Illusory-Correlation Effects on Implicit and Explicit Evaluation

Pieter Van Dessel[1]

Kate Ratliff[2]

Skylar M. Brannon[3]

Bertram Gawronski[3]

Jan De Houwer[1]

[1]Department of Experimental-Clinical and Health Psychology, Ghent University, Belgium

[2]Department of Psychology, University of Florida, US

[3]Department of Psychology, University of Texas at Austin, USA

Author note: Correspondence concerning this article should be addressed to Pieter Van Dessel (Pieter.VanDessel@UGent.be) or Kate Ratliff (Ratliff@ufl.edu).

The research reported in this paper was supported by the Scientific Research Foundation, Flanders under Grant FWO16/PDO/201 to PVD, the Ghent University Methusalem Grant BOF16/MET_V/002 to JDH, and by Project Implicit.

Abstract

Research suggests that people sometimes perceive a relationship between stimuli when no such relationship exists (i.e., illusory correlation). Illusory-correlation effects are thought to play a central role in the formation of stereotypes and evaluations of minority versus majority groups, often leading to less favorable impressions of minorities. Extant theories differ in terms of whether they attribute illusory-correlation effects to processes operating during learning (belief formation) or measurement (belief expression), and whether different evaluation measures should be differentially sensitive to illusory-correlation effects. Past research found mixed evidence for dissociative effects of illusory-correlation manipulations on measures of implicit (i.e., automatic) and explicit (i.e., controlled) evaluation. Four high-powered studies obtained illusory-correlation effects on explicit evaluations, but not implicit evaluations probed with an Implicit Association Test, Evaluative Priming Task, and Affect Misattribution Procedure. The results are consistent with theories that attribute illusory-correlation effects to processes during belief expression.

Keywords: automatic evaluation, dissociation, illusory correlation, implicit measures

Illusory-Correlation Effects on Implicit and Explicit Evaluation

Research suggests that beliefs influence behavior, even when these beliefs are erroneous (e.g., in consumer research: Geraerts et al., 2008; in politics: Wells, Reedy, Gastil, & Lee, 2009). In social psychology, the effect of false beliefs on behavior is predominantly studied in research on social stereotypes (i.e., beliefs about the traits of members of social categories). It is widely assumed that stereotypical beliefs can lead people to act in a biased manner towards certain social groups (Wheeler & Petty, 2001). It is therefore important to know how these beliefs might arise and what underlies their impact.

Illusory-correlation effects

One type of beliefs that are thought to contribute to social stereotypes are illusory-correlation beliefs, which refer to beliefs about the contiguous relationship between two variables (e.g., the presence of social category members and the presence of valenced events) when no such relationship exists. In a foundational study by Hamilton and Gifford (1976), participants read a series of statements describing either desirable or undesirable behaviors performed by the members of two fictitious social groups. More statements were presented for one group (majority group) than for the other group (minority group) but, importantly, the overall proportion of positive to negative behavioral statements was the same for both groups. When more positive than negative behavioral statements were presented for both groups, an illusory-correlation effect was observed such that participants overestimated the proportion of negative statements about the minority compared to the majority group. Conversely, when more negative than positive behavioral statements were presented for both groups, participants overestimated the proportion of positive statements about the minority compared to the majority group. The authors argued that illusory-correlation effects might also arise in real life. Because negative behaviors tend to be

less frequent than positive behaviors in most real-world contexts, people may have a tendency to form negative stereotypes of minorities even when their behavior does not differ from the behavior of majorities. Later studies have established the robustness of illusory-correlation effects and found that effects on proportion estimates transfer to other stimulus-related behavior such as social group evaluation (for reviews, see Fiedler & Walther, 2004; Mullen & Johnson, 1990).

Explanations of illusory-correlation effects

Broadly, two different classes of explanations have been proposed for illusory-correlation effects. One type of explanations attributes effects of illusory-correlation manipulations to processes operating during learning (i.e., belief formation). For example, Hamilton and Gifford's (1976) distinctiveness account postulates that uncommon events are more salient than frequent events. Due to stronger memory encoding of salient, infrequent events (e.g., negative behavior of minority members when negative behavior is less frequent than positive behavior; Johnson & Mullen, 1994), people will overestimate the frequency of these events, which leads to illusory-correlation effects. Another explanation that refers to learning-related processes indicates that there are more learning trials for frequent behavior of the majority group (Fiedler, 1991). During learning, people might extract information more robustly from a larger sample of exemplars, such that frequent information about the majority group is more strongly represented in memory (there is less information loss). Participants might draw on this difference when asked to report proportion judgments or evaluations for the two groups.

A different perspective is provided by explanations of illusory-correlation effects that draw on processes during measurement (i.e., belief expression; e.g., Klauer & Meiser, 2000; Eder, Fiedler, & Hamm-Eder, 2011). When asked to report proportion judgments or evaluations

for two groups, people may have a tendency to meaningfully distinguish between the groups (Berndsen & Spears, 1997). However, due to limitations in memory, participants may not remember to which group some of the positive and negative behaviors referred, and therefore engage in guessing processes. Evidence suggests that, in a typical illusory-correlation paradigm, participants choose the majority group with a higher probability when guessing the origin of a positive behavior and with a smaller probability when guessing the origin of a negative behavior (Bulli & Primi, 2006). One reason for this might be that participants estimate information by drawing on fast and frugal heuristics. They may predominantly use the heuristic that things that occur less often (i.e., statements about minority group members and statements about negative behavior) belong together, leading to the typical error in proportion estimates and evaluations.

Effects of illusory-correlation manipulations on different behavioral measures

An important distinction between these two classes of theories lies in the prediction of differential effects of illusory-correlation manipulations on different measures of evaluation. Theories that attribute illusory-correlation effects to processes that occur during learning (e.g., Hamilton & Gifford, 1976) assume that participants incorrectly represent the proportion of valenced information. Because this representation should transfer to any behavior for which this proportion is relevant, these theories do not distinguish different ways in which illusory correlations impact different types of behavior. Once represented in memory, the proportion of valenced statements is transferred to behavior for which this information is relevant irrespective of how these types of behavior are measured.

In contrast, theories that attribute illusory-correlation effects to processes that occur during measurement assume that different behavioral measures can be differentially sensitive to illusory-correlation effects. For example, guessing bias theories (e.g., Klauer & Meiser, 2000)

predict that dissociations could be observed on measures that evoke different guessing processes or heuristics.

Ratliff and Nosek (2010) provided preliminary evidence that responses on different evaluation measures can be differentially affected by illusory correlations. In two experiments, participants read more positive than negative behavioral statements about members of two fictitious social groups (i.e., Niffians and Laapians). As in a typical illusory-correlation paradigm, there were more statements about the majority group compared to the minority group, but the proportion of positive to negative information was identical for the two groups. In line with prior research, participants reported a preference for the majority over the minority group on self-reported liking ratings. In contrast, no such preference was observed on an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998).

Ratliff and Nosek (2010) explained their findings in terms of dual-process theories of evaluation (Gawronski & Bodenhausen, 2006; Strack & Deutsch, 2004), postulating that the IAT registers implicit attitudes that reflect automatic activation of associations between representations in memory (e.g., between representations of a social group and positivity), whereas self-report measures register explicit attitudes that reflect belief-based processes. Illusory-correlation manipulations might require belief-based processes and therefore influence self-report measures of evaluation but not the IAT (Ratliff & Nosek, 2010). However, this interpretation is based on the assumption that dissociations between measures of evaluation can be interpreted as proxies for functionally distinct learning mechanisms and their resulting representations-an assumption that has been challenged by an accumulating body of evidence (see Corneille & Mertens, in press; De Houwer, 2014; Gawronski, De Houwer, & Sherman, in press; Kurdi & Dunham, in press; Van Dessel, Gawronski, & De Houwer, 2019).

In light of this evidence, the most parsimonious interpretation of Ratliff and Nosek's (2010) finding is that illusory correlation effects result from processes operating during measurement (belief expression) rather than learning (belief formation). For example, guessing-based theories predict dissociative effects on different measures of evaluation when these measures are differentially sensitive to guessing-related processes. Notably, guessing the source of behavioral information might require evaluation under optimal conditions such as sufficient time and attention or the intention to provide an adequate (evaluative) response. For instance, evaluation measures for which guessing the differential proportion of valenced information is relevant should show stronger illusory-correlation effects. In contrast, measures in which evaluative behavior occurs under some of the conditions of automaticity (i.e., automatic or implicit evaluation measures: see De Houwer et al., 2013) might show little evidence for illusory-correlation effects. The IAT is a measure in which evaluation is automatic in the sense that evaluative responding is faster and less controlled than in a self-report measure (De Houwer et al., 2009). Hence, guessing-based theories would predict only weak or absent illusory-correlation effects on the IAT.

