Does Information about Bias Attenuate Selective Exposure? The Effects of Implicit Bias Feedback on the Selection of Outgroup-Rich News

Kroon, A.C.; van der Meer, T.G.L.A.; Pronk, T.

DOI
10.1093/hcr/hqac004

Publication date
2022

Document Version
Final published version

Published in
Human Communication Research

Human Communication Research 0360-3989

ORIGINAL RESEARCH

Does Information about Bias Attenuate Selective Exposure? The Effects of Implicit Bias Feedback on the Selection of Outgroup-Rich News

Anne C. Kroon[1], Toni G. L. A van der Meer[1], & Thomas Pronk[2,3]

1 Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, The Netherlands
2 Faculty of Social and Behavioural Sciences, University of Amsterdam, Amsterdam, The Netherlands
3 Open Science Tools (PsychoPy) lab, School of Psychology, University of Nottingham, Nottingham, UK

People's news diets are shaped by a diverse set of selection biases that may be unconscious in nature. This study investigates whether providing individuals with information about such unconscious biases attenuates selective exposure. More specifically, in two selective-exposure experiments among Dutch ingroup members focusing on ethnic ($N = 286$) and religious ($N = 277$) minorities, we expose individuals to their unconscious prejudices as measured by the Implicit Association Test (IAT) before documenting their news-selection patterns. Findings indicate that the effectiveness of this awareness-inducing strategy depends upon existing levels of implicit and explicit prejudice and overly expressed acceptance of the IAT scores. This implies that raising awareness of implicit prejudice works as an effective strategy for fighting biased news selection for some, but may backfire for others, and should therefore only be implemented with caution and attention for explicit considerations.

Keywords: Selective Exposure, Outgroups, Media Literacy, Implicit prejudice, Implicit Bias Training

https://doi.org/10.1093/hcr/hqac004

In multicultural societies, public and scholarly concerns about the role of media in fueling ethnic and religious conflicts and divides have proliferated. Strong evidence supports the notion that media exposure plays a significant role in the formation and maintenance of perceptions of self and others, shaping decision-making processes and intergroup biases (Mastro, 2015). The choices that audience members make themselves largely determine to what extent negative outcomes of media bias may materialize: In today's high-choice media environments, people are given

Corresponding author: Anne C. Kroon; e-mail: a.c.kroon@uva.nl

unprecedented agency in crafting potentially biased media diets by self-selecting into ingroup-rich content, and preferring negative (vis-à-vis positive) content about outgroup members (Appiah et al., 2013; Knobloch-Westerwick & Hastall, 2010). The current research is primarily designed to understand how to counter the origins of biased selection patterns to ensure less stereotypical news exposure.

Biased news selection decisions are profoundly channeled by implicit biases that operate largely under the radar of conscious awareness and may persist even in the presence of nonprejudiced attitudes on the conscious level. Unconscious mechanisms have been argued to be especially powerful in predicting news choice in socially sensitive domains such as intergroup prejudice, where self-reports likely suffer from impression-management bias (Arendt et al., 2016; Knobloch-Westerwick, 2015). As a result, unconscious forces may frustrate the selection of diverse perspectives, angles, and sources in news about outgroup members (Arendt et al., 2016, 2019; Kroon et al., 2020). It has been argued that a *lack of awareness* of implicit prejudice obstructs individuals to fully apprehend and dread its impact on judgment and behavior (Pronin & Kugler, 2007), such as daily (news) selection decisions (e.g., Devine, Forscher, Austin, & Cox, 2012).

This study investigates the potential of confronting audience members with their implicit ethnic and religious prejudice as a remedy for biased selectivity in the context of outgroup-rich news messages. The potential of implicit bias training for behavioral change in diverse domains of social life has sparked the rife interest of policymakers, practitioners, and researchers. Such trainings have been developed for the workplace (e.g., Jackson et al., 2014), yet, its potential for news media literacy interventions has not been established. More particular, and despite robust support for the drivers of biased selection patterns, empirical research into *how* the selection of auspicious outgroup-rich news can be fostered is virtually non-existent. As a consequence, critical questions remain as to how to design interventions stimulating more favorable selection patterns regarding outgroup-rich news. The current study connects and expands on theoretical conceptualizations of media literacy, implicit and explicit prejudice, and selective exposure. We test our hypotheses in a series of two experiments among Dutch ingroup members across two domains of intergroup bias: Study 1 tests the effects of exposure (vs. no exposure) to feedback on implicit prejudice towards *ethnic outgroups* ($N = 286$). Study 2 replicates the findings in the context of prejudice towards *religious outgroups* and tests how a news media literacy message intervenes with receiving feedback on implicit prejudice scores ($N = 277$).

We add to the literature in the following ways. Our innovative design, which allowed for real-time exposure to implicit prejudice as registered with a set of Implicit Association Tests (IATs), permits us to uniquely trace the effects of exposure to individualized and genuine levels of implicit bias on selection patterns. The study's findings offer empirical insight into our theoretical expectations by rigorously testing hypotheses across multiple news selection tasks in the dual context of racial and religious prejudice. Together, we further our understanding of how the

selection of favorable news about outgroups can be fostered, by dissecting under which circumstances individualized feedback on implicit prejudice may serve as an effective intervention tool swaying people towards more constructive news selection decisions. On a more general level, the findings offer a starting point to think about how to remedy selection biases in different domains, herewith speaking to the literature on selective exposure more broadly.

Selective Exposure to News about Outgroups

News messages represent an important source of information about the characteristics of socially 'others': Mediated exposure to outgroup members has the potential to decrease prejudice—but such favorable outcomes are only to be expected if individuals self-select into favorable outgroup-rich content at sufficient rates (Schieferdecker & Wessler, 2017). Unfortunately, early studies on group-related selection biases suggest that people commonly gravitate towards media messages that boost self-esteem and elevate social identity (Appiah et al., 2013; Knobloch-Westerwick & Hastall, 2010). These media choice patterns can be explained from a social identity perspective (Tajfel & Turner, 1979): People seek to boost self-esteem by preferring media content that makes membership of their social group seem relatively more desirable (Harwood, 1999). Dominant patterns of selective media choice tend to exacerbate pre-existing (potentially prejudiced) beliefs about outgroup members, especially when individuals strongly identify with their ingroup (Schieferdecker & Wessler, 2017).

The well-established literature on selective exposure posits that in order to successfully break the self-fueling cycle of biased selectivity and effects (Schemer, 2012), one should withstand the guiding influence of a broad set of human biases that simultaneously drive news selection. First, classic news factor research predicts *negative* events are particularly pertinent for news coverage *and* selection (Knobloch-Westerwick et al., 2020). This *negativity bias* is also explained from the evolutionary perspective that humans are wired to be more attentive and responsive to negative cues in order to avoid threats (Soroka & McAdams, 2015). Second, *confirmation bias* fosters the selection of news content that resonates with our perceptional screens. The vast body of literature on this phenomenon relies on overtly expressed (i.e., explicit) measures to tap into attitudinal (dis)congruence as drivers of selection.

Yet, rather than an entirely controlled and rational process recent empirical evidence suggests that, especially in today's high-choice media environments, audience members are generally not aware of the mechanisms channeling news media choice (Knobloch-Westerwick, 2015). Implicit drivers of confirmation bias are considered especially daunting from a societal perspective, as they may prevent exposure to opposing perspectives without individuals' awareness of such drivers (Arendt et al., 2019).

Consequently, scholars have argued that selective exposure designs should acknowledge predictors governing news choice on two levels: Overtly expressed, 'explicit' measures—considered as more or less deliberate appraisals—and automatically activated 'implicit' evaluations, often conceptualized as individual's 'gut-feelings' (Arendt et al., 2019; Gawronski & Bodenhausen, 2006). Particularly, sharing and selecting news messages is informed by both conscious, verbalized ('explicit') and impulsively activated ('implicit') evaluations (Arendt et al., 2016, 2019; Arendt & Karadas, 2020; Galdi et al., 2012; Kroon et al., 2020). Empirical research confirms that both concepts predict unique variance in media choice, and should thus be used in sync when studying media choice (Arendt et al., 2016, 2019; Arendt & Karadas, 2020).

