PAPER IN PRESS AT THE JOURNAL OF APPLIED RESEARCH IM MEMORY AND COGNITION

Do False Allegations Persist? Retracted Misinformation Does Not Continue to Influence Explicit Person Impressions

Ullrich K. H. Ecker & Arnold E. Rodricks

School of Psychological Science

The University of Western Australia

Word Count: 7,422 (main text including footnotes and figure captions); 3,000 (Introduction and Discussion sections)

Corresponding author: Ullrich Ecker, School of Psychological Science, University of Western Australia (M304), 35 Stirling Hwy, Perth WA 6009, Australia; ullrich.ecker@uwa.edu.au

Abstract: Corrected misinformation often continues to influence reasoning; this is known as the continued-influence effect (CIE). It is unclear whether this effect also occurs in impression formation, with some arguing that person impressions are readily updated. The present study tested if a retracted allegation influences person impressions. Participants received examples of behaviors that a fictitious person had allegedly engaged in. The set did or did not include a domestic-violence behavior, which subsequently was or was not retracted. Discredited misinformation was found to influence neither trait ratings of the person nor behavior predictions. This held even when the person’s name implied a cultural background stereotypically associated with domestic violence. This provides evidence that under some circumstances, people can fully discount discredited misinformation when building person impressions. However, there was some tentative evidence that corrected misinformation did influence a more indirect measure, namely ratings of the fictitious person’s face.

Keywords: Continued influence effect; misinformation; false allegations; person impressions; impression formation

General audience summary: In the contemporary information landscape, misinformation can spread quickly and widely. This is of concern because research has shown that misinformation often continues to influence people's reasoning even after clear and credible retractions are provided—in other words, misinformation has the tendency to "stick." For example, if people initially receive information that a plane crash was caused by a terror attack but later learn that this information was false, they will often continue to make inferences regarding the event as if it were indeed terror-related. One type of misinformation that has the potential to be especially harmful are false allegations. However, it is unclear whether false allegations are equally "sticky" as other types of misinformation, such as misinformation relating to the cause of an event. Some researchers have even argued that person impressions are particularly malleable, and that people can readily update their impressions of others when they receive new information about the person. Therefore, the present study tested whether a retracted allegation influences person impressions. Participants received examples of behaviors that a fictitious person had allegedly engaged in. These behaviors did or did not include a domestic-violence behavior, which subsequently was or was not retracted. It was found that discredited misinformation had no ongoing influence on person impressions or behavior predictions. This held even when the person's name implied a cultural background stereotypically associated with domestic violence. This suggests that at least under some circumstances, people can fully discount discredited misinformation when building person impressions. However, there was some tentative evidence that corrected misinformation did result in the person's alleged face being rated more negatively.

Do False Allegations Persist? Retracted Misinformation Does Not Continue to Influence Explicit Person Impressions

How do false allegations influence your impressions of people? If you heard that a person behaved in an unacceptable manner, but later learned that this information was false, would the misinformation continue to influence your impression of this person, your predictions of their future behavior, and your perception of their face? The current study aimed to answer these questions in an investigation into the influence of corrected misinformation on impression formation.

False allegations are a type of misinformation. It is known that discredited misinformation often continues to influence reasoning despite clear retractions and intact subsequent memory of the retractions (e.g., Ecker, Lewandowsky, & Tang, 2010; Rich & Zaragoza, 2016). This has been termed the continued-influence effect (CIE; Chan, Jones, Hall Jamieson, & Albarracín, 2017; Johnson & Seifert, 1994; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Walter & Tukachinsky, 2020). The CIE is typically concerned with memory for and reasoning about events—in a standard CIE experiment, often the cause of a fictitious event is retracted, and inferential reasoning about the event continues to be influenced by the initially-presented cause despite the retraction, relative to a no-misinformation control condition (e.g., Ecker, Lewandowsky, & Apai, 2011).

One theoretical account of the CIE has argued that the effect arises from memory-updating failures (Ecker, Lewandowsky, Chang, & Pillai, 2014; Gordon, Brooks, Quadflieg, Ecker, & Lewandowsky, 2017). It assumes that when people encode information about an event, they gradually build a mental model of the event (Bower & Morrow, 1990; van Dijk & Kintsch, 1983). A credible retraction thus requires updating of the event model and removal of the retracted information from the model, or at least its tagging as false (see Ecker, Lewandowsky, Swire, & Chang, 2011; Morrow, Bower, & Greenspan, 1989; Rapp &

Kendeou, 2007). As retractions threaten model coherence—because removing a critical piece of information from a mental model can cause a model gap and thus render the model incomplete—updating can be inadequate (i.e., there is only partial updating, or updating only in participants not overly concerned by model incoherence or those who find the retraction fully credible; Connor Desai, Pilditch, & Madsen, 2019; Ecker et al., 2014; O'Rear & Radvansky, 2020; Rich & Zaragoza, 2016; Seifert, 2002).

In line with the CIE literature, research on belief perseverance has suggested that corrected allegations and rumors continue to influence judgments of a target person (Asch, 1946; Dreben, Fiske, & Hastie, 1979; also see Hogarth & Einhorn, 1992; van Overwalle & Labiouse, 2004). To illustrate, false feedback regarding a person's performance can continue to affect evaluations after the feedback has been declared false (Ross, Lepper, & Hubbard, 1975), incriminating evidence about a defendant that is later ruled inadmissible can affect jury decisions (Steblay, Hosch, Culhane, & McWethy, 2006), and unsubstantiated rumors can affect hiring decisions (Dalal, Diab, & Tindale, 2015; Miron-Shatz & Ben-Shakhar, 2008). Thorson (2016) found continued influence of negative misinformation on evaluations of political candidates.[1] These findings imply that person impressions are difficult to change, potentially resulting in continued influence of false allegations.

However, there is also reason to question whether CIEs occur reliably with person misinformation. First, on a theoretical level, the impression-formation literature suggests that building a mental model of a person may differ from building a mental model of an event. While building an event model requires that the temporal sequence of sub-events and their causal interrelatedness is maintained, impression formation refers to the integration of various pieces of information to form a more global representation (Fiske & Neuberg, 1990; Kashima

[1] We note that effects of *positive* misinformation about political figures may be more readily corrected or even over-corrected (Cobb, Nyhan, & Reifler, 2013).

& Kerekes, 1994). When people observe or learn about another person's behavior, they spontaneously make inferences about the person's traits (Srull & Wyer, 1989; Uleman, Newman, & Moskowitz, 1996), gradually forming a multi-faceted impression that is used to understand and predict their behavior (Park, 1986; Sjovall & Talk, 2004). The retraction of a specific trait description or behavior may thus prompt model updating without any threat to model coherence, as coherence can be maintained through other—unchallenged—dimensions of the trait model.

Second, findings of continued influence of false allegations may depend on specific task features. For example, a jury may find value in evidence even if it is ruled inadmissible (Kassin & Sommers, 1997), and rumors might turn out to be true even if they are unsubstantiated. It may be worthwhile under such circumstances to keep the initial information active (see Mensink & Rapp, 2011).

Third, much research in the person-impression literature has demonstrated that trait models *can* be readily updated when refutational information is presented. For example, in Gregg, Seibt, and Banaji (2006), participants learned about two groups, one described as aggressive and the other as pacific. After initial impressions were formed, it was clarified that group descriptions had erroneously been reversed; person impressions were successfully updated. In Rapp and Kendeou (2007), participants read passages providing trait-relevant behavioral information about a character. Results from a behavior-prediction task showed that readers readily revised their trait models when given refutational information.

It is thus unclear whether corrected misinformation has the same kind of ongoing influence in impression formation as it does in event-related reasoning. It is important to note in this regard that some studies suggesting that impressions are readily updated have relied on attitude formation through group memberships (e.g., Gregg et al., 2006), which may be less susceptible to primacy effects than attitudes regarding specific individuals (Hamilton &

Sherman, 1996). Also, declaring an entire trait model false due to reversal of group descriptions—as opposed to retraction of a specific behavior or trait—will make trait-model updating particularly likely (also see Ecker, Oberauer, & Lewandowsky, 2014; Kessler & Meiran, 2008). Studies investigating attitudes regarding individuals (e.g., McConnell et al., 2008) have often employed a variant of the attitude-learning paradigm (Kerpelman & Himmelfarb, 1971), in which counter-attititudinal behaviors are provided to change impressions (e.g., a person is first associated with negative, then positive behaviors) rather than direct corrections. Some studies have also applied weak corrections, declaring the misinformation not explicitly-false but merely "inadmissible" or "unsubstantiated." The main aim of the present study was thus to investigate whether continued influence occurs in a person impression task that is methodologically similar to a standard CIE task in that a specific piece of misinformation is explicitly provided and directly retracted.

We applied a variety of measures that range from direct to more indirect. The reason for this is that the person-impression literature suggests that explicit impressions are revised more readily than implicit impressions, which may be less malleable (Golding, Fowler, Long, & Latta, 1990; McConnell, Rydell, Strain, & Mackie, 2008; Rydell, McConnell, Mackie, & Strain, 2006; Wyer, 2010; Wyer, 2016). For example, in Gregg et al.'s (2006) experiments, only explicit but not implicit impressions were successfully updated. In Rapp and Kendeou (2007), while the direct behavior-prediction task suggested efficient trait-model revision, an indirect measure (reading times of trait-consistent versus inconsistent outcome statements) provided only mixed evidence of trait-model revision following refutational statements. In a conceptual replication, Rapp and Kendeou (2009) found evidence for trait-model revision in indirect reading-time and lexical-decision measures, but only after particularly strong refutations that offered an alternative explanation for the initial behavioral information, not after plain refutations.