However, findings by Carraro and colleagues (2014) give reason to be cautious about the conclusion that illusory correlation effects result from measurement-related rather than learning-related processes. Different from Ratliff and Nosek (2010), Cararro et al. obtained strong illusory correlation effects on the IAT. In this study, participants were shown 39 sentences each describing a behavior performed by a member of one of two groups (Group A and B). Though the ratio of positive to negative behaviors was identical for both groups, IAT scores reflected a preference for Group A, the group for which participants had seen twice the number of sentences.

The discrepancy between the findings of Carraro et al. (2014) and Ratliff and Nosek (2010) is important, because it obscures whether illusory-correlation manipulations affect IAT

scores, and thus whether illusory-correlation effects arise from learning-related or measurement-related processes. Because both studies included a confound that is unrelated to illusory correlations, this question is even more difficult to answer. In both studies, participants saw positive statements more frequently than negative behavior, but there was no control condition in which negative statements were presented more frequently than positive statements.[1] Because illusory-correlation effects should lead to a reversed preference for the minority over the majority group, it cannot be ruled out that the preference reported for the majority group reflects a mere-exposure effect (i.e., a preference for stimuli that are presented more frequently; Zajonc, 1968).

The current study

The aim of the current study is to examine effects of illusory-correlation manipulations on implicit evaluation measures controlling for differences in mere exposure. We performed four high-powered pre-registered experiments. Experiments 1 and 2 provide conceptual replications of the study by Ratliff and Nosek (2010), additionally allowing independent tests of mere-exposure and illusory-correlation effects. Participants first read 36 statements about the valenced behaviors of two fictitious social groups (Niffites and Laapians), with twice the number of statements for one group (majority) compared to the other group (minority). Different from Ratliff and Nosek, we manipulated between subjects whether positive or negative behavioral statements were more frequent overall. In this design, a general preference for the majority group would reflect a mere-exposure effect, but not an illusory-correlation effect. The latter would be reflected in a preference for the majority group when positive statements are more frequent overall, and a

[1] Note that Ratliff and Nosek's (2010) study included two different types of control condition, one in which there was a higher ratio of positive to negative statements about the minority group and one in which there was a higher ratio of positive to negative statements about the majority group (but the statements were predominantly positive).

preference for the minority group when negative statements are more frequent overall (see Hamilton & Gifford, 1976).

Because (in)sensitivity of the IAT to illusory correlation effects could reflect non-evaluative processes, Experiments 3 and 4 aimed to replicate the findings of Experiments 1 and 2 with two other measures of implicit evaluation: The Affect Misattribution Procedure (AMP, Payne, Cheng, Govorun, & Stewart, 2005) and the Evaluative Priming Task (EPT, Fazio, Sanbonmatsu, Powell, & Kardes, 1986). In the IAT, participants perform two binary categorizations (e.g., the categorization of valenced words as positive or negative and the categorization of the evaluation stimuli on the basis of their identity) and evaluation is inferred on the basis of differential performance when categorizations are performed using the same versus a different response key. This feature makes the IAT sensitive to factors unrelated to evaluation, such as asymmetries in the salience of stimuli (Rothermund & Wentura, 2004) or extra-personal knowledge (Olson & Fazio, 2004), calling for replications with measures that do not suffer from this limitation.

The AMP and the EPT are widely used measures in which evaluative responding is also thought to occur under suboptimal conditions, that is under some of the conditions of automaticity (e.g., evaluative responding when there is little time to process stimuli or little opportunity or motivation to control responding; see De Houwer et al., 2009). In contrast to the IAT, however, evaluation is inferred on the basis of the effect of primes that precede the presentation of the target stimuli (i.e., valenced words or Chinese ideographs) on evaluation of these target stimuli. Furthermore, AMP scores are calculated on the basis of the number of positive and negative categorization responses rather than response latencies. These procedural differences might prevent the observation of variance in implicit evaluation scores that is due to non-evaluative processes. Moreover, by including measures other than the IAT, we can verify

whether the obtained results are specific to this one measure or whether they generalize to other measures that capture evaluative responding under suboptimal conditions.

Experiments 1 and 2

Method

Participants. A total of 3557 (Experiment 1) and 3825 (Experiment 2) English-speaking volunteers were recruited to participate online via the Project Implicit research website (https://implicit.harvard.edu). There were 289 (8.1%) and 450 participants (11.8%) who decided not to complete the experiment after receiving information about the duration and the nature of the study. A total of 1343 (37.8%) and 1569 participants (41.0%) dropped out during the experiment, leaving 1925 participants in Experiment 1 and 1806 participants in Experiment 2. The dropout rates were comparable across the experimental conditions, $\chi^2(3)$s < 2.25, *ps* > .52. Hence, there was no evidence for condition-dependent attrition. A target sample size of 1800 completed contributions was determined based on an a priori power analysis such that we would have sufficient power (i.e., power > 0.95) to detect a small between-subjects effect (d = 0.20) with an alpha criterion p < .05 in a two-tailed between-subjects *t*-test. Prior to data-collection, the target sample size was pre-registered together with the study design and data-analytic plans. Experiment 1 did not follow our planned study design, in that the order of the IAT and the self-report rating task was not counterbalanced due to a programming error. We therefore performed a second experiment that did fully implement our intended design plan. The preregistered plans, raw data, experimental and analytic scripts of these and all other experiments are available at https://osf.io/ry9v3/?view_only=e6e26932b7bc47b98e99bd03603efb65.

Following our preregistered data analysis plan, we excluded the data from participants who (a) did not fully complete all questions and tasks (Experiment 1: 90 participants; 4.9%;

Experiment 2: 76 participants; 4.2%), (b) had error rates above 30% when considering all IAT blocks or above 40% for any one of the critical IAT test blocks (Experiment 1: 285 participants; 14.8%; Experiment 2: 288 participants; 15.9%), or (c) responded faster than 400ms on more than 10% of the IAT trials (Experiment 1: 15 participants; 0.8%; Experiment 2: 10 participants; 0.6%).

Analyses were performed on the data of 1535 participants in Experiment 1 (950 women, mean age = 35, *SD* = 14) and 1432 participants in Experiment 2 (934 women, mean age = 35, *SD* = 15). Table 1 provides an overview of the number of participants in the different between-subject conditions.

Materials and Procedure. Participants first provided informed consent. In line with recommendations by Zhou and Fishbach (2016) to prevent selective attrition, participants were first (1) informed about the duration of the experiment and (2) asked to their best to complete all tasks in a thoughtful manner to help facilitate scientific advance.

Impression formation task. Participants were given the following instructions:

> *The purpose of this experiment is to find out how people process and retain information visually. The information you will read in the next part of the study consists of behaviors performed by members of two social groups - NIFFIANS and LAAPIANS. The groups described here are real groups that exist in society, but we are calling them NIFFIANS and LAAPIANS. When you are reading, try to form an impression of the two groups. Try to remember as much as you can because you will be asked about it later, but do not be discouraged if this seems difficult. Just do your best. Each sentence will appear on the screen for several seconds before automatically moving onto the next. Press the SPACEBAR when you are ready to begin.*

During the impression formation task, positive and negative statements about four Niffians and four Laapians were presented in random order. An example of a negative statement about a Niffian is ‘Ibonnif, a Niffian, did not offer his guests anything to drink”. An example of a positive statement about a Laapian is ‘Zinaalap, a Laapian, sent his mother flowers for Mother’s Day”. The positive and negative behavioral statements were the same as in Ratliff and Nosek (2010, Experiment 2). Because our design included more negative statements than Ratliff and Nosek used, 8 negative items were taken from other research on attitude formation and change (Rydell & Gawronski, 2009).

Each statement was presented for four seconds with an inter-trial interval of 1 second. In total, there were 24 statements about the majority group (6 about each group member) and 12 statements about the minority group (3 about each group member). For half the participants, 16 majority group statements were positive and 8 were negative whereas 8 minority group statements were positive and 4 were negative (Positive frequent condition). For the other participants, 16 majority group statements were negative and 8 were positive whereas 8 minority group statements were negative and 4 were positive (Negative frequent condition). It was counterbalanced between participants whether Niffians or Laapians were the majority group.

IAT. During the IAT, participants categorized eight attribute words (i.e., antisocial, social, irresponsible, responsible, unlikeable, likeable, unpleasant, pleasant, unfriendly; friendly, popular, unpopular) as ‘positive’ or ‘negative’ and the names of the four Niffians and Laapians as ‘Niffians’ or ‘Laapians’. Participants began the IAT with 20 practice trials sorting the Niffians and Laapians names and 20 practice trials sorting the positive and negative words with left and right key presses (keys E and I). Next, participants completed one block of 20 and one block of 40 trials in which Niffians and positive stimuli shared one response key and Laapians and negative stimuli shared another response key (or vice versa). Participants then practiced sorting

Niffians and Laapians on 40 trials with a reversed response key assignment. Finally, participants completed one block of 20 and one block of 40 trials in which Niffians shared a response key with negative and Lappians shared a response key with positive (or vice versa). If participants made an error in the categorization task, a red "X" appeared on the screen until participants provided the correct response. Latencies were recorded until a correct response was made.

