40*, 361–387. https://doi.org/10.1080/02724988843000168

Ybarra, O. (2001). When first impressions don't last: The role of isolation and adaptation processes in the revision of evaluative impressions. *Social Cognition*, *19*(5), 491–520. https://doi.org/10.1521/soco.19.5.491.19910