The discrepancy between direct and indirect measures may arise from the more deliberate nature of responses in direct tasks, which can take into account the perceived validity of evidence and more meta-cognitive influences such as social desirability (Gawronski & Bodenhausen, 2006; Petty, Tormala, Briñol, & Jarvis, 2006). However, we also note that work by Ferguson and colleagues has found that initial implicit evaluations can be undone as long as the corrective information is highly plausible (Cone, Flaharty, & Ferguson, 2019) and participants have sufficient cognitive resources and motivation for reinterpretation (Mann & Ferguson, 2015). Indeed, both task characteristics other than their direct/indirect nature (e.g., whether a judgment is required) and participant motivation to maintain an accurate model have been discussed as important factors determining whether trait-model updating occurs (Mensink & Rapp, 2011; Rapp & Kendeou, 2007, 2009).

In sum, it is unclear whether corrected negative misinformation about a person will continue to affect direct person-related judgments—as suggested by the CIE literature (see especially, Thorson, 2016)—or whether people can readily discount misinformation about a person—as suggested by the broader person-impression literature. In light of the available evidence (e.g., Mann & Ferguson, 2015; Rapp & Kendeou, 2007, 2009), it is also not fully established that corrected misinformation will have continued impact on more indirect person-impression measures.

To investigate these questions, the present study used an impression-formation task, during which a target person was or was not accused of an aggressive behavior; if this critical information was provided, it was or was not later retracted directly. In addition to standard direct trait ratings (how aggressive is the person?), we used behavior predictions (how will the person respond in this conflict situation?), which can be considered slightly less direct (McCarthy & Skowronski, 2011; Newman, 1996), and face ratings (how aggressive is this face?), which are decoupled from trait-based person impression and thus provide a more indirect measure (Paunonen, 2006). Valence ratings of faces have been shown to be sensitive to corrected misinformation (Ecker et al., 2014).

In line with previous CIE research, it was hypothesized that retractions would only be partially effective; we expected that violence-related information would have a negative effect on trait ratings, behavior predictions, and face ratings, and that a clear retraction would only partially offset these effects relative to a no-misinformation control condition. However, based on the impression-formation literature, the alternative hypothesis was entertained that retracted misinformation may only affect face ratings but not the more direct trait ratings and behavior predictions.

Experiment 1

Participants were provided with information about a fictitious person named “John” and were asked to form an impression of this person. The information comprised descriptions of specific behaviors that John had allegedly engaged in. These descriptions were purportedly provided by John’s peers. The information given depended on the experimental condition: in the control condition (C), the information referred only to neutral behaviors, whilst the information in the negative (N) and retraction (R) conditions also referred to a negative behavior. In all conditions, participants were subsequently presented with updated information. During this phase, the negative behavior was discredited in the retraction condition but not the negative condition. The experiment thus used a one-factorial between-subjects design with three levels (C, N, R). Using their impression of John, participants then (1) completed a questionnaire rating John’s likability and aggressiveness, (2) made predictions as to how John would behave in a range of scenarios, and (3) rated the perceived dominance, aggressiveness, attractiveness, and trustworthiness of a face allegedly belonging to John.

Method

Participants. A-priori power analysis (using G*Power 3; Faul, Erdfelder, Lang, & Buchner, 2007) suggested a minimum sample size of 111 participants in order to detect a medium-size effect, f = .30 (α = .05, 1-β = .80). A total of N = 142 participants were pseudo-randomly allocated to one of the three conditions (n_C = 48, n_N = 47, n_R =47). Participants were undergraduate students from the University of Western Australia (UWA), who participated for course credit. Participants provided informed consent after reading an ethically-approved information sheet. The sample comprised 33 males and 108 females; age range was 17-57 years (M = 20.43; SD = 6.55).

Materials.

Behavior set. The behavior set used in the impression-formation task comprised 11 specific behaviors that John had allegedly engaged in over the past month (10 neutral; 1 negative). To avoid having only non-diagnostic items, some of the neutral behaviors were slightly positive or negative. Example neutral behaviors were that John “went to the local bar on a Friday night”, “parked his car in a no-parking zone at the shops”, or “gave an engaging presentation to his class.” In comparison, the negative behavior—that John had “slapped his girlfriend during an argument”—was clearly negative and indicative of high levels of aggression. The full list of behaviors is provided in Appendix A.

Reysen Likability Scale. The Reysen Likability Scale (henceforth referred to simply as the “likability scale”) measures the degree of likability of a target source (Reysen, 2005; see Appendix B). Participants responded to 11 items (e.g., “This person is friendly”) on a seven-point Likert scale from “very strongly disagree” (0) to “very strongly agree” (6). Higher scores indicated higher levels of likability. The likability scale was found to be highly reliable (Cronbach’s α = .88).

Aggressiveness scale. The aggressiveness scale was designed to measure perceived aggressiveness of a target source (see Appendix C). Participants responded to 10 items (e.g., "He often struggles with controlling his temper") adapted from the Buss Perry Aggression Scale (Buss & Perry, 1992) using a five-point Likert scale from "extremely uncharacteristic" (0) to "extremely characteristic"(4). Higher scores on these items indicate higher levels of perceived aggressiveness. Items were selected from the full scale depending on their ability to be modified in order to measure the perceived aggressiveness of a target person. The 10 aggressiveness items were interspersed with eight distractor items, adapted from the Boredom Proneness Scale (Farmer & Sundberg, 1986), to obscure the purpose of the scale and reduce potential demand characteristics. The distractor items were not included in the subsequent analyses. The aggressiveness scale was found to be highly reliable (Cronbach's α = .86).

Behavior-prediction task. Participants were presented with nine scenarios involving John. Each scenario described a social situation; five scenarios were related to the negative behavior of interest (i.e., they were aggression-related), while the remaining four were filler scenarios, which were included to mask the true purpose of the task (see Appendix D for the full description of scenarios). For example, one of the aggression-related scenarios involved someone spilling a drink on John at a bar; one of the filler scenarios involved a co-worker asking John for a favor. The scenarios were presented in a random order.

For each scenario, participants were presented with three possible behavioral responses, namely an aggressive response, a passive response, and an affirmative but non-aggressive response. For example, in response to a stranger spilling a drink on John, it was suggested John could (1) push the stranger, (2) ignore the stranger, or (3) alert the stranger to what they have done and then make a joke about it. Participants were asked to predict the likelihood with which John would show each behavior, on a scale ranging from 0 (very unlikely) to 10 (very likely). Ratings of the aggressive responses were used to calculate predicted aggression (ratings of the passive and non-aggressive responses were discarded).

Face-rating task. Participants were presented with an image of a face that allegedly belonged to John. To avoid face-specific effects, three different faces were used, with one being selected at random for each participant. The three face images were taken from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015). The images were colored frontal head shots of neutral-expression faces of young adult Caucasian males against a white background. The faces were selected such that their database validation ratings (from Ma et al., 2015) were close to average on four dimensions considered central to the formation of facial impressions: dominance, aggressiveness, attractiveness, and trustworthiness (Oosterhof & Todorov, 2008). The selected faces had the following validation ratings on 1-7 scales: dominance, 2.63-3.10 (overall mean was *M* = 2.94 [*SE* = .07]); aggressiveness, 2.29-2.48 (*M* = 2.40 [*SE* = .06]); attractiveness, 2.80-3.05 (*M* = 2.96 [*SE* = .06]); and trustworthiness, 2.95-3.42 (*M* = 3.21 [*SE* = .04]). In the current study, participants rated “John’s” face on these four dimensions (on 0-10 scales), following Ecker et al. (2014). In line with previous research, mean ratings across dominance and aggressiveness scales were used to calculate a “dominance” factor, and mean ratings across attractiveness and trustworthiness scales were used to calculate a “valence” factor (see Oosterhof & Todorov, 2008; Todorov, Said, & Verosky, 2011). In light of Ecker et al.’s findings, we predicted misinformation effects primarily on the valence factor (also see Gross & Crofton, 1977; Paunonen, 2006); however, given that the misinformation in the present study related directly to aggression/dominance, we also tested for effects on the dominance factor.

Recognition test. A recognition test was administered to ensure adequate encoding of behaviors. The test contained 11 multiple-choice questions about the initial and updated behaviors. Questions regarding the initial behaviors required participants to select the correct answer from three choices. For example, participants were asked “What day did he (John) go to the local bar?”. Questions regarding the updated information required participants to select one or more answers from four choices. For example, participants were asked “Which of these behaviors were retracted?” and were then provided with four choices (see Appendix E for the full list of questions). Scores were calculated by dividing the total number of correct responses by the total number of questions.

Procedure. Participants arrived in the lab in groups of 1-5, but were tested on individual computers. The experiment was administered using Qualtrics survey software (Qualtrics, Provo, UT). All verbatim task instructions are provided in Appendix F.