IAT scores were calculated using the D2-algorithm (Greenwald, Nosek, & Banaji, 2003), such that higher scores indicate a stronger preference for Niffians over Laapians. The Spearman-Brown corrected split-half reliability of the IAT score, calculated on the basis of an odd-even split, was $r(1533) = .78$ in Experiment 1 and $r(1430) = .77$ in Experiment 2. Across groups, participants displayed a preference for Laapians over Niffians (Experiment 1: $M = -0.11$, $SD = 0.42$, $t[1534] = -9.96$, $p < .001$, $d = 0.25$; Experiment 2: $M = -0.11$, $SD = 0.43$, $t[1431] = -9.58$, $p < .001$, $d = 0.25$).

Trait rating task. Explicit evaluations were measured with eight-point semantic differential ratings of both Niffians and Laapians on six traits each. The rated traits were identical to the attribute words in the IAT: antisocial-social, irresponsible-responsible, unlikeable- likeable, unpleasant-pleasant, unfriendly-friendly, and popular- unpopular. The order of IAT and trait rating task was counterbalanced across participants in Experiment 2 but not Experiment 1.

We computed separate trait rating scores for Niffians and Laapians by averaging responses on the trait rating questions for each of the groups. An explicit evaluation score was computed by subtracting the resulting trait rating score for Laapians from the score for Niffians such that a positive value indicates a relative preference for Niffians over Laapians. Internal consistency of the self-reported evaluation score was high (Experiment 1: Cronbach's Alpha = .87; Experiment 2: Cronbach's Alpha = .79), and this score correlated significantly with the IAT

score (Experiment 1: r[1533] = .23, p < .001; Experiment 2: r[1430] = .19, p < .001). Across groups, participants displayed a small but robust preference for Laapians over Niffians (Experiment 1: M = -0.14, SD = 1.41, t[1534] = -3.97, p < .001, d = 0.10; Experiment 2: M = -0.10, SD = 1.15, t[1431] = -3.17, p = .002, d = 0.08).

Proportion test. Participants were informed about the number of statements they read about Niffians and Laapians at the beginning of the experiment and they were asked to indicate at the end how many of those they thought were positive for each group. Subsequently, participants were debriefed and thanked for their participation.

A proportion positive score was calculated by subtracting the proportion of statements that participants indicated were positive for Laapians from the proportion of statements that participants indicated were positive for Niffians. Overall, participants indicated a higher proportion of positive statements for Laapians than for Niffians but this was a small effect (Experiment 1: M = -2%, SD = 23%, t[1534] = 3.53, p < .001, d = 0.09; Experiment 2: M = -1%, SD = 22%, t[1431] = -2.00, p = .046, d = 0.05). This proportion positive score showed a small but significant correlation with IAT scores (Experiment 1: r[1533] = .08, p = .002; Experiment 2: r[1430] = .07, p = .011) and with self-reported evaluation scores (Experiment 1: r[1533] = .37, p < .001; Experiment 2: r[1430] = .34, p < .001).

Results

IAT scores.

Experiment 1. We performed a 2 (Majority Group: Niffians, Laapians) x 2 (Valence Frequency: Positive frequent, Negative frequent) x 2 (IAT Order: Block with Niffians and positive together first/ Block with Laapians and positive together first) analysis of variance (ANOVA) on the IAT scores of Experiment 1.

The ANOVA revealed a main effect of Majority Group, $F(1,1527) = 35.97$, $p < .001$, indicating that IAT scores were higher when Niffians were the majority group ($M = -0.04$, $SD = 0.42$) than when Laapians were the majority group ($M = -0.17$, $SD = 0.41$), $d = 0.30$. A Bayes factor (BF_1) was calculated to evaluate how strongly the data support either the null or the alternative hypothesis (with $BF_1 < 0$ reflecting stronger evidence for the absence of an effect and $BF_1 > 0$ reflecting stronger evidence for the presence of an effect) with Cauchy prior width = 0.20 (expected small effect). The Bayes factor indicates strong evidence for the presence of the main effect of Majority Group, $BF_1 > 1000$. We also observed an effect of IAT Order, $F(1,1527) = 44.37$, $p < .001$, indicating that IAT scores were higher when participants first categorized Laapians and positive together in the IAT ($M = -0.04$, $SD = 0.42$) than when they first categorized Niffians and positive together ($M = -0.17$, $SD = 0.41$), $d = 0.34$, $BF_1 > 1000$. We did not observe the crucial interaction of Majority Group and Valence Frequency, $F(1,1527) = 0.26$, $p = .61$, $BF_1 = 0.08$ (Table 2), nor any other significant two-way interaction effects, Fs < 3.18, ps > .074, BF_1s < 0.45. However, we did observe a significant but small three-way interaction effect, $F(1,1527) = 6.49$, $p = .011$, $BF_1 = 3.13$.

For participants who first categorized Niffians and positive together in the IAT, we observed the interaction of Majority Group and Valence Frequency, $F(1,750) = 4.61$, $p = .032$, $BF_1 = 0.98$, indicating that for the positive frequent group, IAT scores were higher when Niffians were the majority group ($M = -0.09$, $SD = 0.42$) than when Laapians were the majority group ($M = -0.25$, $SD = 0.40$), $t(372) = 3.55$, $p < .001$, 95% $CI_{diff} = [0.07, 0.24]$, $d = 0.37$, $BF_1 = 53.33$. We did not observe a Majority Group effect for the negative frequent group (Niffians Majority: $M = -0.17$, $SD = 0.40$, Laapians Majority: $M = -0.20$, $SD = 0.40$), $t(378) = 0.58$, $p = .56$, 95% $CI_{diff} = [-0.06, 0.10]$, $d = 0.06$, $BF_1 = 0.39$. For participants who first categorized Laapians and positive

together in the IAT, we did not observe a significant interaction effect of Majority Group and Valence Frequency, $F(1,777) = 2.10$, $p = .15$, $BF_1 = 0.30$.

Experiment 2. For Experiment 2, the ANOVA on IAT scores also included the factor Task Order (IAT first/self-report rating task first). We observed a main effect of Majority Group, $F(1,1416) = 42.01$, $p < .001$, indicating that IAT scores were higher when Niffians were the majority group ($M = -0.04$, $SD = 0.43$) than when Laapians were the majority group ($M = -0.18$, $SD = 0.41$), $d = 0.34$, $BF_1 > 1000$. We also observed a small but significant main effect of IAT Order, $F(1,1416) = 11.46$, $p < .001$, indicating higher IAT scores when participants first categorized Laapians and positive together in the IAT ($M = -0.07$, $SD = 0.43$) than when they first categorized Niffians and positive together ($M = -0.15$, $SD = 0.42$), $d = 0.17$, $BF_1 = 19.61$. We did not observe the crucial interaction of Majority Group and Valence Frequency, $F(1,1416) = 1.13$, $p = .29$, $BF_1 = 0.14$, but we did observe a significant (but small) three-way interaction effect of Majority Group, Valence Frequency, and Task Order, $F(1,1416) = 6.57$, $p = .010$, $BF_1 = 3.65$. For participants who first completed the IAT, we did not observe a significant interaction effect of Majority Group and Valence Frequency, $F(1,729) = 1.13$, $p = .29$, $BF_1 = 0.22$. In contrast, we did observe a significant interaction of Majority Group and Valence Frequency for participants who first completed the self-report rating task, $F(1,687) = 6.54$, $p = .011$, $BF_1 = 3.26$. For the positive frequent group, IAT scores were higher when Niffians were the majority group ($M = -0.04$, $SD = 0.41$) than when Laapians were the majority group ($M = -0.22$, $SD = 0.41$), $t(348) = 4.08$, $p < .001$, 95% $CI_{diff} = [0.09, 0.27]$, $d = 0.44$, $BF_1 = 288.30$, whereas we did not observe a Majority Group effect for the negative frequent group (Niffians Majority: $M = -0.09$, $SD = 0.42$, Laapians Majority: $M = -0.12$, $SD = 0.42$), $t(343) = 0.49$, $p = .63$, 95% $CI_{diff} = [-0.07, 0.11]$, $d = 0.05$, $BF_1 = 0.39$. The ANOVA did not reveal any other significant main or interaction effects, Fs < 3.03, ps > .082, BF_1s < 0.12.

Self-reported evaluation scores.