Implicit evaluations are often measured with the IAT (Greenwald et al., 2003), which relies on response latency and taps into the strength of associations between social categories (e.g., 'Muslims') and evaluative attributes (e.g., 'good' or 'bad'). Implicit measures like the IAT are seen as especially insightful to measure socially sensitive concepts—such as intergroup prejudice—as the activation occurs automatically and is at least partly beyond people's control (Arendt, 2013).

It has long been recognized that implicit prejudices are important predictors of subsequent decision-making (e.g., Nosek, Graham, & Hawkins, 2010). Likewise, the idea that automaticity plays an important part in media choice is not new: Earlier inquiries into news choice already argue that audience members are often not aware of the processes channeling media choice (Shoemaker & Vos, 2009). In recent years, these assumptions have been supported by empirical evidence showing that both implicit and explicit evaluations guide which news people choose to read (Arendt et al., 2016, 2019; Kroon et al., 2020). It follows that prejudices on the explicit and implicit level causes individuals to be at risk of crafting a biased news diet, triggering exposure to content that reinforces biased beliefs—further fueling anti-outgroup attitudes. We expect

H1: Implicit prejudice (a) negatively predicts the selection of positive news about minorities and (b) positively predicts the selection of negative news about minorities.

H2: Explicit prejudice (a) negatively predicts the selection of positive news about minorities and (b) positively predicts the selection of negative news about minorities.

Fostering Selection into Auspicious News about Outgroups

News Media Literacy

To stimulate 'healthier' news selection behavior, news media literacy (NML) interventions have been proposed as a valuable journalistic tool to help audience members make more mindful selection decisions (Livingstone, 2004). In the context of

news about minorities, NML interventions have been proposed as a potentially effective way to implement evidence-based solutions to stereotypical thinking by bolstering critical evaluations of media portrayals of outgroup members (Ramasubramanian, 2007)—although such interventions may also backfire (see Nathanson, 2002; Steinke et al., 2007). For example, explicit instructions to suppress stereotypical thinking has been associated with *increased* levels of stereotypical thinking (Monteith, Sherman & Devine, 1998).

Yet, we know strikingly little about how NML interventions may help transform audience members' selection patterns in the context of intergroup news. Available evidence regarding the effectiveness of NML interventions in intergroup settings pertains exclusively to *forced-exposure* designs: Despite conclusive support for the various factors driving media choice, research into the effectiveness of coordinated media literacy interventions to foster the selection of more auspicious news about outgroup members has remained at its infancy.

Designed with the aim to reduce the harmful effects of the media and prevent undesirable behavior, NML interventions typically educate the audience about diverse aspects of the media (Jeong et al., 2012). Recent empirical studies on the selection of political news have started to investigate the potential of NML interventions to promote the skills and knowledge needed for people to navigate their information environments more mindfully. This work suggests the audience's news literacy orientations shape incidental and selective exposure and news sharing on social media (Vraga & Tully, 2019b). Furthermore, NML interventions are helpful in fostering skepticism towards misinformation (Tully et al., 2020) and limiting partisan selective exposure among Republicans (but not Democrats) (Vraga & Tully, 2019a).

The conditional nature of NML interventions' success is further confirmed by Van der Meer and Hameleers (2020), who demonstrated differences in effectiveness across issue publics and party affiliations. The authors aimed to stimulate cross-cutting media diets using injunctive and descriptive normative language. This attempt proved successful for pro-immigrant partisans but backlashed for partisans with anti-immigrant attitudes. For the latter group, the NML intervention triggered *less* cross-cutting news selection. The scholars argue that NML messages should be tailored to prior issue-beliefs to be successful.

In conclusion, although it is challenging to change selection behavior with a single message (Tully et al., 2020), NML interventions may—when tailored towards audience members' perceptional screens—effectively change selection behavior. Similar interventions might be promising to bolster a more auspicious news diet when it comes to the selection of outgroup-rich news content. As individuals tend to underestimate their own levels of bias—partly due to contemporary egalitarian norms (West & Eaton, 2019)—audience members may believe they are free from bias and prejudice; therefore, it would be insufficient to focus on explicit prejudices. Our aim is, consequently, to investigate to what extent confronting individuals with their level of implicit prejudice might be a useful element of NML interventions.

Implicit Bias Interventions

Implicit biases are influential behavioral determinants *precisely because* they operate largely under the radar of conscious awareness. Therefore, it is often assumed by both researchers and policymakers that individuals should acknowledge that they harbor implicit intergroup biases in order to counteract biased behavior (e.g., Devine et al., 2012). In line with this idea, Implicit Bias Training (IBT) has been proposed as a promising solution to contest intergroup bias (e.g., Jackson et al., 2014). IBT-interventions vary in format and style, but they typically focus on the acknowledgment of unconscious bias as the key to behavioral change (Hahn & Gawronski, 2019). More specifically, IBT represents an intervention program or session that often confronts individuals with information on unconscious and unintentional biases and offer strategies for reducing unintentional biased cognitions and behavior. Some programs simply inform individuals about implicit bias (Atewologun et al., 2018; Lee, 2017). Other programs administer an IAT at the start of the training program, after which strategies are discussed that might alleviate individual's level of bias or its impact on behavior. Afterwards, another IAT might be administered to track changes in implicit prejudice (Applebaum, 2019).

Despite its appeal to organizations, politicians, and the general public, the scientific community is in dispute about the effectiveness and fitness of IBT to tackle inequality. Some have argued that IBT is not an appropriate strategy to resolve intergroup issues in the first place, as it accentuates agency at the individual level over structure on the institutional, organizational, and political level—herewith obscuring structural sources of bias as a site for solutions (Pritlove et al., 2019). Moreover, as empirical work on the effects of IBT is just starting to take shape, the circumstances under which exposure to implicit bias feedback may foster corrective action are not well understood.

In the current study, we examine whether the *potential* merits of IBT interventions may travel to the domain of news selection. More specifically, adopting the argument that awareness is a necessary pre-condition for behavioral change (e.g., Pronin & Kugler, 2007), the current study asks whether providing individuals with feedback on their performance on the IAT may offer a fruitful intervention tool to fight biased news diets for individuals with varying levels of implicit prejudice. When individuals are aware of their own implicit biases, and how such biases may affect their subsequent news selection decisions, they may feel inclined to change their selection behavior. Specifically, they may aim to alleviate the impact of their implicit biases by avoiding negative stereotypical news about outgroup members.

In favor of this line of thought, a recent experimental study among Indian journalists suggests that exposure to implicit gender prejudice discourages bias in journalistic productions (Kalra & Boukes, 2020). As the production of journalistic content is a deliberative and time-consuming task for which journalists are liable, awareness of implicit bias might channel into motivation to reduce the production of biased content.

Whether such a strategy also works for audience members is an open empirical question. The following research question is central to the current inquiry:

RQ1: *Does providing feedback on the strength of one's individual implicit prejudice elicit a beneficial (dampening) effect on the predictive power of implicit biases on news selection?*

The Moderating Role of Implicit, Explicit Prejudice, and Acceptance of IAT-Feedback

Moving towards a more sophisticated understanding of the effects of IAT-feedback, we examine to what extent these effects are contingent upon self-reported attitudes towards outgroups and acceptance of IAT feedback. Herewith, we aim to better understand for whom exposure to implicit prejudice scores might be a viable element of media literacy interventions.

First, the extent to which IAT feedback may trigger changes in news selection behavior may depend upon the (mis)alignment between implicit and explicit levels of prejudice. At a societal level, overtly expressed prejudiced attitudes have steadily declined over time (Dovidio et al., 2016; Gaertner & Dovidio, 2005), despite that subtle, implicit forms of bias persist (Pettigrew & Meertens, 1995). People are generally aware of such societal norms favoring egalitarianism (West & Eaton, 2019), and therefore aim to avoid appearing prejudiced. To maintain "better-than-average" nonprejudiced views of oneself, people regularly assess themselves as less biased, but also as less susceptible to cognitive biases than others (Howell & Ratliff, 2017; Pronin & Kugler, 2007; Saucier, 2002).