Phase 1: Cover story. Participants were told that the study investigated how people form impressions of others; they were informed they would be given information about a person named “John”. It was implied that John was a student peer, and the information had been collected from four of his acquaintances who had each provided examples of his behavior that they had observed over the past month. Participants were instructed to use this information to form an impression of John.

Phase 2: Presentation of initial behavior set. Eight behavior descriptions were presented one-by-one for participants to read and form an impression of John. Participants were instructed to think about the significance of the behaviors and how they shaped their impression of John. Each behavior was presented for a minimum of five seconds. All conditions featured seven neutral behaviors, presented in a random order, but conditions differed with respect to the behavior presented in position 4, which was the critical negative behavior in the negative and retraction conditions, but another neutral behavior in the control condition (see Appendix A for details). Following the presentation of all behaviors, participants were asked to generate three words that described John in order to ensure they integrated the received information into an impression (the words were not analyzed).

Phase 3: Presentation of updated information. Participants were informed that in order to verify the initially presented information, and to learn more about John, him and his acquaintances were interviewed a second time as a group. They were told that based on this, they would receive additional information, and that some behaviors from the initial set may be confirmed, some may be retracted, and there may also be some new behaviors not previously mentioned. Participants were instructed to use this new information to update their impression of John. The updated information comprised six statements, presented one-by-one. In all conditions, two new additional neutral behavior descriptions were presented, two neutral behaviors from the initial set were confirmed, and two behaviors from the initial set were retracted. For example, a confirmation was provided in the following format: “CONFIRMATION: John DID go to the bar on a Friday night”; similarly, a rejection was provided as “REJECTION: John DID NOT slap his girlfriend during an argument.” Conditions differed in regards to which behaviors were retracted: in the control and negative conditions, two neutral behaviors were retracted; in the retraction condition, one neutral behavior and the critical negative behavior were retracted (see Appendix A for details). The order of the six items was random; however, the retraction of the negative behavior (in the retraction condition) always occurred in the fourth position.

Phase 4: Test phase. After completing a short unrelated distractor task, participants completed the likability scale, the aggressiveness scale, the behavior-prediction task, the face-rating task, and the recognition test, in that order. Participants were then fully debriefed.

Results

Recognition. Scores from the recognition test were used mainly to ensure adequate encoding of behaviors. Performance on the test was relatively high (*M* = .92, *SE* = .01). Only one participant scored < 0.5; excluding them did not affect the outcome of any analysis, so only analyses of the full-sample data will be reported. A single-factor between-subjects

ANOVA on recognition scores yielded a non-significant main effect, *F*(2,139) = 1.14, *MSE* = .01, *p* = .32, indicating that there were no significant differences between conditions.

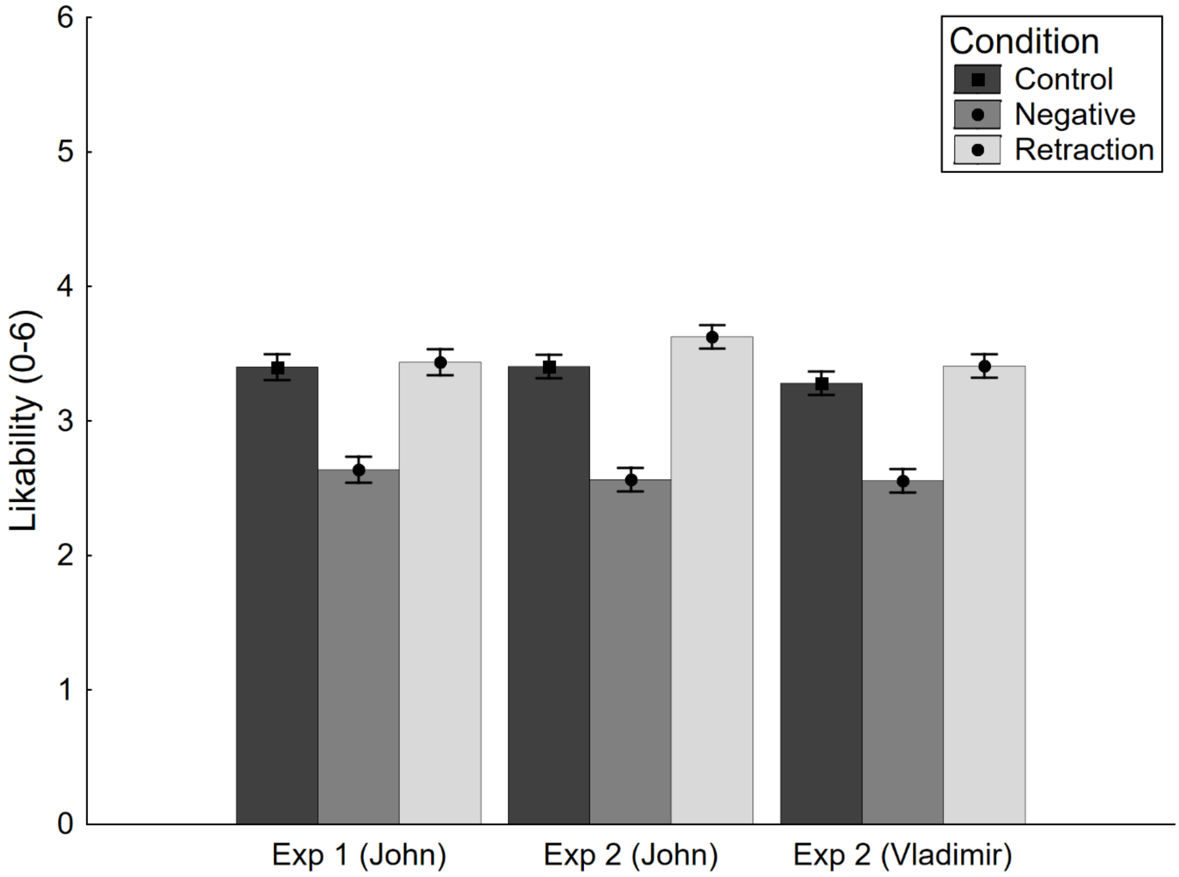

Figure 1. Mean likability scores across conditions in Experiments 1 and 2. Error bars indicate standard errors of the mean.

Likability. Likability scores were calculated as the mean of the 11 items on the likability scale, and the score thus ranged from 0-6. The mean likability scores across conditions are presented in Figure 1 (left-hand panel). A between-subjects ANOVA yielded a significant main effect of condition, *F*(2,139) = 21.29, *MSE* = 0.45, *p* < .001, η_p^2 = .23.[2] Planned contrasts confirmed lower likability in the negative condition (M_N = 2.64; *SE* = .10)

[2] As the negative behavior was a male-perpetrated violent behavior against a woman, we speculated post-hoc that participant gender may have an impact, and thus repeated this and all following analyses with a (dichotomous) gender factor included. There were no interactions with gender, with one minor exception reported in Experiment 2.

relative to control (M_C = 3.40; *SE* = .10), *F*(1,139) = 30.59, *p* < .001, as well as lower likability in the negative condition relative to the retraction condition (M_R = 3.44; *SE* = .10), *F*(1,139) = 33.33, *p* < .001. There was no difference between retraction and control conditions, *F* < 1, indicating the retraction was fully effective, and retracted misinformation had no impact on likability.