Experiment 1. The 2 (Majority Group) x 2 (Valence Frequency) x 2 (IAT Order) ANOVA on the self-reported evaluation scores of Experiment 1 did not show a main effect of Majority Group nor any other main or interaction effects, Fs < 2.60, ps > .10, BF_1s < 0.20, with the exception of a significant interaction between Majority Group and Valence Frequency, $F(1,1527) = 14.31$, $p < .001$, $BF_1 = 97.44$. For the positive frequent group, self-reported evaluation scores were higher when Niffians were the majority group ($M = 0.02$, $SD = 1.38$) than when Laapians were the majority group ($M = -0.37$, $SD = 1.40$), $t(755) = 3.86$, $p < .001$, 95% $CI_{diff} = [0.19, 0.59]$, $d = 0.28$, $BF_1 = 170.54$, but not for the negative frequent group (Niffians Majority: $M = -0.19$, $SD = 1.38$, Laapians Majority: $M = -0.03$, $SD = 1.40$), $t(776) = -1.54$, $p = .12$, 95% $CI_{diff} = [-0.35, 0.04]$, $d = 0.11$, $BF_1 = 0.70$.

Experiment 2. The 2 (Majority Group) x 2 (Valence Frequency) x 2 (IAT Order) x 2 (Task order) ANOVA on the self-reported evaluation scores of Experiment 2 revealed a small but significant main effect of IAT Order, $F(1,1416) = 4.14$, $p = .042$, indicating higher self-reported evaluation scores when participants first categorized Niffians and positive together in the IAT ($M = -0.02$, $SD = 1.22$) than when they first categorized Laapians and positive together ($M = -0.16$, $SD = 1.08$), $d = 0.12$, $BF_1 = 1.89$. Crucially, we also observed the interaction between Majority Group and Valence Frequency, $F(1,1416) = 13.51$, $p < .001$, $BF_1 = 54.97$. For the positive frequent group, self-reported evaluation scores were higher when Niffians were the majority group ($M = 0.03$, $SD = 1.11$) than when Laapians were the majority group ($M = -0.19$, $SD = 1.15$), $t(721) = 2.69$, $p = .007$, 95% $CI_{diff} = [0.06, 0.39]$, $d = 0.20$, $BF_1 = 11.50$, whereas for the negative frequent group self-reported evaluation scores were lower when Niffians were the majority group ($M = -0.21$, $SD = 1.17$) than when Laapians were the majority group ($M = 0.00$, $SD = 1.15$), $t(707) = -2.44$, $p = .015$, 95% $CI_{diff} = [-0.21, -0.01]$, $d = 0.18$, $BF_1 = 2.32$. We also

observed a small interaction effect of Majority Group, Valence, and Task Order, $F(1,1416) = 6.03$, $p = .014$, $BF_1 = 1.85$, indicating that the interaction effect of Majority Group and Valence Frequency was significant for participants who first completed the self-report rating task, $F(1,687) = 21.41$, $p < .001$, $BF_1 > 1000$, but not for participants who first completed the IAT, $F(1,729) = 0.67$, $p = .41$, $BF_1 = 0.16$. The ANOVA did not reveal any other significant effects, Fs < 3.70, ps $> .054$, BF_1s < 0.11.

Combined analyses of IAT and self-reported evaluation scores.

Experiment 1. To directly compare effects on IAT and self-reported evaluation scores, we performed an ANOVA on standardized scores that included Majority Group, Valence Frequency, and IAT Order as between-subject factors and Measure (IAT, self-reported evaluation score) as within-subjects factor. We observed a main effect of IAT Order, $F(1, 1526) = 22.55$, $p < .001$, $BF_1 > 1000$, which was qualified by an interaction of IAT Order and Measure, $F(1, 1526) = 21.26$, $p < .001$, $BF_1 > 1000$. More importantly, we also observed a main effect of Majority Group, $F(1, 1526) = 23.29$, $p < .001$, $BF_1 > 1000$, an interaction of Majority Group and Valence Frequency, $F(1, 1526) = 7.64$, $p = .006$, $BF_1 = 7.50$, a three-way interaction of Majority Group, Valence Frequency, and Measure, $F(1, 1526) = 7.13$, $p = .008$, $BF_1 = 8.09$, and a four-way interaction of Majority Group, Valence Frequency, Measure, and IAT Order, $F(1, 1526) = 7.52$, $p = .006$, $BF_1 = 4.12$. The latter interaction revealed that the Majority Group x Valence Frequency x Measure interaction was present for participants who first categorized Laapians and positive together in the IAT, $F(1, 777) = 14.80$, $p < .001$, $BF_1 = 983.10$, but not for participants who first categorized Niffians and positive together in the IAT, $F(1, 750) = 0.00$, $p = .96$, $BF_1 < 0.01$.

Experiment 2. The Majority Group x Valence Frequency x IAT Order x Task Order x Measure ANOVA revealed an interaction of IAT Order and Measure, $F(1, 1416) = 18.11$, $p <$

.001, $BF_1 > 1000$. More importantly, we also observed a main effect of Majority Group, $F(1, 1416) = 17.59$, $p < .001$, $BF_1 > 1000$, an interaction of Majority Group and Valence Frequency, $F(1, 1526) = 9.49$, $p = .002$, $BF_1 = 5.11$, an interaction of Majority Group and Measure, $F(1, 1526) = 25.41$, $p < .001$, $BF_1 > 1000$, a three-way interaction of Majority Group, Valence Frequency, and Task Order, $F(1, 1526) = 10.56$, $p = .001$, $BF_1 = 9.21$, and of Majority Group, Valence Frequency, and Measure, $F(1, 1526) = 4.31$, $p = .040$, $BF_1 = 1.59$, and finally, a five-way interaction of Majority Group, Valence Frequency, Measure, Task Order, and IAT Order, $F(1, 1526) = 5.62$, $p = .020$, $BF_1 = 2.19$. The latter interaction revealed that the crucial Majority Group x Valence Frequency x Measure interaction was present for (1) participants who first categorized Laapians and positive together in the IAT and who started with the IAT, $F(1, 381) = 4.91$, $p = .030$, $BF_1 = 2.98$, and (2) participants who first categorized Niffians and positive together in the IAT and who started with the trait rating task, $F(1, 325) = 4.67$, $p = .032$, $BF_1 = 3.06$, but not for participants who first categorized Niffians and positive together in the IAT and who started with the IAT, $F(1, 348) = 0.00$, $p = .96$, $BF_1 < 0.01$, or for participants who first categorized Laapians and positive together in the IAT and who started with the trait rating task, $F(1, 362) = 0.07$, $p = .79$, $BF_1 = 0.02$.

Proportion positive scores.

Experiment 1. A 2 x 2 x 2 ANOVA in Experiment 1 revealed a main effect of Majority Group, $F(1, 1527) = 165.33$, $p < .001$, indicating that participants reported a lower proportion of positive statements for Niffians when Niffians were the majority group (M = -9%, SD = 22%) than when Laapians were the majority group (M = 5%, SD = 22%), $d = 0.66$, $BF_1 > 1000$. We also observed an interaction of Majority Group and Valence Frequency, $F(1, 1527) = 15.83$, $p < .001$, $BF_1 > 1000$, indicating that the main effect of Majority Group was reduced when the majority of the statements were positive, $t(755) = -6.04$, $p < .001$, 95% $CI_{diff} = [-0.13, -0.07]$, $d = 0.44$, $BF_1 >$

1000, compared to when the majority of the statements were negative, $t(776) = -12.41$, $p < .001$, 95% $CI_{diff} = [-0.22, -0.16]$, $d = 0.89$, $BF_1 > 1000$.

Experiment 2. A 2 x 2 x 2 x 2 ANOVA in Experiment 2 also revealed a main effect of Majority Group, $F(1, 1416) = 108.67$, $p < .001$, indicating that participants reported a lower proportion of positive statements for Niffians when Niffians were the majority group ($M = -7\%$, $SD = 21\%$) than when Laapians were the majority group ($M = 5\%$, $SD = 21\%$), $d = 0.55$, $BF_1 > 1000$. We also observed an interaction of Majority Group and Valence Frequency, $F(1, 1416) = 37.91$, $p < .001$, $BF_1 > 1000$, indicating that the main effect of Majority Group was reduced when the majority of the statements were positive, $t(721) = -3.72$, $p < .001$, 95% $CI_{diff} = [-0.09, -0.03]$, $d = 0.28$, $BF_1 = 104.69$, compared to when the majority of the statements were negative, $t(707) = -10.83$, $p < .001$, 95% $CI_{diff} = [-0.21, -0.14]$, $d = 0.81$, $BF_1 > 1000$. We also observed an interaction of Valence, IAT Order, and Task Order, $F(1, 1416) = 5.43$, $p = .020$, $BF_1 = 1.35$.