Yet, differences exist in the extent to which individuals aspire to (appear as) an unbiased, egalitarian individual. Within certain social groups, the expression of blatant and stereotypical views of outgroups members remain uncontested. For example, right-wing populist campaigns have been shown to blatantly dehumanize immigrants and other minority groups (Ahmed & Matthes, 2017; Betz, 2013; Schemer, 2012). Likewise, overt expressions of racism is increasingly normalized in online anti-immigrant communities (Ekman, 2019; Rogers, 2020).

Arguably, people that openly express prejudiced views of minority members might respond differently to implicit bias feedback, as they do not feel threatened in their unprejudiced self-concept. On the contrary, they may simply recognize themselves in such feedback, and may not feel the urge to change their (selection) behavior accordingly. For individuals that openly express weak levels of explicit prejudice, on the other hand, exposure to implicit prejudice scores may threat egalitarian self-views and incite feelings of cognitive inconsistency. These individuals may be inclined to adjust their selection behavior to fit their unprejudiced self-concept, for example by avoiding negative news. In short, when implicit bias feedback diverges from explicit attitudes, individuals may feel motivated to adjust their behavior, while this might not be the case when explicit and implicit prejudice levels coincide. We ask:

RQ2: *Does explicit prejudice moderate the interaction between implicit prejudice and feedback, so that the dampening effect of providing feedback is more pronounced in individuals with weak rather than strong levels of explicit prejudice?*

Second, whether IAT feedback leads to different selection patterns may depend upon the extent to which individuals accept IAT feedback as a credible reflection of their unconscious prejudices. Empirical work in the field of social psychology has found that individuals who do not recognize themselves in implicit bias feedback may respond defensively and reject the IAT's ability to accurately capture unconscious bias (Sukhera et al., 2019). Particularly, if the IAT feedback suggested more bias than individuals attributed to themselves, they were more prone to distrust the IAT's validity (Howell et al., 2015). Seeking to explain these findings, qualitative research has found that individuals experience frictions between acceptance and justification after being exposed to their implicit biases, ultimately revealing conflicts between individuals' actual and idealized identities (Sukhera et al., 2018).

Based on this literature, we expect that implicit prejudice interacts with the degree to which individuals recognize themselves in the IAT feedback and accept such feedback as and accurate and credible reflection of one's unconscious biases. This, in turn, may lead to differential news selection outcomes. More in particular, news users without implicit prejudice who accept their IAT feedback may feel strengthened and validated in their egalitarian self-concept, which may motivate them to align their behaviors, accordingly, resulting in favorable selection patterns. Acceptance, on the other hand, will mean something different for individuals with implicit bias. For these individuals, acceptance indicates explicit agreement with the IAT scores and tolerance towards their unconscious ethnic and racial prejudices. If implicitly prejudiced individuals agree that IATs genuinely and accurately reflect their unconscious self, they may appreciate implicit prejudice as a legitimate basis for behavior. Contrarily, individuals that feel that feedback indicating implicit prejudice is not representative of their (unprejudiced) self-concept, might be more inclined to align future news selection decisions with their desired self-image. We ask:

RQ3: *Does acceptance of feedback regarding one's implicit prejudice increase the unfavorable influence of implicit biases on news selection?*

Study 1

Method

Two online experiments were registered during 14 and 19 December 2018. Both experiments build on insights from the field of selective exposure, and explicitly model factors that are known to affect news selection. More specifically, in our model predicting selectivity, we include *explicit* and *implicit prejudice, negativity bias*, and *source bias*. By doing so, we robustly test whether exposure to feedback on implicit biases predict selectivity *above* and *beyond* these biases in news selection.

Table 1. Predicting the Selection of Negative (Models-I) and Positive (Models- II) News about Ethnic Outgroups with Exposure to IAT-Feedback (Study 1)

	Selection of negative news				Selection of positive news			
	Model 1 I *b* (SE)	Model 2 I *b* (SE)	Model 3 I *b* (SE)	Model 4 I *b* (SE)	Model 1 II *b* (SE)	Model 2 II *b* (SE)	Model 3 II *b* (SE)	Model 4 II *b* (SE)
Intercept	52.14***	46.65***	43.84***	42.58***	57.80***	62.24***	58.82***	56.90***
	(2.86)	(3.11)	(3.86)	(4.33)	(2.86)	(3.09)	(3.84)	(4.33)
Exposure to IAT feedback	2.33	2.45	6.03	7.32	−2.77	−3.02	1.36	3.30
	(3.20)	(3.12)	(4.28)	(4.82)	(3.20)	(3.11)	(4.26)	(4.82)
Implicit prejudice		7.66*	14.50*	23.96**		−0.37	7.97	14.65†
		(3.08)	(6.38)	(7.92)		(3.07)	(6.35)	(7.92)
Explicit prejudice		0.12*	0.12*	0.24		−0.23***	−0.23***	−0.05
		(0.06)	(0.06)	(0.20)		(0.06)	(0.06)	(0.20)
Exposure to IAT feedback * Implicit prejudice			−8.67	−20.28*			−10.58	−17.89*
			(7.09)	(8.85)			(7.05)	(8.84)
Exposure to IAT feedback * Explicit prejudice				−0.14				−0.19
				(0.22)				(0.22)
Implicit prejudice * Explicit prejudice				−0.42				−0.37
				(0.26)				(0.26)
Exposure to IAT feedback * Implicit prejudice * Explicit prejudice				0.54†				0.40
				(0.30)				(0.30)
Log-likelihood	−1283.7	−1275.6	−1274.8	−1272.4	−1283.7	−1274.4	−1273.3	−1272.1

Note. †$p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$. $N = 286$.

Table 2. Predicting the Selection of Negative (Models-I) and Positive (Models- II) News about Ethnic Outgroups with Acceptance of IAT feedback scores (Study 1)

	Selection of negative news			Selection of positive news		
	Model 1 I *b* (SE)	Model 2 I *b* (SE)	Model 3 I *b* (SE)	Model 4 I *b* (SE)	Model 1 II *b* (SE)	Model 2 II *b* (SE)
Intercept	43.09***	35.93***	49.72***	52.28***	53.43***	53.87***
	(4.40)	(5.08)	(6.78)	(4.43)	(5.06)	(6.89)
Acceptance of IAT feedback	2.42**	2.79**	0.48	0.59	1.37	1.29
	(0.89)	(0.96)	(1.22)	(0.90)	(0.96)	(1.24)
Implicit prejudice		9.09*	−20.89†		−0.87	−1.82
		(3.60)	(10.60)		(3.58)	(10.77)
Explicit prejudice		0.09	0.05		−0.25***	−0.25***
		(0.07)	(0.07)		(0.07)	(0.07)
Acceptance of IAT feedback * Implicit prejudice			5.85**			0.18
			(1.95)			(1.98)
Log-likelihood	−1019.1	−1012.3	−1007.8	−1020.7	−1011.4	−1011.4

Note. †$p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$. The presented findings are tested among the sample of participants that were exposed to their implicit biases: Participants in the non-exposure condition are not included. $N = 229$.

Study 1 focuses on the selection of messages about one of the main ethnic outgroups in the Netherlands: Moroccans. This group is subject to negatively biased news coverage and persistent prejudice and discrimination in the Netherlands. The aim of Study 1 was to test the effects of awareness of implicit biases across message types (negative *vs.* positive) and individuals. The design concerned a single factor (exposure *vs.* no exposure to IAT feedback) varied between subjects.