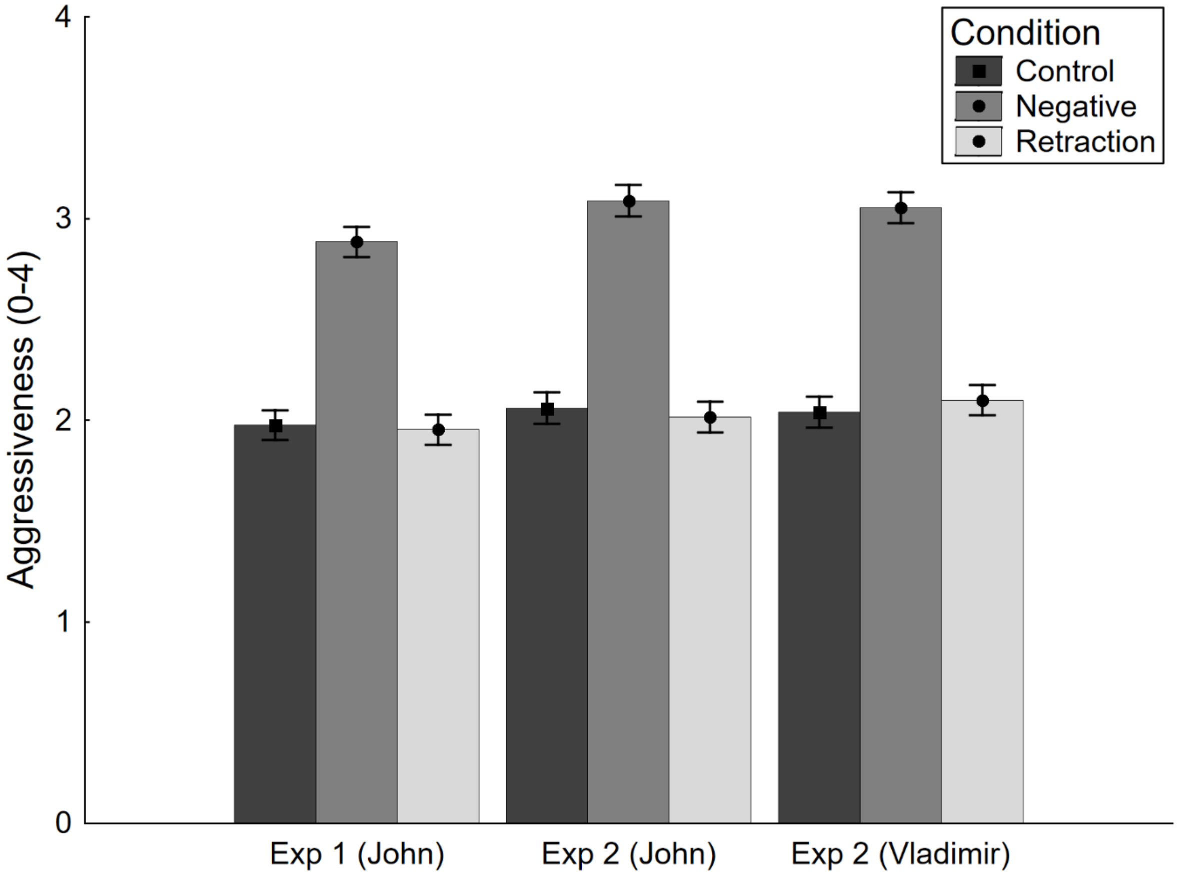

Figure 2. Mean aggressiveness scores across conditions in Experiments 1 and 2. Error bars indicate standard errors of the mean.

Aggressiveness. Aggressiveness scores were calculated as the mean of the 10 items on the aggressiveness scale, and scores thus ranged from 0-4. Mean aggressiveness scores across conditions are presented in Figure 2 (left-hand panel). A between-subjects ANOVA returned a main effect of condition, *F*(2,139) = 50.68, *MSE* = 0.26, *p* < .001, η_p^2 = .42. Mimicking the likability data, planned comparisons revealed that rated aggressiveness was significantly greater in the negative condition (M_N = 2.89; *SE* = .07) compared to both control (M_C = 1.98; *SE* = .08), *F*(1,139) = 74.56, *p* < .001, and retraction conditions (M_R = 1.95; *SE* = .07), F(1,139) = 77.72, *p* < .001, which did not differ from each other, *F* < 1. This indicated that retracted misinformation had no influence on perceived aggressiveness.

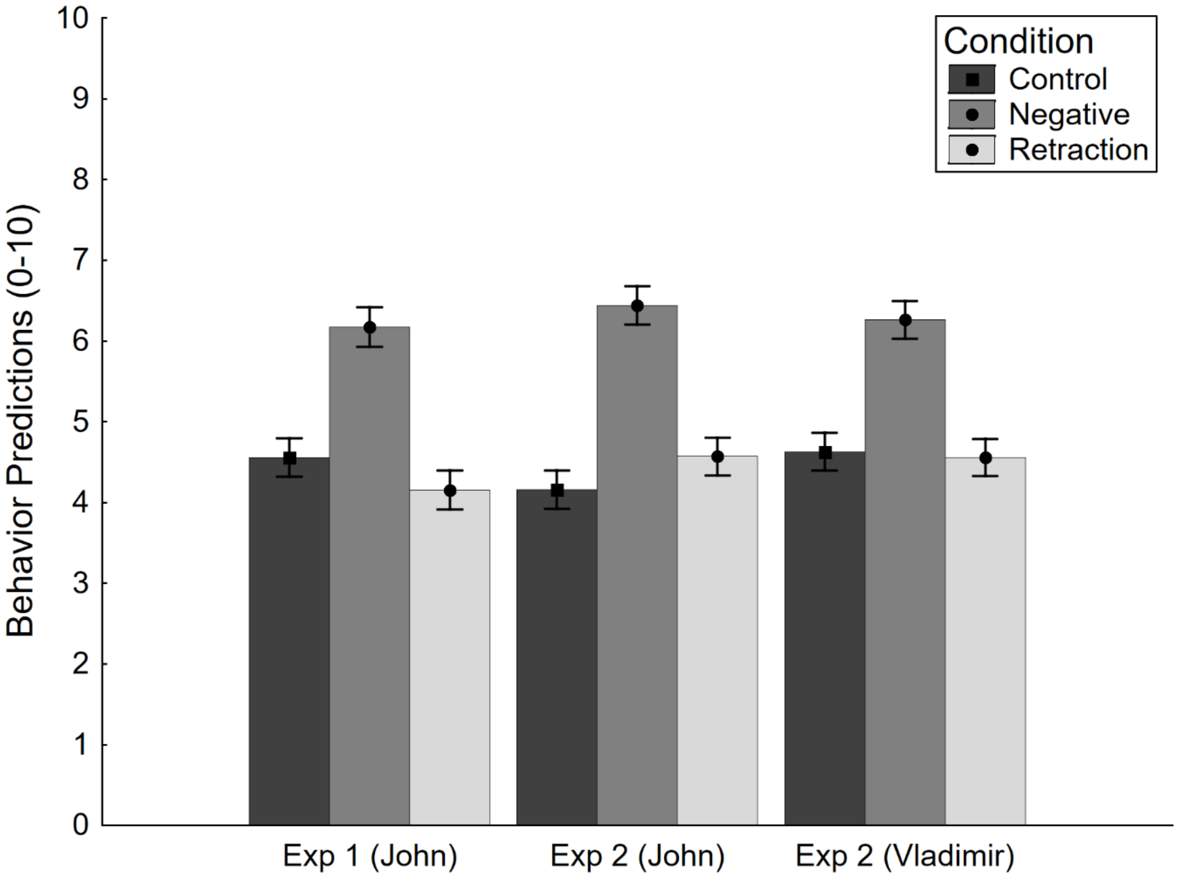

Figure 3. Mean aggressive-behavior prediction scores across conditions in Experiment 1. Error bars indicate standard errors of the mean.

Behavior predictions. Predicted-aggression scores were calculated as the mean of the five aggression-related items in the behavior-prediction task, and scores thus ranged from 0-10. Mean predicted-aggression scores across conditions are presented in Figure 3 (left-hand panel). A between-subjects ANOVA yielded a significant main effect of condition, *F*(2,139) = 19.45, *MSE* = 2.77, *p* < .001, η_ρ^2 = .22. Planned comparisons showed that predicted aggression was significantly greater in the negative condition (M_N = 6.17; *SE* = .24) in comparison to both the control condition (M_C = 4.56; *SE* = .24), *F*(1,139) = 22.41, *p* < .001, and the retraction condition (M_R = 4.15; *SE* = .24), *F*(1,139) = 34.68, *p* < .001, which did not differ from each other, *F*(1,139) = 1.41, *p* = .24. This indicates that retracted misinformation did not influence behavior predictions.

Face ratings. A dominance score was computed by calculating mean ratings across dominance and aggressiveness scales; a valence score was computed by calculating mean ratings across attractiveness and trustworthiness scales. Scores varied from 0-10. Observed mean dominance and valence scores and standard errors are provided in Table 1. A between-subjects ANOVA on dominance scores did not yield a significant effect of condition, *F* < 1. Planned contrasts were not run due to the clear non-significance of the main effect.

Table 1

Descriptive Face Rating Statistics from Experiments 1 and 2

Face Ratings	Condition	*M*	*SE*
Dominance			
Exp. 1 (John)	Control	4.92	0.28
	Negative	5.21	0.22
	Retraction	5.07	0.24
Exp. 2 (John)	Control	4.38	0.24
	Negative	5.06	0.29
	Retraction	5.15	0.29
Exp. 2 (Vladimir)	Control	4.81	0.23
	Negative	5.19	0.23
	Retraction	5.30	0.26
Valence			
Exp. 1 (John)	Control	4.54	0.25
	Negative	3.85	0.22
	Retraction	4.00	0.22
Exp. 2 (John)	Control	4.71	0.20
	Negative	3.69	0.21
	Retraction	4.11	0.20
Exp. 2 (Vladimir)	Control	4.10	0.23
	Negative	3.89	0.22
	Retraction	4.95	0.21

A between-subjects ANOVA on valence scores returned a marginal main effect of condition, $F(2,139) = 2.54$, $MSE = 2.48$, $p = .08$, $\eta_p^2 = .04$. Planned contrasts showed that John's face was rated more negatively in the negative condition ($M_N = 3.85$; $SE = .23$) than the control condition ($M_C = 4.54$; $SE = .23$), $F(1,139) = 4.57$, $p = .03$—however, applying Holm-Bonferroni correction, this can be considered a statistically non-significant result. There was no difference between the retraction condition ($M_R = 4.00$; $SE = .23$) and the negative condition, $F < 1$. The difference between retraction and control conditions was also statistically non-significant, $F(1,139) = 2.81$, $p = .10$.