Discussion

Experiments 1 and 2 showed the expected illusory-correlation effect on explicit evaluations: participants preferred the group for which they learned more information (the majority group) when the majority of the statements were positive but not when the majority of the statements were negative. Importantly, this illusory-correlation effect was not observed on IAT scores which only revealed a general preference for the group participants learned more information about, irrespective of valence frequency (i.e., a mere-exposure effect). Unexpectedly, we also did not observe a typical illusory-correlation effect on estimated proportions of positive information. There was a difference in proportion estimates depending on the valence of the majority of statements, but both groups reported a lower proportion of positive statements for the majority group. One possible explanation is that the proportion measure asked participants to indicate the total number of pieces of positive information about each group. However, the total

number of pieces of information was different for both groups and, though the instructions emphasized this difference, some participants might not have taken this into account when providing their answers.

Experiment 3

Experiments 1 and 2 failed to provide support for illusory-correlation effects on implicit evaluations. It is possible, however, that specific characteristics of the implicit evaluation measure (i.e., the IAT) preclude observation of such an effect. In Experiments 3 and 4, we therefore extend our investigation to two other popular measures of implicit evaluation. The design of Experiment 3 was similar to that of Experiments 1 and 2, with two exceptions. First, the implicit evaluation measure consisted of an Affect Misattribution Procedure (AMP: Payne et al., 2005). In this task, participants evaluate Chinese ideographs that are preceded by brief presentations of a prime (i.e., the name of Niffians or Laapians) that participants are instructed to ignore. Implicit evaluation scores were computed based on the difference in the proportion of positive responses in the context of the different types of primes. Second, we used a different measure for the estimated proportion of positive information in which participants used a slider to indicate the percentage of positive to negative pieces of information for both groups.

Method

Participants. A total of 3334 English-speaking volunteers were recruited to participate online via the Project Implicit research website (https://implicit.harvard.edu). There were 307 (9.2%) participants who decided not to complete the experiment after receiving information about the duration and the nature of the study. A total of 1198 (35.9%) participants dropped out during the experiment, leaving 1829 participants. The dropout rates were not significantly different

across the experimental conditions, $\chi^2(3)$s < 5.10, *ps* > .18. Hence, there was no evidence for condition-dependent attrition. Sampling plan was the same as for Experiments 1 and 2.

Data were excluded for participants who (a) did not fully complete all questions and tasks (134 participants; 7.3%), or (b) showed the same response on all trials in the AMP blocks (300 participants; 14.8%). Analyses were performed on the data of 1395 participants (846 women, mean age = 32, *SD* = 14). Table 3 provides an overview of the number of participants in the different between-subject conditions.

Procedure. Procedure was identical to Experiments 1 and 2, with three exceptions. First, behavioral statements were now presented for 5000ms. This change was made because some participants in Experiments 1 and 2 indicated that they were unable to read the entire statement for all statements and we wanted to ensure this.

Second, participants completed an AMP instead of an IAT. The AMP procedure followed Cone and Ferguson (2015) and consisted of one practice block consisting of three trials that used the word table or chair as primes and three critical blocks of 40 trials that included the names of the Niffians and Laapians as primes. On each trial, participants were presented with a prime stimulus for 75ms, a blank screen for 125ms, and a Chinese ideograph for 100ms, which was then covered with a black-and-white pattern mask. Participants were asked to indicate if they considered the Chinese ideograph more or less visually pleasant than average by pressing either "E" or "I", respectively.

AMP scores were calculated by subtracting the proportion of pleasant responses on trials with Laapian primes from the proportion of pleasant responses on trials with Niffian primes, such that higher scores indicate a stronger preference for Niffians over Laapians. The Spearman-Brown corrected split-half reliability of the AMP score was low, $r(1261) = .18$. Across groups,

participants displayed a small preference for Laapians over Niffians (M = -0.01, SD = 0.10), $t(1261) = -2.44$, $p = .015$, $d = 0.07$. The AMP score showed a small but significant correlation with the self-reported evaluation scores (internal consistency: Cronbach's Alpha = .78), $r(1260) = .06$, $p = .028$, which also revealed a small preference for Laapians over Niffians (M = -0.10, SD = 1.24), $t(1534) = -2.74$, $p = .006$, $d = 0.08$.

A third change to the procedure was in the proportion test. Participants were now asked to indicate the proportion (rather than the total number) of statements about Niffians and Laapians that they thought were positive. Responses were given on a slider scale ranging from 0% to 100%. A proportion positive score was calculated by subtracting the proportion of statements that participants indicated were positive for Laapians from the proportion of statements that participants indicated were positive for Niffians. Overall, participants indicated a slightly higher proportion of positive statements for Laapians than for Niffians (M = -1%, SD = 20%), $t(1261) = -2.00$, $p = .047$, $d = 0.06$. This proportion positive score correlated significantly with AMP scores, $r(1260) = .10$, $p < .001$, and with self-reported evaluation scores, $r(1260) = .64$, $p < .001$.

Results

AMP scores. We performed a 2 (Majority Group) x 2 (Valence Frequency) x 2 (Task Order) ANOVA on AMP scores. The ANOVA revealed a small but significant main effect of Majority Group, $F(1,1254) = 4.19$, $p = .041$, indicating that AMP scores were lower when Niffians were the majority group (M = -0.01, SD = 0.10) than when Laapians were the majority group (M = 0.00, SD = 0.10), $d = 0.11$, $BF_1 = 1.24$. We also observed a small effect of Valence Frequency, $F(1,1254) = 6.95$, $p = .008$, indicating that AMP scores were lower when the majority of information was positive (M = -0.01, SD = 0.10) than when the majority of information was negative (M = 0.00, SD = 0.09), $d = 0.14$, $BF_1 = 2.88$. We did not observe the crucial interaction

of Majority Group and Valence Frequency, $F(1,1254) = 0.72$, $p = .40$, $BF_1 = 0.13$ (Table 4), nor any other significant interaction effects, Fs < 1.69, ps > .19, BF_1s < 0.23.

Self-reported evaluation scores. The 2 (Majority Group) x 2 (Valence Frequency) x 2 (Task Order) ANOVA on the self-reported evaluation scores did not show a main effect of Majority Group nor any other main or interaction effects, Fs < 2.79, ps > .095, BF_1s < 0.54, with the exception of a significant interaction between Majority Group and Valence Frequency, $F(1,1254) = 18.82$, $p < .001$, $BF_1 = 398.56$. For the positive frequent group, self-reported evaluation scores were higher when Niffians were the majority group ($M = 0.04$, $SD = 1.22$) than when Laapians were the majority group ($M = -0.23$, $SD = 1.13$), $t(633) = 2.87$, $p = .004$, 95% $CI_{diff} = [0.08, 0.45]$, $d = 0.23$, $BF_1 = 8.92$, whereas the opposite pattern was observed for the negative frequent group (Niffians Majority: $M = -0.23$, $SD = 1.27$, Laapians Majority: $M = 0.08$, $SD = 1.30$), $t(625) = -3.01$, $p = .003$, 95% $CI_{diff} = [-0.51, -0.11]$, $d = 0.24$, $BF_1 = 12.76$.

Combined analyses of AMP and self-reported evaluation scores. The ANOVA on standardized AMP and self-reported evaluation scores revealed a main effect of Valence Frequency, $F(1, 1253) = 4.07$, $p = .043$, $BF_1 = 1.18$, which was qualified by an interaction of Majority Group and Valence Frequency, $F(1, 1253) = 12.60$, $p < .001$, $BF_1 = 89.12$, and a three-way interaction of Majority Group, Valence Frequency, and Measure, $F(1, 1253) = 6.43$, $p = .011$, $BF_1 = 8.57$.

Proportion positive scores. A 2 x 2 x 2 ANOVA revealed a main effect of Majority Group, $F(1, 1254) = 13.87$, $p < .001$, indicating that participants reported a lower proportion of positive statements for Niffians when Niffians were the majority group ($M = -3\%$, $SD = 20\%$) than when Laapians were the majority group ($M = 1\%$, $SD = 22\%$), $d = 0.21$, $BF_1 = 140.64$. Crucially, we also observed an interaction of Majority Group and Valence Frequency , $F(1, 1527)$

= 15.83, $p < .001$, indicating that the main effect of Majority Group was observed when the majority of the statements were negative, $t(625) = -5.07$, $p < .001$, 95% CI_{diff} = [-11.68, -5.16], $d = 0.41$, $BF_1 > 1000$, but not when they were positive, $t(633) = -0.21$, $p = .83$, 95% CI_{diff} = [-3.33, 2.68], $d = 0.02$, $BF_1 = 0.29$.

Discussion

Experiment 3 showed illusory-correlation effects on self-reported evaluation scores and on proportion estimates. Importantly, however, implicit evaluation scores as obtained with the AMP did not reveal illusory-correlation effects, in line with the results for IAT scores in Experiments 1 and 2.