Sample

An international polling agency (SSI/Dynata) was used to recruit a diverse sample of Dutch citizens (>18 years of age). A total of 567 participants started the experiment. We ensured that the final sample attentively read the instructions and news headlines and leads by removing participants that failed either the first ($n = 119$) or second ($n = 95$) attention check at the beginning of the experiment. Thirty-nine participants did not fully complete the experiment. Two observations were removed because 1 person participated twice. For 20 participants, valid *D*-scores could not be calculated.[1] Additionally, if participants indicated that they themselves or their parents are of Moroccan descent ($n = 6$), they were removed from the sample. A final sample of $N = 286$ participants met the inclusion criteria and successfully and

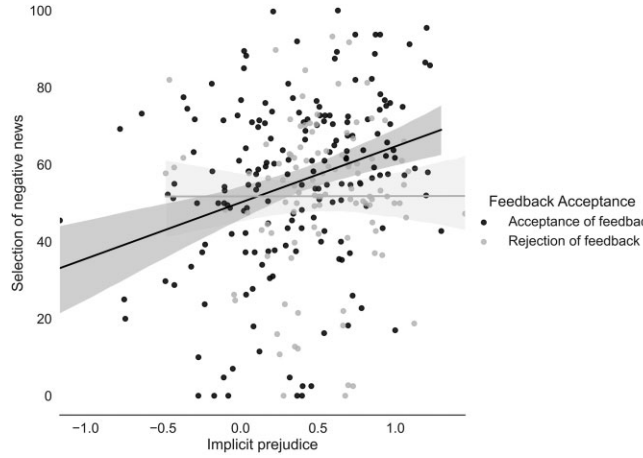

Figure 1 The two-way interaction effects of acceptance of IAT feedback and implicit prejudice on the selection of negative news about ethnic outgroups (Study 1).

attentively completed the experiment (completion rate: 50.8%). The mean age of the sample is $M = 40.65$ ($SD = 13.61$), of which 155 (54.2%) self-identified as female. 17.1% of participants were lower educated, 42.7% had a moderate level of education, and 40.2% were higher educated.

Procedure

The online experiments were administered in Qualtrics. Upon entering the online survey environment, participants were informed that completion of an IAT was part of the study. They were asked to comply with the informed consent and were randomly assigned to the exposure or no-exposure condition and asked to comply with the informed consent procedure. Subsequently, participants were asked to complete a series of questions measuring demographic, moderating, and control variables. All participants were subsequently automatically redirected to the external online environment JASMIN (Pronk, 2014) implementing the IAT measuring their implicit associations[2]. Upon completion, they were redirected back to the Qualtrics environment. Participants in the exposure condition received feedback on their implicit prejudice score and an interpretation of the score was provided. Participants in the non-exposure condition proceeded immediately to the selection of news headlines and leads. We asked participants to rate a feed of online headlines and leads on the likelihood of selection. Finally, participants were debriefed, and informed that the headlines and teasers they read were manipulated for the purpose of the study to investigate the influence of implicit and explicit stereotypes on the selection of news. Participants were redirected to the panel company website, where they received incentive gift vouchers.

Pilot Test

A broad set of negative and positive headlines and leads about ethnic (Moroccans) and religious (Muslims) outgroups were developed based on actual Dutch news and

subject to extensive pre-testing in a pilot study among 36 Dutch citizens (56.9% female, M_{age}=43.44, SD = 13.76). More specifically, a series of ten headlines and five leads × two outgroups were evaluated. The final selected negative and positive headlines and leads differed significantly regarding perceived support (vs. lack of support) for societal participation of Moroccans/Muslims, tolerance (vs. lack of tolerance) towards Moroccans/Muslims, and valence (negative vs. positive)—confirming the effectiveness of the manipulation. The final headlines and leads, on the other hand, proved comparable regarding perceived arousal and likelihood of selection, further confirming internal validity. In addition, generally participants indicated to find the headlines and leads somewhat comparable to other news about Moroccans/Muslims. Supplementary Appendix A, Table A1 details final headlines and leads, Table A2 provides the summary statistics of the pilot data.

Dependent Variables

To measure news selection likelihood, participants were presented with four (two negative and two positive) headlines and four (two negative, two positive) leads about Moroccans. The headlines and leads appeared in random order and resembled a social media newsfeed (see Figure A1, Supplementary Appendix A). For each headline and lead, participants were asked to indicate how likely they would read the related article if they came across it in their everyday life (*0 = not likely at all*, *100 = very likely*).

Negative News Selection

The items measuring reading likelihood of the four negative news items (two headlines, two teasers) were averaged for an overall index of negative news selection. Higher scores indicate a stronger likelihood to self-select into negative news about ethnic outgroup members.

Positive News Selection

Likewise, for the measurement of positive news selection, we took the composite mean scores of the four positive news items (two headlines, two teasers). Higher scores indicate a stronger likelihood to select positive news about ethnic outgroup members.

Independent Variables

Exposure (vs. No Exposure) to IAT Feedback

Implicit prejudice towards racial outgroups was assessed with a Moroccan-Dutch IAT specifically designed for the current study. This IAT measured implicit association between Moroccan (vs. Dutch) individuals and good (vs. bad) attributes. The IAT consisted of seven blocks (Greenwald et al., 2003). For the target categories, verbal ('Moroccan' vs. 'Dutch') and visual (photographic) stimuli were used. For the visual stimuli, we relied on the Radboud Faces Database (RaFD) to select photos of Arabic and Caucasian males with neutral facial expressions (Langner et al.,

2010). For the attribute categories, we relied on commonly used stimuli to measure 'good' (wonderful, lovely, fantastic, excellent) and 'bad' (rotten, dirty, horrible, awful) categories. An automatized script computed the *D*-score (Greenwald et al., 2003) for individual participants under the hood and immediately after completion of the IAT. The *D*-score indicates the level of implicit prejudice, with higher levels indicating faster cognitive pairing of Moroccans (vs. Dutch) and bad (vs. good). We used the D_{2SD} scoring algorithm to transform the data (Greenwald et al., 2003).

After completing the IAT, *D*-scores were converted into personalized feedback statements, consisting of the actual test score and an interpretation, starting with the following statement: "*Test score: [.]. You have just completed an Implicit Association Test. Your (unconscious) prejudices against Moroccans were measured*". Based on the standard cut-off points of the *D*-scores (Hahn & Gawronski, 2019; see http://projectimplicit.net/) the interpretation messages relied on the following format: *No implicit bias* for $D \leq .15$[3], *slight bias* for $.15 < D \leq .35$, *moderate bias* for $.35 < D \leq .65$, and *strong bias* for $D > .65$. For example, participants with moderate implicit bias received the following message: "The test score indicates moderate implicit bias toward Moroccans. More specifically, the test just conducted shows that you have a moderate automatic preference for other people over Moroccans. This means that you (at an implicit level) associate Moroccans more strongly with bad qualities than the Dutch." Only participants in the exposure condition saw their test scores and personalized feedback statements. Relying on the *D*-score and associated feedback statement, four relevant exposure groups were imputed (*no bias*, *slight bias, moderate bias, strong bias).* Participants in the no-exposure condition continued immediately to the selection of headlines and leads.

Implicit Prejudice

As an indication of implicit confirmation bias, we measured implicit prejudice towards Moroccans. Specifically, the continuous D_{2SD}-scores resulting from the Moroccan-Dutch IAT were used. Higher scores indicate stronger implicit prejudice. We estimated reliability of the IAT scores via a Spearman–Brown adjusted Pearson correlation, averaged over 10,000 replications of permutated split-halves, stratified by IAT blocks (Pronk, Molenaar, et al., 2020). The IAT scores formed a reliable scale ($r_{sb} = 0.79$).

Explicit Prejudice

Explicit prejudice towards Moroccans was measured with two items asking participants to indicate how coldly—warmly they felt towards *Moroccans* and *Other people* (*0= extremely cold, 100 = extremely warm*). Warm feelings towards Moroccans were subtracted from warm feelings towards others. Scores > 50 indicate relatively prejudiced (cold) feelings towards Moroccans, while scores < 50 indicate relatively positive (warm) feelings towards Moroccans.