Discussion

Results from the trait and behavior-prediction measures did not support the hypothesis that corrected misinformation would continue to influence person impressions and impression-related reasoning. On the contrary, all direct measures consistently showed full discounting of the misinformation; as such, in this experiment, participants were able to successfully update their explicit person impressions following a retraction. To some extent, methodological differences between studies may be able to explain some of the discrepancy with previous results. For example, in Dalal et al. (2015), the rumors that affected hiring decisions were not explicitly retracted but merely labeled as "unconfirmed." In jury studies investigating the impact of inadmissible evidence, the mock jurors may believe that retracted evidence is still relevant (Kassin & Sommers, 1997). In Thorson (2016), misinformation about a political candidate was always worldview-congruent—that is, participants were told that the political candidate belonged to the party opposite to their own political partisanship. As a result, pre-existing attitudes may have contributed to the observed CIE, which has been shown to be sensitive to worldview effects (see Ecker & Ang, 2019). By contrast, in the present study, misinformation was explicitly retracted, participants had no reason to second-guess the retraction, and no pre-existing attitudes towards "John"; these factors may have facilitated impression updating. The results of Experiment 1 were, however, consistent with the person-impression literature, which has suggested that people can readily update their impressions of people when information conflicting with the existing impression is presented (e.g., Golding et al., 1990; Wyer, 2016).

Experiment 1 also provided some suggestion that misinformation about a person can influence face ratings (in line with Paunonen, 2006); similar to Ecker et al. (2014), this influence seemed limited to the valence factor (i.e., attractiveness and trustworthiness dimensions). However, it must be acknowledged that the effect was statistically non-significant after correcting for multiple comparisons, and thus requires replication before drawing any conclusions. Unlike the direct measures, on this more indirect measure, the retraction condition did not differ from the negative condition—that is, assuming there was a true effect of the misinformation, the retraction was ineffective. While this supports the hypothesis that explicit impression measures are more readily updated than indirect measures (Gregg et al., 2006; Wyer, 2010), the retraction condition also did not differ significantly from control, precluding strong conclusions to be drawn.

To corroborate the findings from Experiment 1, it was decided to run a replication study. Experiment 2 replicated Experiment 1 exactly, but also added an additional name manipulation. As suggested earlier, one reason why Experiment 1 may have yielded different results from the study by Thorson (2016) may lie in differences in participants' pre-existing attitudes. Thus, in Experiment 2, participants were either asked to form an impression of "John" or "Vladimir"—a name that pilot testing found to be associated with a more negative attitude, and specifically more stereotypically linked to domestic-violence behavior. This was done in order to test whether under such conditions a CIE of violence-related misinformation on impression formation may arise.

Experiment 2

Pilot testing was conducted to find a name that may be more stereotypically associated with domestic violence than “John”. It is well-known that people infer personality characteristics from first names (e.g., Leirer, Hamilton, & Carpenter, 1982; Mehrabian, 2001; Sidhu, Deschamps, Bourdage, & Pexman, 2019). Also, names serve as proxies for socio-cultural background (e.g., Cotton, O’Neill, & Griffin, 2008), and while domestic violence is a serious problem in many Western countries including Australia (Phillips & Vandenbroek, 2014), domestic violence and in particular violence against women is even more common in some parts of Africa, South Asia, and some countries of the former Soviet Union (OECD, 2018; UN Women, 2011). We thus selected a range of common male names associated with these regions (specifically: John, Rahul, Ahmed, and Vladimir) and conducted a pilot survey.

A total of $N = 31$ participants were told that we were investigating the role that names play in the perception of people; they were asked to make 2AFC responses to five questions, including the question “Who is more likely to slap their girlfriend in an argument?”[3]. Each question was posed six times: once with each possible name pair; each name therefore appeared with the violence-related question three times. The overall expected frequency of any particular name being chosen by chance in response to the violence question was thus $31 \times 3 / 2 = 46.5$. The names John, Rahul, Ahmed, and Vladimir were chosen 34, 41, 35, and 76 times, respectively. There was a significant difference between the expected and observed frequency distributions, $\chi^2(3) = 25.57$, $p < .001$. This means that on average, a person called Vladimir was seen as more likely to slap his girlfriend, and therefore Experiment 2 utilized “John” and “Vladimir” as names. Experiment 2 thus used a 2 (name: John vs. Vladimir) × 3 (condition: C, N, R) between-subjects design. We hypothesized that corrected violence-

[3] Additional decoys were “Who is more likely to drop everything to help a friend in need?”; “Who is more likely to be caught speeding?”; “Who is more likely to make a charitable donation?”; “Who is more likely to be married with children?”

related misinformation may have continued influence only when “Vladimir” was accused, based on pre-existing stereotypical attitudes related to the name. We expected that when using the name “John”, the results of Experiment 2 would replicate the results of Experiment 1.

Method

Participants. A total of N = 287 participants were pseudo-randomly allocated to one of the six conditions (n_{JC} = 47, n_{JN} = 47, n_{JR} =48; n_{VC} = 48, n_{VN} = 48, n_{VR} =49). Participants were again undergraduate students from the University of Western Australia, who participated for course credit and had not participated in Experiment 1. The sample comprised 92 males, 194 females, and 1 person of undisclosed gender; age range was 17-50 years (M = 20.28; SD = 4.80).

Materials and procedure. Materials and procedure were identical to Experiment 1, apart from use of the name “Vladimir” in the respective conditions of Experiment 2.

Results

Recognition. Performance on the test was again high (M = .93, SE = .01). All participants scored at least .50 and so all participants were retained for analyses. A between-subjects ANOVA on recognition scores yielded no significant effects, all Fs < 1.21, ps > .30, indicating that there were no significant differences between conditions.

Likability. Mean likability scores across conditions are presented in Figure 1 (middle and right-hand panel). A between-subjects ANOVA yielded a significant main effect of condition, $F(2,281) = 69.05$, $MSE = 0.36$, $p < .001$, $\eta_\rho^2 = .33$. The main effect of name did not reach significance, $F(1,281) = 2.70$, $MSE = 0.36$, $p = .10$, $\eta_\rho^2 = .01$, and neither did the interaction, $F < 1$. Planned contrasts confirmed lower likability in the negative condition (M_{JN} = 2.56, SE = .09; M_{VN} = 2.55, SE = .09) relative to control (M_{JC} = 3.40, SE = .09; M_{VC} = 3.28, SE = .09), $F(1,281) = 80.83$, $p < .001$, as well as lower likability in the negative condition relative to the retraction condition (M_{JR} = 3.63, *SE* = .09; M_{VR} = 3.41, *SE* = .09), *F*(1,281) = 122.34, *p* < .001. There was a marginal difference between retraction and control conditions, *F*(1,281) = 4.10, *p* = .04, indicating slightly greater likability in the retraction condition. Thus, the retraction was fully effective, and retracted misinformation had no negative impact on likability, closely replicating Experiment 1.

Aggressiveness. Mean aggressiveness scores across conditions are presented in Figure 2 (middle and right-hand panel). A between-subjects ANOVA returned a main effect of condition, *F*(2,281) = 117.79, *MSE* = 0.28, *p* < .001, η_ρ^2 = .46. The main effect of name and the interaction effect were non-significant, *F* < 1. Planned comparisons revealed that rated aggressiveness was significantly greater in the negative condition (M_{JN} = 3.09, *SE* = .08; M_{VN} = 3.05, *SE* = .08) compared to both control (M_{JC} = 2.06, *SE* = .08; M_{VC} = 2.04, *SE* = .08), *F*(1,281) = 177.60, *p* < .001, and retraction conditions (M_{JR} = 2.02, *SE* = .08; M_{VR} = 2.10, *SE* = .08), *F*(1,281) = 176.39, *p* < .001, which did not differ from each other, *F* < 1. This replicated Experiment 1 exactly and indicated that retracted misinformation had no influence on perceived aggressiveness.

Behavior predictions. Mean predicted-aggression scores across conditions are presented in Figure 3 (middle and right-hand panel). A between-subjects ANOVA yielded a significant main effect of condition, *F*(2,281) = 42.80, *MSE* = 2.62, *p* < .001, η_ρ^2 = .23. The main effect of name and the interaction were non-significant, *F* < 1.03. Planned comparisons showed that predicted aggression was significantly greater in the negative condition (M_{JN} = 6.44, *SE* = .24; M_{VN} = 6.26, *SE* = .23) compared to both control (M_{JC} = 4.16, *SE* = .24; M_{VC} = 4.63, *SE* = .23), *F*(1,281) = 69.54, *p* < .001, and retraction conditions (M_{JR} = 4.57, *SE* = .24; M_{VR} = 4.56, *SE* = .23), *F*(1,281) = 58.61, *p* < .001, which did not differ from each other, *F* < 1. This replicated Experiment 1 and indicates that retracted misinformation did not influence behavior predictions.