Experiment 4

Experiment 4 included another common measure of implicit evaluation: the Evaluative Priming Task (EPT: Fazio, Sanbonmatsu, Powell, & Kardes, 1986). In this task, participants evaluate positive and negative words that are preceded by the presentation of a prime that participants are instructed to ignore. For the sake of consistency with the self-report task, the prime consisted of the social group category names (the word Niffian or Laapian: see also Van Dessel et al., 2015). Implicit evaluation scores were computed based on differences in reaction times for positive and negative responses in the context of the different types of primes.

Method

Participants. A total of 3675 English-speaking volunteers were recruited to participate online via the Project Implicit research website (https://implicit.harvard.edu). A total of 369 (10.0%) participants decided not to complete the experiment after receiving information about the duration and the nature of the study and 1415 participants (38.50%) dropped out during the experiment, leaving 1829 participants. The dropout rates were not significantly different across

the experimental conditions, $\chi^2(3)$s < 4.20 , *ps* > .24. Hence, there was no evidence for condition-dependent attrition. Sampling plan was the same as for Experiments 1-3.

Data were excluded for participants who (a) did not fully complete all questions and tasks (314 participants; 16.6%), or (b) had an excessive number of errors (>60%) in the EPT or did not have any trials left in each of the trial conditions following outlier treatment (20 participants; 1.1%). Analyses were performed on the data of 1551 participants (978 women, mean age = 38, *SD* = 15) (Table 3).

Procedure. Procedure was identical to Experiment 3 with the exception that an EPT was completed instead of an AMP. The EPT procedure followed Hu, Gawronski, and Balas (2017) and comprised 3 blocks of 40 test trials. A single trial consisted of the presentation of a fixation cross for 500 ms, a prime for 200 ms, and the presentation of a target word for a maximum of 1500 ms. Targets consisted of 10 positive words (e.g., the words pleasant and good) and 10 negative words (e.g., the words *unpleasant* and *bad*). The prime stimuli were the group names (Niffians and Laapians) (in line with the targets used in the self-report rating task).

Latencies from trials with errors (4.5% of trials) and trials with latencies lower than 300ms (0.4% of trials) or higher than 1000ms (18.6% of trials) were removed (Koppehele-Gossel, Hoffmann, Banse, & Gawronski, 2020). A score was calculated for each social group by subtracting the mean response latency to positive target words preceded by primes related to the social group from the mean response latency to negative target words preceded by the same prime. Scores for Laapians were subtracted from scores for Niffians such that higher values indicate more favorable implicit evaluations of Niffians over Luupians. The split-half reliability of the EPT score was poor, $r(1869) = -.06$. Across groups, participants displayed a small

preference for Laapians over Niffians (M = -3.07, SD = 44.93), t(1261) = -2.69, p = .007, d = 0.07.

The EPT scores correlated significantly with the self-reported evaluation scores (internal consistency: Cronbach's Alpha = .79), r(1549) = .05, p = .056, which also revealed a small preference for Laapians over Niffians (M = -0.08, SD = 1.19), t(1534) = -2.58, p = .010, d = 0.07. The EPT scores also correlated significantly with proportion positive scores, r(1549) = .05, p = .035, which indicated a slightly higher proportion of positive statements for Laapians than for Niffians overall (M = -2%, SD = 19%), t(1550) = -3.18, p = .001, d = 0.08.

Results

EPT scores. We performed a 2 (Majority Group) x 2 (Valence Frequency) x 2 (Task Order) ANOVA on EPT scores. The ANOVA did not show main effects of Majority Group, F(1,1542) = 0.10, p = .75, BF_1 = 0.20, or of Valence Frequency, F(1,1542) = 0.08, p = .77, BF_1 = 0.20. We also did not observe the crucial interaction of Majority Group and Valence Frequency, F(1,1542) = 0.86, p = .36, BF_1 = 0.12, (Table 4), nor any other significant effects, Fs < 0.60, ps > .43, BF_1s < 0.13.

Self-reported evaluation scores. The 2 (Majority Group) x 2 (Valence Frequency) x 2 (Task order) ANOVA on the self-reported evaluation scores revealed a small main effect of Majority Group, F(1,1542) = 6.97, p = .008, indicating that self-reported evaluation scores were higher when Niffians were the majority group (M = 0.00, SD = 1.18) than when Laapians were the majority group (M = -0.17, SD = 1.19), d = 0.14, BF_1 = 7.21. We also observed a significant interaction between Majority Group and Valence Frequency, F(1,1542) = 19.09, p < .001, BF_1 = 925.52. For the positive frequent group, self-reported evaluation scores were higher when Niffians were the majority group (M = 0.14, SD = 1.23) than when Laapians were the majority

group (M = -0.27, SD = 1.33), $t(830) = 4.65$, $p < .001$, 95% CI_{diff} = [0.24, 0.58], d = 0.33, $BF_1 >$ 1000, whereas there was no significant difference for the negative frequent group (Niffians Majority: M = -0.13, SD = 1.11, Laapians Majority: M = -0.03, SD = 1.01), $t(717) = -1.32$, p = .19, 95% CI_{diff} = [-0.26, 0.05], d = 0.10, $BF_1 = 0.55$. There was also an interaction of Valence Frequency x Task Order, $F(1,1542) = 4.25$, p = .039, $BF_1 = 0.62$, but no other significant effects, Fs < 3.61, ps > .057, BF_1s < 0.55.

Combined analyses of EPT and self-reported evaluation scores. The ANOVA on standardized EPT and self-reported evaluation scores revealed an interaction of Majority Group and Task, $F(1, 1542) = 4.55$, p = .030, $BF_1 = 1.32$, an interaction of Majority Group and Valence Frequency, $F(1, 1542) = 13.31$, $p < .001$, $BF_1 > 1000$, and a three-way interaction of Majority Group, Valence Frequency, and Measure, $F(1, 1542) = 6.12$, p = .010, $BF_1 = 3.62$.

Proportion positive scores. The 2 x 2 x 2 x 2 ANOVA revealed a small main effect of Majority Group, $F(1, 1542) = 8.78$, p = .003, indicating that participants reported a lower proportion of positive statements for Niffians when Niffians were the majority group (M = -3%, SD = 20%) than when Laapians were the majority group (M = 0%, SD = 19%), d = 0.12, BF_1 = 2.08. Crucially, we also observed an interaction of Majority Group and Valence Frequency , $F(1, 1542) = 24.14$, $p < .001$, $BF_1 > 1000$, indicating that the main effect of Majority Group was observed when the majority of the statements were negative, $t(717) = -5.18$, $p < .001$, 95% CI_{diff} = [-10.03, -4.51], d = 0.39, $BF_1 > 1000$, but not when they were positive, $t(830) = 1.57$, p =.12, 95% CI_{diff} = [-0.53, 4.74], d = 0.11, $BF_1 = 0.72$.

Discussion

Similar to Experiment 3, Experiment 4 revealed illusory-correlation effects on self-reported evaluation scores and on proportion estimates but not on implicit evaluation scores (as

probed with an EPT). Notably, the internal consistency of the EPT (and of the AMP in Experiment 3) was very low which could represent a limitation of the study. Low reliabilities, however, are not uncommon for EPT and AMP scores (Gawronski & De Houwer, 2014; Van Dessel et al., 2017). Moreover, effects of experimental manipulations can be robustly established even with measures that show suboptimal performance in picking up differences between individuals (De Schryver, Hughes, Rosseel, & De Houwer, 2015), which is reflected in the fact that AMP, EPT, and self-reported measures revealed a preference for Laapians over Niffians.

General Discussion

Four experiments investigated illusory-correlation effects on implicit and explicit evaluations. Participants first read valenced statements about two fictitious social groups, with more total statements about one group (majority) than about another group (minority). Even though the proportion of positive to negative statements was the same for both groups, participants reported a preference for the majority group when there were more positive statements overall, and a preference for the minority group when there were more negative statements. Importantly, implicit evaluations did not reflect this illusory correlation effect. Instead, participants exhibited an overall preference for the majority group on the IAT (i.e., a mere-exposure effect), and no preference for either group on the EPT and AMP.

The current results replicate and extend earlier findings by Ratliff and Nosek (2010), indicating that illusory-correlation effects do not emerge for implicit evaluations. Importantly, this result does not appear to be limited to the IAT and it is not the result of confounds in the original study design. Our results also partially replicate findings by Carraro et al. (2014), showing an effect on IAT scores in an illusory-correlation paradigm, but our modified design

suggests that this effect likely reflects a mere-exposure effect rather than an illusory-correlation effect.