Acceptance of IAT Feedback

Participants in the exposure condition were asked to indicate to what extent they "agreed with the interpretation" and "recognized" themselves in the presented IAT score (*1 = disagree completely, 7 = agree completely*). The items form a reliable scale (Cronbach's α=.91) and were averaged for an overall index of acceptance of IAT feedback. Summary statistics can be found in Supplementary Table B1, Appendix B.

Analyses

Linear regression models are used to estimate positive and negative news selection as a function of exposure to IAT feedback, implicit and explicit prejudice (H1, H2, RQ1 and RQ2). To estimate the (interaction) effects of acceptance of IAT feedback on news selection (RQ3), analyses are performed using only participants that were exposed to IAT feedback.

Results of Study 1

Table 1 displays the linear regression models predicting selection likelihood of negative (Models I) and positive (Models II) news as a function of exposure (vs. no exposure) to IAT feedback, implicit and explicit prejudice. We discuss the results related to the predictive value of implicit (H1) and explicit (H2) prejudice with regards to news selection: More specifically, we expected that both constructs would (a) positively predict the selection of negative news; and (b) negatively predict the selection of positive news about minorities. The results presented in Model I-2 show that both implicit and explicit prejudice significantly and positively predicts the selection of negative news. The results in Model II-2 indicate that explicit, but not implicit, prejudice negatively predicts the selection of positive news. Consequently, we reject H1b, and accept H1a and H2ab.

We now discuss our primary research question (RQ1): *Does providing feedback on the strength of an individual's implicit bias elicit a beneficial (dampening) effect on the predictive power of implicit biases on news selection?* Model I-1 and Model II-1 of Table 1 indicate that exposure to IAT feedback in itself does not affect selection behavior. Supplementary Table C1, Appendix C compares the effects of the different subgroups within the exposure to feedback condition (i.e., individuals exposed to feedback indicating no, moderate, slight, and strong prejudice). These results confirm that individual subgroups do not significantly differ from the no-exposure to feedback group, indicating that, across the board, exposure to IAT feedback does not affect selection behavior. In addition, the two-way interaction between implicit prejudice and exposure to IAT feedback is not significant. Hence, results suggest that across-the-board mere exposure to one's implicit prejudice is an insufficient incentive to dampen the negative influence of implicit prejudice on news selection choices.

Next, we asked whether explicit prejudice moderates the interaction between implicit prejudice and exposure to IAT feedback (RQ2). The results presented in Table 1, Model I-3 indicate that this three-way interaction is not significant ($p = 0.07$).

Last, we asked whether the effect of receiving feedback on news selection likelihood depends on the extent to which people explicitly accept such feedback (RQ3). We test this hypothesis among the sample of participants that were exposed to their implicit prejudice scores, as presented in Table 2. Model I-1 displays a significant and positive effect of acceptance of IAT feedback on the selection of negative news, indicating that accepting individuals are more likely to self-select into negative news. In addition, Model I-3 displays a significant two-way interaction of implicit prejudice and acceptance of IAT feedback on the selection of negative news. Figure 1 visualizes this interaction effect. As can be seen, the positive effect of implicit prejudice on the selection of negative news is more pronounced for individuals that accept rather than reject their IAT feedback. These results suggest that high levels of acceptance foster favorable news selection for individuals receiving IAT-feedback indicating weak (rather than strong) implicit prejudice. Conversely, low levels of acceptance dampen the undesirable influence of implicit prejudice on news selection among individuals with strong (rather than weak) implicit prejudice. We conclude that indeed, the effect of exposure to IAT feedback depends on the extent to which individuals accept this feedback.

Study 2

Study 1 revealed that the effects of exposure to IAT-feedback on news selection are driven by acceptance of this feedback, in such a way that implicitly biased individuals will self-select into more auspicious news *only* if they accept their IAT-feedback. The main aim of Study 2 is to replicate Study 1's findings in a different intergroup context. Additionally, and in response to Study 1's finding that providing IAT-feedback indicating implicit prejudice can backfire for endorsing individuals, Study 2 aims to explore the need for more explicit NML instructions educating participants on the role of media selection in implicit prejudice activation. More specifically, Study 2 investigates the effects of an integrated NML intervention, consisting of personalized feedback to implicit bias and an NML *message*. Empirical evidence suggests that in addition to *awareness*, people should receive *strategies* on how to mitigate or diminish implicit prejudice (Carnes et al., 2016). The type of information provided with IAT feedback scores is, in fact, crucial regarding the extent that people are planning to change their future behavior (Scaife et al., 2020). NML messages can offer explicit strategies on how to weaken or avoid the activation of implicit bias, which may help and motivate audience members to adopt bias reducing behavioral change (i.e., self-select into more auspicious news about outgroup members).

Method

In the second experiment, news about religious outgroups took central place. We focused on Islam, as one of the primary outgroup religions in the Netherlands. The design concerned a single factor (no exposure to IAT feedback *vs.* exposure to IAT feedback vs. exposure to IAT feedback + NML message) between-subjects design. A total of 583 participants started the experiment. Participants that failed the first ($n = 103$) or second ($n = 115$) attention check, did not fully complete the survey ($n = 51$), or for whom no valid D_{2SD}-scores could be calculated ($n = 23$) were removed from the sample. Fourteen participants that self-identified as Muslim were removed from the sample. A total of 277 participants attentively and successfully completed the second experiment (completion rate: 47.5%). Participants were on average 40.97 years old ($SD = 13.62$) of which 148 (53.4%) were female. A 16.6% of participants were lower educated, 43.7% were higher educated and 39.7% had a moderate level of education.

Largely the same procedure as Study 1 was followed. This time, one additional condition was added: exposure to implicit prejudice paired with a literacy message. In this condition, participants were exposed to a short literacy message about the potential influence of news exposure on implicit prejudice (see Supplementary Appendix A, Table A3 for the full message). Afterwards, they received feedback on their implicit scores.

Measures

We rely largely on the same measures as reported in Study 1. A Muslim-Other People IAT was designed to measure implicit prejudice towards Muslims. Targets were manipulated using words ("Muslims" and "Other people") and photos of Arabic and Caucasian males with neutral facial expressions (Langner et al., 2010). For the attribute categories, we relied again on the "good" vs. "bad" categories. The same cut-off points and feedback messages as reported in Study 1 were used, this time tailored towards implicit prejudice of Muslims. Explicit prejudice was measured by asking participants to indicate how coldly (0)—warmly (100) they feel towards Muslims and others. Warm feelings towards Muslims were subtracted from warm feelings towards others. Implicit prejudice, acceptance of IAT-feedback (r_{sb}=0.82) were all measured as reported in the Method section of Study 1.

Results of Study 2

Table 3 displays the linear regression models predicting selection likelihood of negative (Models I) and positive (Models II) news as a function of exposure (vs. no exposure) to IAT feedback, exposure to an NML message, implicit and explicit prejudice. Again, the models show that selection is positively predicted by source bias.