Face ratings. Observed mean dominance scores and standard errors are provided in Table 1. A between-subjects ANOVA returned a small but significant main effect of condition, $F(2,281) = 3.35$, $MSE = 3.21$, $p = .04$, $\eta_\rho^2 = .02$. Planned comparisons showed that perceived dominance was lower in the control condition than both the negative condition, $F(1,281) = 4.13$, $p = .04$, and the retraction condition, $F(1,281) = 5.81$, $p = .02$, which did not differ from each other, $F < 1$. While only the control versus retraction condition difference remained significant after controlling for multiple comparisons, this result deviates from Experiment 1, where no condition differences were found. It suggests that a retraction was ineffective in reducing the impact from negative misinformation on perceived dominance.[4]

Observed mean valence scores and standard errors can also be found in Table 1. A between-subjects ANOVA returned a main effect of condition, $F(2,281) = 7.10$, $MSE = 2.14$, $p = .001$, $\eta_\rho^2 = .05$, which was qualified by a significant interaction, $F(2,281) = 5.87$, $p = .003$, $\eta_\rho^2 = .04$. Due to the interaction, planned contrasts were run separately for each name condition. Contrasts showed that John's face was rated more positively in the control condition compared to the negative condition, $F(1,281) = 11.48$, $p = .001$. The difference between the control condition and the retraction condition was marginal, $F(1,281) = 3.98$, $p \leq .05$, and non-significant after Holm-Bonferroni correction. There was no significant difference between the retraction condition and the negative condition, $F(1,281) = 1.99$, $p = .16$. The observed pattern thus closely resembled the pattern found in Experiment 1, but provided stronger statistical evidence that negative misinformation affected face-valence ratings, and that this effect was at least not fully undone by a retraction. By contrast, Vladimir's face was rated more positively in the retraction condition than the negative condition, $F(1,281) = 12.85$, $p < .001$, or the control condition,

[4] We note that this was the only analysis that yielded a significant interaction with gender. The condition main effect was driven mainly by male participants; restricted to the female sample, no main effects or interactions were significant, *F*s(1/2,188) < 2.14, *p*s > .12.

F(1,281) = 8.10, *p* = .005. The difference between the negative condition and control was non-significant, *F* < 1. This shows that the negative misinformation had no impact on valence ratings of Vladimir's face—although the data also suggest that a face is seen more negatively at baseline if it is assumed to belong to "Vladimir" as opposed to "John" (post-hoc test of name effect in control condition: *F*[1,281] = 4.12, *p* = .04). A retraction seemed to then undo the impact of the negative stereotype (and misinformation), as the valence rating of Vladimir's face in the retraction condition was comparable to the rating of John's face in the control condition.

Discussion

Overall, Experiment 2 replicated Experiment 1 closely, with retractions fully effective in counteracting the influence of negative misinformation on direct person-impression measures. The name manipulation had no effect, suggesting that retractions of negative misinformation can be effective even if there is congruence between the misinformation and stereotypical attitudes towards the evaluated person. The evidence regarding the more indirect face-rating measures was weaker, but suggested that retractions can be ineffective in counteracting the influence of negative misinformation on implicit impressions. The only exception was the face-valence ratings of "Vladimir", where the data suggested (as in the pilot study) that the name was associated with negative valence, which negative misinformation did not decrease further; a retraction here had the effect of boosting face-valence ratings up to the baseline level associated with "John". We can only speculate that this occurs because the positive retraction message counteracted not only the misinformation but also the negative stereotype.

Bayesian Analyses

As Bayesian analyses can quantify evidence in favor of a null hypothesis (see Wagenmakers et al., 2018), we ran a series of Bayesian *t*-tests on data combined across the identical control and retraction conditions of both experiments (i.e., including only the “John” conditions of Experiment 2). For each dependent variable, these tested whether a model including a condition factor that specified a continued-influence effect—that is, a difference between the control and retraction conditions in the direction of the negative condition (H_1)—would be preferable to a null model. The resulting Bayes factors, indicating evidence for or against inclusion of the condition factor, are shown in Table 2. A $BF_{01} < 1$ suggests evidence in favor of including a condition factor; a $BF_{01} > 1$ suggests evidence in favor of the null model. For example, $BF_{01} = 10$ would suggest that the data are 10 times more likely to have occurred under the null hypothesis; $BF_{01} = 0.10$ would suggest that the data are 10 times more likely to have occurred under the alternative hypothesis. *BF* values between 0.33 and 3 are taken to only provide anecdotal evidence; *BF* values between 0.1 and 0.33, or 3 and 10 constitute moderate/substantial evidence; *BF* values < 0.1 or > 10 provide strong to very strong evidence (Jeffreys, 1961; Wagenmakers et al., 2018).

Results showed that the two experiments yielded substantial to strong evidence *against* continued influence on direct trait and behavior measures, in line with the frequentist analyses that detected no difference between control and retraction conditions. By contrast, the evidence *for* continued influence on face-valence ratings was substantial.

Table 2

Results from Bayesian Analyses across Experiments 1-2

dV	Effect direction under H_1	BF_{01}
Likability	Control > Retraction	14.75**
Aggressiveness	Control < Retraction	8.48*
Behavior predictions	Control < Retraction	6.26*
Face dominance	Control < Retraction	0.83
Face valence	Control > Retraction	0.14*

Note. H_1, alternative hypothesis; * indicates substantial and ** indicates strong evidence for ($BF_{01} > 0$) or against ($BF_{01} < 0$) the null.

General Discussion

This study tested whether misinformation would produce a continued-influence effect on person-impression measures. Two experiments produced a clear result: A strongly-negative piece of misinformation had the expected effect on trait ratings and behavior-prediction measures, but no continued impact after being retracted. Thus, we demonstrate that direct person-impression measures can be unaffected by person-specific misinformation after its retraction. This suggests that previous findings of continued influence on person impressions may have been driven by worldview (Thorson, 2016), and potentially points to a boundary condition of the CIE (in line with Golding et al., 1990).

As outlined in the Introduction, one way to interpret this result is that forming and updating an impression of a person may differ from building and updating a mental model of an event. Specifically, while model-updating theory assumes that retraction of a core piece of information can threaten event-model coherence (e.g., Ecker et al., 2010), impression formation is a more holistic process (e.g., Park, 1986), where retraction of one piece of information may not produce comparable incoherence. Thus, continued influence may arise more readily with event reports because events by their very nature demand an explanation (*something* must have started the fire!), and it may therefore be rational to hold onto initially-provided causal information even if it is credibly retracted (Connor Desai et al., 2019), whereas credibly-retracted information about a person is less likely to still be of relevance post-retraction. This supports the notion of efficient impression updating (also see Mende-Siedlecki, 2018) and is in line with theoretical models that specify belief updating as a sequential anchoring-and-adjustment (Hogarth & Einhorn, 1992) or averaging process (Kashima & Kerekes, 1994). To corroborate this conclusion, future research could seek to develop materials that would allow testing for CIEs on event-related reasoning and person impression in the same study.

In contrast to the results from the direct tasks, we found tentative evidence for a CIE on face ratings. This meshes well with the claim that implicit impressions are more change-resistant than explicit impressions (e.g., Wyer, 2016) and broadly supports theories that assume fundamental differences between explicit and implicit attitudes (e.g., Wyer, 2010). Alternatively, it is possible that demand characteristics and other meta-cognitive factors may have had a stronger impact on the direct measures. Participants were explicitly instructed to update their impressions and respond based on the information provided. Thus, participants may have felt unable to justify negative-impression responses. They may have also tried to appear non-judgmental in order not to violate socio-cultural norms (e.g., Paulhus, 2002). Given the relatively-ambiguous evidence obtained, we can make no strong claims regarding the observed difference between direct and more indirect measures. We also acknowledge that our face ratings may not have tapped into "truly implicit" impressions, especially as they were administered after the direct measures, and given that task features other than the direct/indirect dimension may have contributed to results (Mensink & Rapp, 2011). Future research could contrast implied versus explicit updating instructions,[5] or incorporate behavioral measures, which may provide another avenue to reduce the impact of socially-desirable responding (e.g., measuring participants' behavior towards a confederate pretending to be "John").

Turning to applications, this research shows that a false allegation does not necessarily do lasting damage if a clear retraction is provided and people are not motivated to hold onto the misinformation. This may have implications for situations ranging from jury trials and defamation lawsuits to interpersonal relationships in social and organizational settings. However, some limitations must be considered. First, one should never place too

[5] We note, however, that in the broader memory-updating literature, updating difficulties are often demonstrated despite explicit instructions to keep mental representations up-to-date; e.g., Allanson & Ecker, 2017).

much emphasis on a single study. It is an open question if the observed pattern would emerge with different allegations, characters, instructions, retraction formats, or measures. Such task parameters should therefore be varied in conceptual replications. Second, our participants had no prior knowledge about the character of interest, and a sole negative behavior was provided in a context-free manner that did not relate to the person’s other behaviors—this will be different in many real-world situations. Future research could build a richer narrative, or use allegations against real-world people (e.g., Cone et al., 2019) to establish more conclusively whether the impact of false allegations on person impressions can be effectively corrected in real-life situations.