Theoretical implications

The fact that illusory-correlation effects did not emerge for implicit evaluations provides support for guessing-based explanations for the illusory-correlation effect. From this perspective, illusory-correlation effects depend on processes operating during measurement (belief expression) rather than processes operating during learning (belief formation). We speculate that, when participants read the behavioral statements, they do not derive or encode differences in the proportion of valenced statements between the groups. Only when they are asked to self-report their preferences, or to estimate the proportions of valenced information, do they try to distinguish between the two groups, essentially guessing about differences in proportions and resulting evaluations. In this guessing process, they then can make a mistake, judging that positive but not negative behavioral statements had more often been assigned to the majority group (Bulli & Primi, 2006). When evaluation occurs under some of the conditions of automaticity (e.g., there is little time or little effort to control evaluative responses to the stimuli), participants do not readily engage in this guessing process and illusory-correlation effects might not show up.

It is important to note that Experiment 1 did find a small effect of Majority Group and Valence Frequency on IAT scores in one IAT order condition. However, the Bayes Factor indicates no evidence for this interaction effect ($BF_1 = 0.98$) and Experiment 2 did not replicate this effect. Notably, Experiment 2 provided stronger evidence ($BF_1 = 3.26$) for an interaction effect of Majority Group, Valence Frequency, and Task Order on IAT scores. This might indicate that IAT scores *could* reveal small illusory-correlation effects. This effect, however, can be easily accommodated by guessing accounts: Completing self-report measures (which invoke guessing)

before completing the IAT might lead to illusory-correlation beliefs that can influence IAT performance. From this perspective, illusory-correlation effects can be observed irrespective of the measure once participants have engaged in guessing processes. However, this post-hoc explanation should be treated with caution, because the interaction effect was weak, and it was not observed in Experiments 3 and 4.

It is further noteworthy that explicit evaluations and proportion estimates did not always accord with one another. Experiments 1 and 2 provided evidence for illusory-correlation effects on explicit evaluations but not proportion estimates. Though this could be due to issues related to participants not being able to adequately report their actual proportion estimates, it might also reflect a true dissociative pattern. Indeed, Experiments 3 and 4 (and a meta-analysis of these results) also showed evidence for a dissociation in proportion estimates and explicit evaluations. Specifically, the effect on proportion estimates was observed only when the majority of information was negative (in contrast with Hamilton & Gifford, 1976, and Ratliff & Nosek, 2010) whereas explicit evaluations showed an effect also when positive information was more frequent. These results further support the important role of measurement-related processes in illusory-correlation effects. At first glance, however, these effects might seem inconsistent with guessing-based theories given that proportion judgment measures require estimation of information and are therefore likely to invoke guessing the source of positive and negative behavioral statements. One explanation is that participants try to distinguish between groups on the basis of the received information (by engaging in effortful guessing-related processes) only when they are urged to evaluate the groups because differential evaluation is considered more important or because people are more accustomed to doing this (Berndsen & Spears, 1997).

The current results do not fit well with illusory-correlation theories that attribute effects to processes during learning (belief formation). These theories assume that differences in beliefs

about the (relative) number of positive and negative statements for the two groups are learned during encoding (e.g., due to differences in information salience: Hamilton & Gifford, 1976, or differences in the number of learning trials: Fiedler, 1991). These beliefs should influence different measures of evaluative behavior to a similar extent (Hamilton & Gifford, 1976) which contrasts with the evidence for illusory-correlation effects on explicit evaluations but not on implicit evaluations (or proportion estimates). These theories could accommodate our results if they assume that the key processes explaining effects occur (also) during retrieval. The distinctiveness account could explain effects in terms of enhanced availability in memory of distinctive information during retrieval (Tversky & Kahneman, 1973). It is possible that salience of the doubly infrequent events affects judgments only when participants retrieve the information (under non-automatic conditions). Fiedler's information loss account (1991) already distinguishes a separate judgment formation stage; however, it is argued that this occurs on-line (during the pairings). A revised account could potentially designate this process to the measurement phase.

Another explanation to consider is that implicit measures such as the IAT, AMP, and EPT measure evaluation in a more noisy manner than self-report measures (Blanton & Jaccard, 2015) and are therefore less sensitive to (subtle) effects like illusory-correlation effects. While this explanation cannot be ruled out, it seems unlikely to provide a full explanation of our results given that (1) illusory-correlation effects on self-report measures revealed illusory-correlation effects with moderate effect sizes (mean difference in effect size d = 0.41), (2) our experiments had ample statistical power for observing small effects, and (3) implicit measures revealed other evaluative effects (e.g., a preference for the majority over the minority group; a preference for Laapians over Niffians).

Implications for stereotype research and future directions

The current results are relevant for stereotype formation research. If illusory correlations impact stereotype formation and related behavior, then our results suggest that this strongly depends on processes during belief expression. For example, people might be more likely to erroneously designate more positive behaviors to (in-group) majorities than to (out-group) minorities when they are motivated or have ample opportunity to distinguish between groups. This might imply a strong relation between illusory-correlation effects (and stereotype formation) with certain (motivational) personality traits such as generalized prejudice (Bergh & Akrami, 2017) and with certain social environments (e.g., whether a person's social environment reacts positively or negatively to group distinction) (Sears & Patrick, 2003).

In line with previous research, the current study examined illusory correlation effects on evaluations of unfamiliar social groups (i.e., Niffians and Laapians). Thus, it is possible that the obtained asymmetry reflects low sensitivity of implicit evaluation measures to novel information that has not been highly overlearned. In line with this idea, it has been argued that implicit evaluation measures reflect deeply-ingrained, longstanding associations that have been established by years of reinforcement (e.g., Rudman, 2004), which would explain why we found illusory correlation effects based on novel information on explicit, but not implicit, evaluations. However, counter to this interpretation, the available evidence suggests that implicit evaluation measures are highly sensitive to novel information that has not been highly overlearned (for a review, see Gawronski & Sritharan, 2010). Based on this evidence, differential sensitivity to novel information seems unlikely to account for the obtained pattern of results. Nonetheless, it is possible that our results might not generalize to other groups for other reasons. For instance, implicit evaluation measures might be more sensitive to factors unrelated to evaluation for unfamiliar social groups, precluding observation of illusory correlation effects. Future studies

might test this idea by examining illusory correlation effects on implicit evaluations of well-known social groups (e.g., racial in- and out-groups).

Our study also did not test which specific conditions (of non-automaticity) are required for the illusory-correlation effect to arise. We used implicit evaluation measures for which evaluative responding is thought to occur under several conditions of automaticity and we therefore do not know which automaticity features are important. For example, effects could depend on controllability, fastness, intentionality, or motivation. These questions might be tested in future studies that manipulate the automaticity conditions of evaluative responding (see Payne et al., 2008; Van Dessel et al., 2020, for examples of studies that used such manipulations). Future studies could also examine whether illusory-correlation effects require effortful processing. For instance, one could test if effects on implicit evaluation measures are observed when the application of illusory-correlation beliefs is automatized on the basis of extensive practice, whether this relates to real-life stereotypes or implicit prejudice as observed for well-known racial groups (Banaji & Greenwald, 2013), and whether this bias can be re-trained. As such, the current study suggests new directions for research on attitude and stereotype formation.

References

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

Bergh, R., & Akrami, N. (2017). Generalized prejudice: Old wisdom and new perspectives. In C. G. Sibley & F. K. Barlow (Eds.), *The Cambridge handbook of the psychology of prejudice* (p. 438–460). Cambridge University Press.

Blanton, H., & Jaccard, J. (2015). Not So Fast: Ten Challenges to Importing Implicit Attitude Measures to Media Psychology. *Media Psychology*, 18, 338-369.

Carraro, L., Negri, P., Castelli, L., & Pastore, M. (2014). Implicit and explicit illusory correlation as a function of political ideology. *PLoS ONE 9(5):e96312*.

Corneille, O., & Mertens, G. (in press). Behavioral and physiological evidence challenges the automatic acquisition of evaluations. *Current Directions in Psychological Science*.

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

De Houwer, J., Beckers, T., & Moors, A. (2007). Novel attitudes can be faked on the Implicit Association Test. *Journal of Experimental Social Psychology, 43,* 972–978.

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology*, *24*(1), 252–287.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*, 347-368.

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. *Advances in Experimental Social Psychology*, *61*, 127-83.

De Schryver, M., Hughes, S., Rosseel, Y., & De Houwer, J. (2015). Unreliable yet still replicable: A comment on LeBel and Paunonen (2011). *Frontiers in Psychology*. 6: 2039.

Eder, A. B., Fiedler, K., Hamm-Eder, S. (2011). Illusory correlations revisited: The role of pseudo-contingencies and working-memory capacity. *The Quarterly Journal of Experimental Psychology, 64*, 517–532.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*, 229–238.

Fiedler, K. (1991). The tricky nature of skewed frequency tables: An information loss account of distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology, 60*, 24–36

Fiedler, K., & Walther, E. (2004). *Stereotyping as inductive hypothesis testing.* New York: Psychology Press.

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.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd edition, pp. 283–310). New York: Cambridge University Press.