We investigate the predictive value of implicit (H1) and explicit (H2) prejudice. Again, we find that implicit prejudice significantly and positively predicts the

Table 3. Predicting the Selection of Negative (Models-I) and Positive (Models- II) News about Religious Outgroups with Exposure to IAT-Feedback (Study 2)

	Selection of negative news				Selection of positive news			
	Model 1 I *b* (SE)	Model 2 I *b* (SE)	Model 3 I *b* (SE)	Model 4 I *b* (SE)	Model 1 II *b* (SE)	Model 2 II *b* (SE)	Model 3 II *b* (SE)	Model 4 II *b* (SE)
Intercept	52.05***	44.50***	42.36***	39.89***	57.25***	60.85***	60.14***	58.54***
	(2.81)	(3.26)	(3.92)	(4.70)	(2.78)	(3.39)	(4.09)	(4.95)
Exposure to IAT feedback	2.44	1.85	4.73	9.75†	−0.21	−0.01	0.95	3.46
	(3.16)	(3.00)	(4.20)	(5.17)	(3.12)	(3.12)	(4.37)	(5.45)
Implicit prejudice		6.99*	11.78*	13.62†		−1.86	−0.27	2.29
		(2.73)	(5.60)	(7.66)		(2.83)	(5.84)	(8.07)
Explicit prejudice		0.23***	0.23***	0.41†		−0.11*	−0.11*	0.00
		(0.05)	(0.05)	(0.21)		(0.05)	(0.05)	(0.22)
Exposure to NML message		−0.21	−0.33	0.11		−1.13	−1.18	−1.05
		(2.43)	(2.44)	(2.44)		(2.53)	(2.54)	(2.56)
Exposure to IAT feedback * Implicit prejudice			−6.20	−14.30†			−2.06	−6.92
			(6.33)	(8.60)			(6.59)	(9.05)
Exposure to IAT feedback * Explicit prejudice				−0.33				−0.17
				(0.22)				(0.24)
Implicit prejudice * Explicit prejudice				−0.16				−0.15
				(0.27)				(0.28)

(*Continued*)

Table 3 (continued)

	Selection of negative news				Selection of positive news			
	Model 1 I *b* (SE)	Model 2 I *b* (SE)	Model 3 I *b* (SE)	Model 4 I *b* (SE)	Model 1 II *b* (SE)	Model 2 II *b* (SE)	Model 3 II *b* (SE)	Model 4 II *b* (SE)
Exposure to IAT feedback * Implicit prejudice * Explicit prejudice				0.46 (0.29)				0.26 (0.31)
Log-likelihood	−1238.6	−1221.6	−1221.2	−1217.5	−1235.6	−1232.5	−1232.4	−1231.9

Note. $^{\dagger}p < .1$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

selection of negative news, but not positive news. Explicit prejudice positively predicts the selection of negative news, and negatively predicts the selection of positive news. These findings fully replicate the results of Study 1: We reject H1b, and accept H1a, H2ab.

Next, we investigate the interaction effect of exposure to IAT feedback and implicit prejudice scores (RQ1). Model 1-I and Model 1-II in Table 3 replicate Study 1's finding that exposure to implicit scores in itself is not enough to cause a change in selection behavior. In addition, Table 3, Model 2-I and Model 2-II indicate that the effect of exposure to IAT feedback on news selection is not dependent upon implicit prejudice scores (RQ1). Likewise, and in relation to RQ2, we do not find a significant three-way interaction effect between exposure to IAT feedback, implicit prejudice, and explicit prejudice.

Next, we investigate the interaction effect of acceptance of IAT feedback and implicit prejudice (RQ3). Table 4 models the effect of acceptance of IAT feedback on news selection patterns. Model 3-II shows that the effect of IAT feedback on the selection of *positive* news depends upon existing levels of implicit bias and overtly expressed acceptance of the IAT feedback scores (RQ3). Figure 2 for a visualization of the interaction effect: the negative effect of implicit prejudice on the selection of positive news dampened for individuals that reject their IAT feedback. Conversely, the negative influence of implicit prejudice on the selection of positive news is more pronounced for individuals that accept their feedback.

Last, we test whether an NML-message, presented with individualized IAT-feedback, will trigger more favorable news selection behavior. The models in Table 3 and Table 4 include exposure to the NML message as a predictor of news selection likelihood. The results indicate that exposure to the NML message did not motivate news users to change their news selection decisions.

Combined Sample Effects

To further investigate the three-way interaction between exposure to IAT feedback, implicit prejudice, and explicit prejudice (RQ2), we re-ran our analyses using the combined sample of Study 1 and Study 2 (see Supplementary Appendix D, Table D1). We do so as our analyses might have been underpowered: In both studies, the group that did not receive feedback was relatively small (Study 1: $n = 57$, Study 2: $n = 57$). We included intergroup domain as a covariate. Results (Supplementary Appendix D, Table D1) reveal a significant three-way interaction effect of exposure to IAT feedback, implicit prejudice, and explicit prejudice on the selection of negative news ($b = 0.51$, $SE = 0.21$, $p < .5$). For people with *weak* explicit prejudice, IAT feedback dampened the negative influence of implicit prejudice on negative news selection. This is *not* the case for people with *strong* levels of explicit prejudice: For those, exposure to IAT feedback does not alter the influence of implicit prejudice on selection (see Figure 3).

Table 4. Predicting the Selection of Negative (Models-I) and Positive (Models- II) News about Religious Outgroups with Acceptance of IAT Feedback Scores (Study 2)

	Selection of negative news			Selection of positive news		
	Model 1 I *b* (SE)	Model 2 I *b* (SE)	Model 3 I *b* (SE)	Model 4 I *b* (SE)	Model 1 II *b* (SE)	Model 2 II *b* (SE)
Intercept	48.83***	43.01***	46.90***	52.28***	61.48***	49.66***
	(3.98)	(4.80)	(7.34)	(4.43)	(4.99)	(7.56)
Acceptance of IAT feedback	1.25	0.83	0.17	0.59	−0.21	1.80
	(0.82)	(0.87)	(1.28)	(0.90)	(0.90)	(1.32)
Implicit prejudice		6.58*	−0.35		−2.65	18.41^{+}
		(3.34)	(10.45)		(3.47)	(10.76)
Explicit prejudice		0.20***	0.19**		−0.11^{+}	−0.07
		(0.06)	(0.06)		(0.06)	(0.06)
Exposure to NML message		0.88	0.76		0.16	0.54
		(2.82)	(2.82)		(2.92)	(2.91)
Acceptance of IAT feedback * Implicit prejudice			1.37			−4.16*
			(1.96)			(2.01)
Log-likelihood	−1019.1	−976.58	−976.33	−1020.7	−984.96	−982.79

Note. $^{+}p < .1$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. The presented findings are tested among the sample of participants that were exposed to their implicit biases: Participants in the non-exposure condition are not included. $N = 229$.

Discussion

To understand the foundations of biased news selection patterns, scholars have recognized that in addition to overtly expressed 'explicit' attitudes, automatically activated 'implicit' processes are important determinants of media choice (Arendt et al., 2016, 2019; Kroon et al., 2020). As such implicit forces operate largely under the radar of conscious awareness, they meaningfully determine what people decide to read, like, or share—without news users' appreciation of such effects. This study was first to ask whether opening news users' eyes to personal levels of implicit bias can help them to change their media diets by making more informed news selection decisions. Concretely, we tested whether providing audience members with feedback on individual-level IAT scores may be a fruitful feature of media literacy interventions aim to alter news consumers' news decisions. We put our hypotheses to the test across two domains of intergroup bias (i.e., ethnic and religious prejudice) using data from two experiments conducted among native Dutch citizens.

The presented results further replicate previous findings that implicit measures explain news choice in addition to explicit measures (Arendt et al., 2016, 2019;

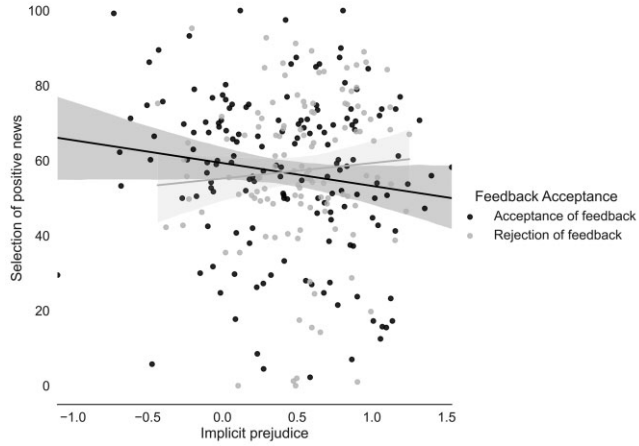

Figure 2 The two-way interaction effects of acceptance of IAT feedback and implicit prejudice on the selection of positive news about religious outgroups (Study 2).