References

Allanson, F., & Ecker, U. K. H. (2017). No evidence for a role of reconsolidation in updating of paired associates. *Journal of Cognitive Psychology, 29*, 912-919. doi:10.1080/20445911.2017.1360307

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258-290. doi:10.1037/h0055756

Bjork, E. L., & Bjork, R. A. (1996). Continuing influences of to-be-forgotten information. *Consciousness & Cognition*, *5*, 176-196. doi:10.1006/ccog.1996.0011

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

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology*, *63*, 452-459. doi:10.1037//0022-3514.63.3.452

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

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*, 307-326. doi:10.1111/j.1467-9221.2012.00935.x

Cone, J., Flaharty, K., & Ferguson, M. J. (2019). Believability of evidence matters for correcting social impressions. *Proceedings of the National Academy of Sciences, 116*, 9802-9807. doi:10.1073/pnas.1903222116

Connor Desai, S., Pilditch, T., & Madsen, J. K. (2019). *The rational continued influence of misinformation.* doi:10.31234/osf.io/cqy6p

Cotton, J., O'Neill, B., & Griffin, A. (2008). The "name game": Affective and hiring reactions to first names. *Journal of Managerial Psychology, 23,* 18-39. doi:10.1108/02683940810849648

Dalal, D. K., Diab, D. L., & Tindale, R. S. (2015). I heard that…: Do rumors affect hiring decisions? *International Journal of Selection and Assessment, 23,* 224-236. doi:10.1111/ijsa.12110

Dreben, E. K., Fiske, S. T., & Hastie, R. (1979). The independence of evaluative and item information: Impression and recall order effects in behavior-based impression formation. *Journal of Personality and Social Psychology, 37,* 1758-1768. doi:10.1037/0022-3514.37.10.1758

Ecker, U. K. H., & Ang, L. C. (2019). Political attitudes and the processing of misinformation corrections. *Political Psychology, 40,* 241-260. doi:10.1111/pops.12494

Ecker, U. K. H., Lewandowsky, S., & Apai, J. (2011). Terrorists brought down the plane!—No, actually it was a technical fault: Processing corrections of emotive information. *Quarterly Journal of Experimental Psychology, 64,* 283-310. doi:10.1080/17470218.2010.497927

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,* 323-335. doi:10.1037/xap0000028

Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Misinformation in memory: Effects of the encoding strength and strength of retraction. *Psychonomic Bulletin & Review, 18,* 570–578. doi:10.3758/s13423-011-0065-1

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

Ecker, U. K. H., Oberauer, K., & Lewandowsky, S. (2014). Working memory updating involves item-specific removal. *Journal of Memory and Language*, *74*, 1-15. doi:10.1016/j.jml.2014.03.006

Farmer, R., & Sundberg, N. D. (1986). Boredom proneness: The development and correlates of a new scale. *Journal of Personality Assessment*, *50*, 4-17. doi:10.1207/s15327752jpa5001_2

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191. doi:10.3758/bf03193146

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, *23,* 1-74. doi:10.1016/s0065-2601(08)60317-2

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

Golding, J. M., Fowler, S. B., Long, D. L., & Latta, H. (1990). Instructions to disregard potentially useful information: The effects of pragmatics on evaluative judgments and recall. *Journal of Memory and Language*, *29,* 212-227. doi:10.1016/0749-596x(90)90073-9

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. doi:10.1016/j.neuropsychologia.2017.10.003

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

Gross, A. E., & Crofton, C. (1977). What is good is beautiful. *Sociometry, 40,* 85-90. doi:10.2307/3033549

Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review, 103*, 336-355. doi:10.1037//0033-295x.103.2.336

Hendrick, C., Franz, C. M., & Hoving, K. L. (1975). How do children form impressions of persons? They average. *Memory & Cognition, 3*, 325-328. doi:10.3758/bf03212919

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1-55. doi: 10.1016/0010-0285(92)90002-J

Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1420-1436. doi:10.1037//0278-7393.20.6.1420

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

Kassin, S. M., & Sommers, S. R. (1997). Inadmissible testimony, instructions to disregard, and the jury: Substantive versus procedural considerations. *Personality and Social Psychology Bulletin, 23*, 1046-1054. doi:10.1177/01461672972310005

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

Kessler, Y., & Meiran, N. (2008). Two dissociable updating processes in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1339-1348. doi:10.1037/a0013078

Krumpal, I. (2011). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*, 2025-2047. doi:10.1007/s11135-011-9640-9

Leirer, V. O., Hamilton, D. L., & Carpenter, S. (1982). Common first names as cues for inferences about personality. *Personality and Social Psychology Bulletin*, *8,* 712-718. doi:10.1177/0146167282084018

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

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago Face Database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*, 1122-1135. doi:10.3758/s13428-014-0532-5

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,* 823-849. doi:10.1037/pspa0000021

McCarthy, R. J., & Skowronski, J. J. (2011). What will Phil do next? The influence of spontaneous trait inferences on behavior predictions. *Journal of Experimental Social Psychology*, *47*, 321-332. doi:10.1016/j.jesp.2010.10.015

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

Mehrabian, A. (2001). Characteristics attributed to individuals on the basis of their first names. *Genetic, Social, and General Psychology Monographs, 127*, 59-88.

Mende-Siedlecki, P. (2018). Changing our minds: The neural bases of dynamic impression updating. *Current Opinion in Psychology, 24*, 72-76. doi:10.1016/j.copsyc.2018.08.007

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

Miron-Shatz, T., & Ben-Shakhar, G. (2008). Disregarding preliminary information when rating job applicants' performance: Mission impossible? *Journal of Applied Social Psychology*, *38*, 1271-1294. doi:10.1111/j.1559-1816.2008.00348.x

Morrow, D. G., Bower, G. H., & Greenspan, S. L. (1989). Updating situation models during narrative comprehension. *Journal of Memory and Language*, *28*, 292-312. doi:10.1016/0749-596x(89)90035-1

Newman, L. S. (1996). Trait impressions as heuristics for predicting future behavior. *Personality and Social Psychology Bulletin*, *22*, 395-411. doi:10.1177/0146167296224006

O'Rear, A. E., & Radvansky, G. A. (2020). Failure to accept retractions: A contribution to the continued influence effect. *Memory & Cognition*, *48*, 127-144. doi:10.3758/s13421-019-00967-9

Oberauer, K., & Vockenberg, K. (2009). Updating of working memory: Lingering bindings. *The Quarterly Journal of Experimental Psychology*, *62,* 967-987. doi:10.1080/17470210802372912

OECD (2018). Violence against women. https://data.oecd.org/inequality/violence-against-women.htm

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*, 11087-11092. doi:10.1073/pnas.0805664105

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

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (p. 49–69). New York: Lawrence Erlbaum Associates.

Paunonen, S. V. (2006). You are honest, therefore I like you and find you attractive. *Journal of Research in Personality*, *40*, 237-249. doi:10.1016/j.jrp.2004.12.003

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

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

Phillips, J. & Vandenbroek, P. (2014). Domestic, family and sexual violence in Australia: an overview of the issues. http://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/rp1415/ViolenceAust

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

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

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

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*, 62-74. doi:10.1037/xlm0000155

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32,* 880-892. doi:10.1037//0022-3514.32.5.880

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

Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? *Psychology of Learning and Motivation, 41*, 265-294. doi:10.1016/S0079-7421(02)80009-3

Sidhu, D. M., Deschamps, K., Bourdage, J. S., & Pexman, P. M. (2019). Does the name say it all? Investigating phoneme-personality sound symbolism in first names. *Journal of Experimental Psychology: General, 148*, 1595-1614. doi:10.1037/xge0000662

Sjovall, A. M., & Talk, A. C. (2004). From actions to impressions: Cognitive attribution theory and the formation of corporate reputation. *Corporate Reputation Review*, *7*, 269-281. doi:10.1057/palgrave.crr.1540225

Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review*, *96*, 58-83. doi: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*, 469-492. doi:10.1007/s10979-006-9039-7

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1948-1961. doi:10.1037/xlm0000422

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

Todorov, A., Said, C. C., & Verosky, S. C. (2011). Personality impressions from facial appearance. In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.), *The Oxford handbook of face perception* (pp. 631-651). Oxford: Oxford University Press doi:10.1093/oxfordhb/9780199559053.013.0032

Uleman, J. S., Newman, L. S., & Moskowitz, G. B. (1996). People as flexible interpreters: Evidence and issues from spontaneous trait inference. *Advances in Experimental Social Psychology*, *28*, 211-279. doi:10.1016/s0065-2601(08)60239-7

UN Women (2011). Violence against women prevalence data: Surveys by country. https://www.endvawnow.org/uploads/browser/files/vaw_prevalence_matrix_15april_2011.pdf

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

van Overwalle, F., & Labiouse, C. (2004). A recurrent connectionist model of person impression formation. *Personality and Social Psychology Review*, *8*, 28-61. doi:10.1207/S15327957PSPR0801_2

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*, 58-76. doi:10.3758/s13423-017-1323-7

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*, 155-177. doi:10.1177/0093650219854600

Wyer, N. A. (2010). You never get a second chance to make a first (implicit) impression: The role of elaboration in the formation and revision of implicit impressions. *Social Cognition*, *28*, 1-19. doi:10.1521/soco.2010.28.1.1

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

Author contributions: UE and AR jointly conceptualized the study, designed the experiment, analyzed the results, and wrote the manuscript.

Acknowledgements: This research was supported by the Australian Research Council under grant DP160103596, awarded to the first author. We thank Charles Hanich for research assistance.

Appendix A

Neutral Behaviors (note *C*, *N*, and *R* refer to control, negative, and retraction conditions)

1. Gave an engaging presentation to his class. *(presented in CNR)*
2. Had one too many drinks at his best friend’s birthday party. *(presented in CNR)*
3. Arrived 10 minutes early to work. *(presented in CNR)*
4. Caught up on the latest episodes of his favorite TV show. *(presented in C)*
5. Went to the local bar on a Friday night. *(presented/confirmed in CNR)*
6. Parked his car in a no-parking zone at the shops. *(presented/confirmed in CNR)*
7. Ate at an expensive Italian Restaurant. *(presented/retracted in CNR)*
8. Played a game of social tennis. *(presented in CNR; retracted in C and N)*
9. Told a funny story from his childhood. *(new information in CNR)*
10. Attended a football game. *(new information in CNR)*

Negative Behavior

1. Slapped his girlfriend during an argument. *(presented in N and R; retracted in R)*

Appendix B

Reysen Likability Scale (Reysen, 2005)

The Reysen Likability Scale was used to measure the likability of “John”. Participants responded to each item on a seven-point Likert scale from “very strongly disagree” to “very strongly agree”.