Gawronski, B., De Houwer, J., & Sherman. J. W. (in press). Twenty-five years of research using implicit measures. *Social Cognition*.

Gawronski, B., & Sritharan, R. (2010). Formation, change, and contextualization of mental associations: Determinants and principles of variations in implicit measures. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 216-240). New York, NY: Guilford Press.

Geraerts, E., Bernstein, D.M., Merckelbach, H., Linders, C., Raymaekers, L., & Loftus, E.F.. (2008). Lasting false beliefs and their behavioral consequences. *Psychological Science, 19*, 749–753.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology, 12,* 392-407.

Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: I. The Effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 43*, 43, 17–32

Johnson, C., & Mullen, B. (1994). Evidence for the accessibility of paired distinctiveness in distinctiveness-based illusory correlation in stereotyping. Personality and Social *Psychology Bulletin, 20,* 65–70.

Klauer, K. C., & Meiser, T. (2000) A source-monitoring analysis of illusory correlations. Personality and Social Psychology Bulletin, 26,1074-1093.

Koppehele-Gossel, J., Hoffmann, L., Banse, R., & Gawronski, B. (2020). Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores. *Journal of Experimental Social Psychology, 87:103905*

Kurdi, B., & Dunham, Y. (in press). Propositional accounts of implicit evaluation: Taking stock and looking ahead. *Social Cognition*.

Meiser, T., & Hewstone, M. (2006). Illusory and spurious correlations: Distinct phenomena or joint outcomes of exemplar-based category learning? *European Journal of Social Psychology, 36,* 315–336.

Mullen, B., & Johnson, C. (1990). Distinctiveness based illusory correlations and stereotyping: A meta-analytic integration. *British Journal of Social Psychology, 29,* 11–28.

Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extra-personal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology, 86*, 653-667.

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology, 94,* 16–31.

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

Ratliff, K. A., & Nosek, B. A. (2010). Creating distinct implicit and explicit attitudes with an illusory correlation paradigm. *Journal of Experimental Social Psychology, 46,* 721–728.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test (IAT): Dissociating salience from associations. *Journal of Experimental Psychology: General, 133,* 139–165.

Rudman, L. A. (2004). Sources of implicit attitudes. *Current Directions in Psychological Science, 13,* 79-82

Rydell, R. J., & Gawronski, B. (2009). I like you, I like you not: Understanding the formation of context-dependent automatic attitudes. *Cognition and Emotion, 23,* 1118–1152

Sears, D. O., & Henry, P. J. (2003). The origins of symbolic racism. *Journal of Personality and Social Psychology, 85,* 259–75.

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

Van Dessel, P., De Houwer, J., Gast, A., & Smith, C. T. (2015). Instruction-based approach–avoidance Effects: changing stimulus evaluation via the mere instruction to approach or avoid stimuli. *Experimental Psychology, 62,* 161-169.

Van Dessel, P., Gawronski, B., & De Houwer, J. (2019). Does explaining social behavior require multiple memory systems? *Trends in Cognitive Sciences, 23,* 368-369.

Van Dessel, P., Mertens, G., Smith, C. T., & De Houwer, J. (2017). The mere exposure instruction effect: Mere exposure instructions influence liking. *Experimental Psychology, 64,* 299-314.

Van Dessel, P., Cone, J., Gast, A., & De Houwer, J. (2020). The impact of valenced verbal information on implicit and explicit evaluation: the role of information diagnosticity, primacy, and memory cueing. *Cognition and Emotion, 34*, 74-85,

Wells, C., Reedy, J., Gastil, J., & Lee, C. (2009). Information distortion and voting choices: The origins and effects of factual beliefs in initiative elections. *Political Psychology, 30*, 953-969.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127*, 797–826

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monographs, 9*, 1-27.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 11*, 493-504.

Tables

Table 1. *Total number of participants who completed Experiment 1 and 2 in the 4 experimental conditions as a function of Task Order (IAT/self-report task first) and IAT Order (Block with Niffians and positive together first/ Block with Laapians and positive together first).*

		IAT first		Self-report first	
		IAT Order 1	*IAT Order 2*	*IAT Order 1*	*IAT Order 2*
		N *(% of total)*	*N* *(% of total)*	*N* *(% of total)*	*N* *(% of total)*
Experiment 1					
Niffians majority	Positive frequent	186 (12.1%)	196 (12.8%)		
	Negative frequent	194 (12.6%)	198 (12.9%)		
Laapians majority	Positive frequent	188 (12.2%)	187 (12.2%)		
	Negative frequent	186 (12.1%)	200 (13.0%)		
Experiment 2					
Niffians majority	Positive frequent	92 (6.4%)	89 (6.2%)	96 (6.7%)	87 (6.1%)
	Negative frequent	91 (6.4%)	101 (7.1%)	82 (5.7%)	95 (6.6%)
Laapians majority	Positive frequent	81 (5.7%)	111 (7.8%)	73 (5.1%)	94 (6.6%)
	Negative frequent	88 (6.1%)	84 (5.9%)	78 (5.4%)	90 (6.0%)

Table 2. *Overview of mean IAT, self-report rating and proportion positive scores in Experiments 1 and 2 and the difference in means between the two Majority Group conditions.*

	Positive frequent			Negative frequent		
	Niffians majority	Laapians majority	Diff	Niffians majority	Laapians majority	Diff
IAT	Exp 1: -0.04 (0.42)	Exp 1: -0.17 (0.41)	0.13	Exp 1: -0.05 (0.43)	Exp 1: -0.17 (0.41)	0.12
	Exp 2: -0.03 (0.41)	Exp 2: -0.20 (0.40)	0.17	Exp 2: -0.05 (0.42)	Exp 2: -0.17 (0.43)	0.12
Self-report	Exp 1: 0.02 (1.38)	Exp 1: -0.37 (1.40)	0.39	Exp 1: -0.19 (1.38)	Exp 1: -0.03 (1.40)	-0.16
	Exp 2: 0.03 (1.11)	Exp 2: -0.19 (1.15)	0.22	Exp 2: -0.22 (1.17)	Exp 2: 0.00 (1.15)	-0.22
Prop pos	Exp 1: -0.07 (0.22)	Exp 1: 0.03 (0.23)	-0.10	Exp 1: -0.11 (0.22)	Exp 1: 0.07 (0.20)	-0.18
	Exp 2: -0.04 (0.20)	Exp 2: 0.02 (0.20)	-0.06	Exp 2: -0.10 (0.22)	Exp 2: 0.08 (0.22)	-0.18

Table 3. *Total number of participants who completed Experiment 3 and 4 in the 4 experimental conditions as a function of Task Order (automatic/self-report task first).*

		Automatic evaluation task first	Self-report task first
		N *(% of total)*	*N* *(% of total)*
Experiment 3			
Niffians majority	Positive frequent	165 (11.8%)	138 (9.9%)
	Negative frequent	187 (13.4%)	156 (11.2%)
Laapians majority	Positive frequent	185 (13.3%)	147 (10.5%)
	Negative frequent	163 (11.7%)	121 (8.7%)
Experiment 4			
Niffians majority	Positive frequent	221 (14.2%)	196 (12.6%)
	Negative frequent	190 (12.3%)	191 (12.3%)
Laapians majority	Positive frequent	194 (12.5%)	221 (14.2%)
	Negative frequent	174 (11.2%)	163 (10.5%)

Table 4. *Overview of mean AMP, EPT, self-report rating and proportion positive scores in Experiments 3 and 4 and the difference in means between the two Majority Group conditions.*

	Positive frequent			Negative frequent		
	Niffians majority	Laapians majority	Diff	Niffians majority	Laapians majority	Diff
AMP	Exp 3: -0.02 (0.11)	Exp 3: -0.01 (0.10)	-0.01	Exp 3: -0.01 (0.09)	Exp 3: 0.01 (0.09)	-0.02
EPT	Exp 4: -2.68 (47.06)	Exp 4: -3.91 (45.67)	1.23	Exp 4: -4.28 (44.95)	Exp 4: -1.16 (41.28)	-3.12
Self-report	Exp 3: 0.04 (1.22)	Exp 3: -0.23 (1.13)	0.27	Exp 3: -0.23 (1.27)	Exp 3: 0.08 (1.30)	-0.31
	Exp 4: 0.14 (1.23)	Exp 4: -0.27 (1.33)	0.41	Exp 4: -0.13 (1.11)	Exp 4: -0.03 (1.01)	-0.10
Prop pos	Exp 3: -0.01 (0.19)	Exp 3: -0.01 (0.20)	0.00	Exp 3: -0.05 (0.20)	Exp 3: 0.04 (0.21)	-0.09
	Exp 4: 0.00 (0.20)	Exp 4: -0.02 (0.19)	0.02	Exp 4: -0.05 (0.19)	Exp 4: 0.02 (0.18)	-0.07