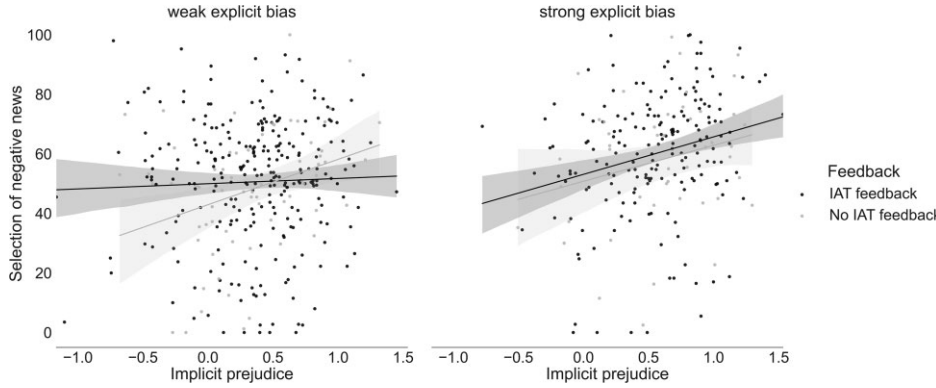

Figure 3 The three-way interaction effects of exposure to IAT feedback, implicit prejudice, and explicit prejudice on the selection of negative news using the combined sample of Study 1 and Study 2 ($N = 563$).

Kroon et al., 2020). Our findings further refine this conclusion by adding that implicit prejudice serves as a significant and important predictor for the selection of negative news, but *not* for positive news. This implies that valence of news serves as a *contingent* moderator (Holbert & Park, 2019) of the relation between implicit prejudice and selection likelihood.

Of key interest to the current study, we found that across-the-board, exposure to IAT-feedback does not change news users' selection of negative and positive news about ethnic and religious outgroups. Rather, a primary conclusion of this study is that exposure effects are contingent upon both implicit prejudice and overtly-expressed attitudes. More specifically, Study 1's findings hinted towards a three-way

interaction effect between exposure to IAT feedback, implicit prejudice, and explicit prejudice on the selection of negative news. Using the combined sample of Study 1 and Study 2, we find empirical support for this interaction. Substantively, these findings indicate that exposure to IAT feedback dampens the influence of implicit prejudice on the selection of negative news for individuals with weak levels of explicit prejudice, but not for those with strong levels of explicit prejudice.

In addition, we found that the influence of implicit prejudice and overly expressed acceptance of the IAT feedback scores interact in predicting selection patterns. More specifically, acceptance fostered favorable news selection for individuals receiving IAT-feedback indicating *weak* (rather than strong) implicit prejudice. Conversely, *low* levels of acceptance dampened the undesirable influence of implicit prejudice on news selection among individuals with *strong* (rather than weak) implicit prejudice.

Based on these empirical findings, our results indicate that a tailored NML intervention using individualized IATs scores may materialize into more favorable news selection patterns for the following groups. First, for individuals with *strong* implicit, but *weak* explicit prejudice. For these individuals, the intervention creates the awareness-inducing effect that was intended; they became aware of their unconscious biases and made favorable adjustments to their selection behavior accordingly. Second, intervention outcomes are favorable for individuals with *weak* (or non-existing) implicit prejudice that explicitly *accept* their implicit feedback scores. For these individuals, the confirmation that they do *not* harbor implicit prejudices further reinforces their appropriate selection behavior by basically stimulating and praising them with an acknowledgement of their prejudice-free attitude. For these type of news users, the presented findings give reason for optimism regarding the effectiveness of NML interventions to successfully break the self-fueling cycle of biased news selectivity and effects (Kroon et al., 2020; Schemer, 2012) by helping them to successfully withstand the guiding influence of implicit drivers of selection biases.

On the contrary, the findings indicate that IAT-feedback should not be used in a tailored NML intervention for individuals with *strong* levels of *explicit and implicit* prejudice. Likewise, such an intervention is not recommended for individuals that explicitly accept feedback indicating strong implicit prejudice. For these individuals, exposure to IAT feedback scores did not alter selection patterns. Apparently, those individuals did not become discouraged, but might have felt reinforced in their prejudiced self-image and aligned their selection behavior accordingly. This finding can be related to previous research that found that uncongenial information in NML messages may trigger feelings of reactance which, in turn, result in the intervention having the opposite effect instead of the desired outcome (Van der Meer & Hameleers, 2020). In sum, our findings suggests that an important group of individuals that may gain most from receiving implicit bias feedback are also among the ones least likely to benefit from it, echoing previous findings (Howell et al., 2015).

Finally, and motivated by the insight that in addition to fostering awareness individuals should receive strategies on how to mitigate implicit prejudice (Carnes et al., 2016), we tested a NML message providing audience members with advice regarding the selection of news messages. Exposure to this message did not cause individuals to change their behavior above and beyond the effect of exposure to the IAT feedback. Apparently, a 'one-message-fits-all' solution did not provide additional benefits over mere exposure to individualized levels of implicit prejudice.

Together, the findings indicate that feedback to implicit prejudice should not be used as an across-the-board intervention tool, but rather be implemented carefully and on the basis of news users' levels of implicit prejudice and overtly-expressed attitudes, in particular explicit prejudice and the acceptance of unconscious bias and motivation for corrective action. Previous research has shown that further tailoring NML interventions to individuals' prior attitudes can be effective in fostering more healthy news selection behavior while avoiding boomerang effects (Van der Meer & Hameleers, 2020). Future research may further investigate the effects of tailoring how IAT feedback is delivered (see Scaife et al., 2020) both in isolation and in sync with NML messages providing strategies for overcoming prejudice.

Limitations

The current study is not without limitations. First, the IAT has been shown to be valid at the group-level, yet it should be acknowledged that the accuracy of IAT scores at the individual level has been criticized (Connor & Evers, 2020). In addition, some participants demonstrated *favorable* outgroup prejudice (i.e., *D*-scores < -0.15), but received a message stating that they had "no bias." Consequently, participants in this study may have been exposed to implicit prejudice scores that did not accurately represent their unconscious prejudices. It is vital that future studies and intervention programs seriously consider the accuracy of the IAT and the ethical implications of exposing individuals to implicit bias scores that might be noisy. In addition, in both studies, only a relatively small group of participants was not exposed to their IAT scores. This approach limited the power of the analysis, especially as effect-sizes of implicit measures tend to be small (Greenwald et al., 2015). In this study, we have therefore re-run the analyses on the combined sample of Study 1 and Study 2. Further, as *all* participants completed an IAT, the current study lacked a solid news preference baseline, as just completing an IAT might have primed participants. Together, future research should further validate the here-reported findings using larger samples, and by including a control group that does not complete an IAT at all.

Moreover, even though our operationalization of news headlines and leads was based on extensive pretesting, this study did not measure news selection in a high-choice, every-day media use context. This is a serious limitation that should be addressed in future research. Finally, the generalizability of the here-reported findings, as well as their applicability to actual interventions, should be discussed. The

current study investigated short-term effects. Yet, to successfully change news consumption patterns, it is important that feedback effects are sustainable and effective over time. To translate the here-tested implicit bias training into a real-world application, it is important to explore its long-term consequences for news selection decisions.

Scholars have argued that unconscious processes are especially powerful in guiding our news selection choices, ultimately affecting what type of news we are exposed to. Although the foundations of biases in news selection are well-established, far less is known about how we can actually change news selection choices. By testing a promising intervention tool, this study aims to further stimulate the scientific discussion about how 'healthier' news selection can be stimulated. Our findings show that raising awareness of implicit prejudice works as an effective strategy for corrective action for some, but may backfire for others, and should therefore only be implemented with caution and attention for explicit considerations.