1. This person is friendly.
2. This person is likeable.
3. This person is warm.
4. This person is approachable.
5. I would ask this person for advice.
6. I would like this person as a co-worker.
7. I would like this person as a roommate.
8. I would like to be friends with this person.
9. This person is physically attractive.
10. This person is similar to me.
11. This person is knowledgeable.

Appendix C

Aggressiveness Scale

The aggressiveness scale was used to measure the perceived aggressiveness of "John". Participants responded to each item on a five-point Likert scale from "extremely uncharacteristic" to "extremely characteristic". Aggression-related items are marked with an *. Reverse-coded items are marked with (R). Non-aggression items were used as distractors.

1. He enjoys taking on new challenges.
2. Others would describe him as being hot-headed.*
3. When other people disagree with him, he will rarely challenge them.* (R)
4. He is always thinking about projects to complete.
5. He needs a lot of variety in his life to keep him satisfied.
6. Many people view him as an argumentative person.*
7. He often has a lot of free time available.
8. He is likely to threaten other people.*
9. It is easy for him to concentrate on activities.
10. He often struggles with controlling his temper.*
11. Most people would describe him as a very creative person.
12. When he is frustrated, he is unlikely to let his frustration show.* (R)
13. He rarely enjoys going to work.
14. If he feels that his rights are violated, he is likely to respond in a hostile manner.*
15. He may purposely tease others to make them angry.*
16. He finds it easy to entertain himself.
17. He is an even tempered person.* (R)
18. If people annoy him, he is likely to tell them what he thinks of them.*

Appendix D

Participants were presented with the following scenarios regarding “John” and were required to predict the likelihood of each behavior response on a scale from 0 (very unlikely) to 10 (very likely). Scenarios 1-5 include an aggressive response (marked with an *). Scenarios 6-9 were used as distractors. All scenarios were presented in a randomized order.

Scenario 1

While John is at the pub, he goes to the bar to order a drink. Suddenly a stranger spills some of his drink over him without noticing. How likely is it that John will…?

1. Alert the stranger to what they have done but then make a joke about of it
2. Push the stranger*
3. Ignore the incident

Scenario 2

John is driving home after a long shift at work. While he is on the freeway, another driver cuts in front of him and John is forced to brake. How likely is it that John will…?

1. Show a rude gesture to the other driver*
2. Briefly turn up his headlights to alert the other driver and then restore a safe distance
3. Change lanes to avoid the other driver

Scenario 3

Early one morning, John is waiting in line at a busy café to buy a cup of coffee. While he is waiting, another customer jumps the queue. How likely is it that John will…?

1. Ignore the other customer
2. Scold the other customer*
3. Politely tell the other customer to move to the back of the line

Scenario 4

John is sitting in his car in a carpark waiting for a friend. While he is waiting, another car parks in the empty bay next to him. As the other driver gets out, his car door accidently touches John's car. How likely is it that John will…?

1. Politely tell the other driver to be more careful next time
2. Ignore the incident
3. Swear at the other driver*

Scenario 5

John is at his local grocery store buying food for the week. While he is at the checkout, he notices that he is being incorrectly charged a small amount of extra money for an item. How likely is it that John will…?

1. Demand to see the manager to complain about the checkout operator*
2. Politely ask for a refund
3. Walk away without requesting a refund

Scenario 6

John had been invited to a short road trip with friends over the weekend. He did, however, have an exam the following Monday and had not yet prepared for it. How likely is it that John will…?

1. Go on the trip without studying for his exam
2. Skip the trip to study for his exam
3. Take his books to study while on the trip

Scenario 7

John has just finished his shift at work and is about to leave. A co-worker asks John is he could give him a lift. John knows that this will take him an extra 10 minutes to reach home. How likely is it that John will…?

1. Give his co-worker a lift home
2. Give his co-worker a lift to the nearest bus station
3. Make an excuse to avoid giving his co-worker a lift

Scenario 8

John is walking across campus to get to his next lecture. He is running late due to difficult finding parking. While he is walking, he recognizes someone who he has spoken to before, but does not know very well. How likely is it that John will…?

1. Avoid making eye-contact with the other person
2. Briefly greet the other person
3. Have a chat with the other person

Scenario 9

It is nearing dinner time and John is alone at home. He looks in the fridge and finds leftovers that do not look particularly appetizing. This is, however, meat and vegetable that have not yet been cooked. How likely is it that John will…?

1. Cook himself a fresh meal
2. Eat the leftovers
3. Order takeaway food

Appendix E

Recognition Test 1: Initial Behavior Questions

Participants were required to select the correct answer out of three choices.

1. What day did he go to the local bar?

   a. Friday

   b. Saturday

   c. Sunday

2. How early did he arrive to work?

   a. 5 minutes

   b. 10 minutes

   c. 15 minutes

3. What sport did he play?

   a. Soccer

   b. Golf

   c. Tennis

4. Which of the following did John do?

   a. Drive 15km/h over the speed limit

   b. Park his car in a no-parking zone at the shops

   c. Avoid paying a parking infringement

5. *In the control condition:* What did he watch on TV?

   a. Movies

   b. Favorite TV show

   c. News

   In the negative and retraction conditions: Who did he slap during an argument?

   a. His brother

b. His girlfriend

c. His friend

6. At which restaurant did he eat?

a. Greek

b. Italian

c. Chinese

7. What did he do in front of his class?

a. Give an engaging presentation

b. Make a distasteful joke

c. Talk about his project

8. What did he do at his best friend’s birthday party?

a. Give a funny speech

b. Make new friends

c. Have one too many drinks

Recognition Test 2: Updated Information Questions

Participants had to select one or more correct answer(s) from four choices.

Questions presented in the control condition:

1. Which of these behaviors were confirmed?

a. Caught up on the latest episodes of his favorite TV show

b. Went to the local bar on a Friday night

c. Arrived 10 minutes early to work

d. Parked his car in a no-parking zone at the shops

2. Which of these behaviors were retracted?

a. Gave an engaging presentation to his class

b. Parked his car in a no-parking zone at the shops

c. Ate at an expensive Italian restaurant

d. Played a game of social tennis

3. Which of these behaviors were added?

a. Went on a holiday

b. Attended a football game

c. Bought a new pair of shoes

d. Told a funny story from his childhood

Questions presented in the negative and retraction conditions:

1. Which of these behaviors were confirmed?

a. Played a game of social tennis

b. Went to the local bar on a Friday night

c. Arrived 10 minutes early to work

d. Parked his car in a no-parking zone at the shops

2. Which of these behaviors were retracted?

a. Gave an engaging presentation to his class

b. Parked his car in a no-parking zone at the shops

c. Ate at an expensive Italian restaurant

d. Slapped his girlfriend during an argument

3. Which of these behaviors were added?

a. Went on a holiday

b. Attended a football game

c. Bought a new pair of shoes

d. Told a funny story from his childhood

Appendix F

Verbatim Task Instructions

Phase 1. When you meet a person for the first time, you are usually given very little information about what that person is like. Physical cues such as appearance and body language are usually the first pieces of evidence that people use to form an impression of a stranger. This first impression influences how we view a person.

Alternatively, we may form an impression of a person based on information about their behaviour. This experiment is interested in **how people use behavioural information to form impressions of others**. You will be provided with **examples of behaviours** that a person has engaged in over the past month. Your task will be to **form an impression of the person** and then respond to some questions regarding the person.

Phase 2. The information we will present to you now was obtained with informed consent. In order to maintain confidentiality, the person of interest will be referred to as "John". We interviewed four acquaintances of John. They were asked to provide any examples of John's behaviour that they had observed over the past month.

You will be presented with some of these behaviours, one by one. **Think about the significance of each behaviour and how it shapes your impression of John**. Imagine that you are about to meet John for the first time at a social gathering. Think about what kind of person he is, and what it would be like to spend time with him.

Phase 3. After interviewing the four acquaintances separately, we spoke to them as a group, and also spoke to John himself, to verify the information from the initial interviews and to learn more about John. Therefore, you will now be presented with some additional information regarding John. **You should use this new information to update your impression of John.**

Phase 4. The following tasks require you to make **judgements** about John. Please provide an **honest opinion** based on your impression of him. There are **no right or wrong answers** and all information provided will remain **confidential.**

Reysen Likability Scale. Using your impression of John, please indicate how much you agree with each description.

Aggressiveness scale. Using your impression of John, please indicate how characteristic each description is of John.

Behavior-prediction task. You will now be presented with a series of scenarios involving John. After each scenario, you will be presented with three possible behaviour responses of how John may choose to react in the given situation. For each response, please predict how likely John would show the specified behaviour in the given situation, on a scale from 0 - 10.

Face-rating task. You will now be presented with an image of John, and you will be asked to rate his face on a number of dimensions from 0 - 10.

Recognition test. Try to remember the initial behaviour examples that were presented (irrespective of later corrections). / Try to remember any information that was updated. You may select more than one.