Supplementary material

Supplementary material is available online at *Human Communication Research* (http://mtp.oxfordjournals.org/)

Notes

1. In line with the recommendations for the improved D_{2SD}-scores, we eliminated subjects for whom more than 10% of trials have latencies less than 300 ms.
2. JASMIN ensured accurate timing by leveraging modern web technologies such as requestAnimationFrame and a high-resolution timer. Since the IAT *D*-score is based on RT differences within task conditions, the impact of noisy RT measurements, as is common in any web application, had relatively little impact on the reliability with which individual differences were measured (see Pronk, Wiers, et al., 2020).
3. We have grouped together participants with strong favorable outgroup bias, moderate favorable outgroup bias, slight favorable outgroup bias and no bias, because we are substantively interested in *unfavorable outgroup* bias, and because we expected to have only few observations in the remaining categories. As can be seen in Table B1 and Figure B1, this expectation was supported by our data: D_{2SD}-scores for most participants that were exposed to the "no bias" message, ranged between 0.15 and −0.15. Yet, some individuals in this category demonstrated slight to strong *favorable* outgroup biases.

References

Ahmed, S., & Matthes, J. (2017). Media representation of Muslims and Islam from 2000 to 2015: A meta-analysis. *International Communication Gazette*, *79*(3), 219–244. https://doi.org/10.1177/1748048516656305

Appiah, O., Knobloch-Westerwick, S., & Alter, S. (2013). Ingroup favoritism and outgroup derogation: Effects of news valence, character race, and recipient race on selective news reading. *Journal of Communication*, *63*(3), 517–534. https://doi.org/10.1111/jcom.12032

Applebaum, B. (2019). Remediating campus climate: Implicit bias training is not enough. *Studies in Philosophy and Education*, *38*(2), 129–141. https://doi.org/10.1007/s11217-018-9644-1

Arendt, F. (2013). Dose-dependent media priming effects of stereotypic newspaper articles on implicit and explicit stereotypes. *Journal of Communication*, *63*, 830–851. https://doi.org/10.1111/jcom.12056

Arendt, F., & Karadas, N. (2020). Implicit and explicit attitudes toward Germany as news-choice predictors among Muslims with migration backgrounds living in Germany. *Communications*, *45*(4), 440–462. https://doi.org/10.1515/commun-2019-2067

Arendt, F., Northup, T., & Camaj, L. (2019). Selective exposure and news media brands: Implicit and explicit attitudes as predictors of news choice. *Media Psychology*, *22*(3), 526–543. https://doi.org/10.1080/15213269.2017.1338963

Arendt, F., Steindl, N., & Kümpel, A. (2016). Implicit and explicit attitudes as predictors of gatekeeping, selective exposure, and news sharing: Testing a general model of media-related selection. *Journal of Communication*, *66*(5), 717–740. https://doi.org/10.1111/jcom.12256

Atewologun, D., Cornish, T., & Tresh, F. (2018). Unconscious bias training: An assessment of the evidence for effectiveness. Research report 113. In *Equality and Human Rights Commission Research Report Series*.

Betz, H. G. (2013). Mosques, minarets, burqas and other essential threats: The populist right's campaign against Islam in Western Europe. In B. M. R. Wodak & M. KhosraviNik (Ed.), *Right-wing populism in Europe: Politics and discourse* (pp. 71–88). London: Bloomsbury Academic

Carnes, M., Middleton, W. S., Veterans, M., Devine, P. G., Manwell, L. B., Office, C., … Kaatz, A. (2016). The effect of an intervention to break the gender bias habit for faculty at one institution: A cluster randomized, controlled trial. *Academic Medicine*, *90*(2), 221–230. https://doi.org/10.1097/ACM.0000000000000552

Connor, P., & Evers, E. R. K. (2020). The bias of individuals (in crowds): Why implicit bias is probably a noisily measured individual-level construct. *Perspectives on Psychological Science*, *15*(6), 1329–1345. https://doi.org/10.1177/1745691620931492

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*(6), 1267–1278. https://doi.org/10.1016/j.jesp.2012.06.003

Dovidio, J. F., Gaertner, S. L., & Pearson, A. R. (2016). Aversive Racism and Contemporary Bias. *The Cambridge Handbook of the Psychology of Prejudice*, 267–294. https://doi.org/10.1017/9781316161579.012

Ekman, M. (2019). Anti-immigration and racist discourse in social media. *European Journal of Communication*, *34*(6), 606–618. https://doi.org/10.1177/0267323119886151

Gaertner, S. L., & Dovidio, J. F. (2005). Understanding and addressing contemporary racism: From aversive racism to the common ingroup identity model. *Journal of Social Issues*, *3*, 615–639. https://doi.org/10.1111/j.1540-4560.2005.00424.x

Galdi, S., Gawronski, B., Arcuri, L., & Friese, M. (2012). Selective exposure in decided and undecided individuals: Differential relations to automatic associations and conscious

beliefs. *Personality and Social Psychology Bulletin*, *38*(5), 559–569. https://doi.org/10.1177/0146167211435981

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*(4), 553–561. https://doi.org/10.1037/pspa0000016

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology*, *116*(5), 769–794. https://doi.org/10.1037/pspi0000155

Harwood, J. (1999). Age identification, social identity gratifications, and television viewing. *Journal of Broadcasting & Electronic Media*, *43*(1), 123–136. https://doi.org/10.1080/08838159909364479

Holbert, R. L., & Park, E. (2019). Conceptualizing, organizing, and positing moderation in communication research. *Communication Theory*, *00*, 1–20. https://doi.org/10.1093/ct/qtz006

Howell, J. L., Gaither, S. E., & Ratliff, K. A. (2015). Caught in the middle: Defensive responses to IAT feedback among whites, blacks, and biracial black/whites. *Social Psychological and Personality Science*, *6*(4), 373–381. https://doi.org/10.1177/1948550614561127

Howell, J. L., & Ratliff, K. A. (2017). Not your average bigot: The better-than-average effect and defensive responding to Implicit Association Test feedback. *British Journal of Social Psychology*, *56*(1), 125–145. https://doi.org/10.1111/bjso.12168

Jackson, S. M., Hillard, A. L., & Schneider, T. R. (2014). Using implicit bias training to improve attitudes toward women in STEM. *Social Psychology of Education*, *17*(3), 419–438. https://doi.org/10.1007/s11218-014-9259-5

Jeong, S. H., Cho, H., & Hwang, Y. (2012). Media literacy interventions: A meta-analytic review. *Journal of Communication*, *62*(3), 454–472. https://doi.org/10.1111/j.1460-2466.2012.01643.x

Kalra, P., & Boukes, M. (2020). Curbing journalistic gender bias: How activating awareness of gender bias in Indian journalists affects their reporting. *Journalism Practice*, *15*(5), 651–668. https://doi.org/10.1080/17512786.2020.1755344

Knobloch-Westerwick, S. (2015). *Choice and preference in media use: Advances in selective exposure theory and research*. New York: Routledge.

Knobloch-Westerwick, S., & Hastall, M. R. (2010). Please yourself: Social identity effects on selective exposure to news about in- and out-groups. *Journal of Communication*, *60*(3), 515–535. https://doi.org/10.1111/j.1460-2466.2010.01495.x

Knobloch-Westerwick, S., Mothes, C., & Polavin, N. (2020). Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. *Communication Research*, *47*(1), 104–124. https://doi.org/10.1177/0093650217719596

Kroon, A. C., van der Meer, T. G. L. A., & Mastro, D. (2020). Confirming bias without knowing? Automatic pathways between media exposure and selectivity. *Communication Research*. https://doi.org/10.1177/0093650220905948

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud faces database. *Cognition and Emotion*, *24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076

Lee, C. (2017). Awareness as a first step toward overcoming implicit bias. *SSRN Electronic Journal*, *289*. https://doi.org/10.2139/ssrn.3011381

Livingstone, S. (2004). Media literacy and the challenge of new information and communication technologies. *Communication Review*, *7*(1), 3–14. https://doi.org/10.1080/10714420490280152

Mastro, D. (2015). Why the media's role in issues of race and ethnicity should be in the spotlight. *Journal of Social Issues*, *71*(1), 1–16. https://doi.org/10.1111/josi.12093

Monteith, J. M., Sherman, J. W., & Devine, P. G. (1998). Suppression as a stereotype control strategy. *Personality and Social Psychology Review*, *2*(1), 63–82. https://doi.org/10.1207/s15327957pspr0201_4

Nathanson, A. L. (2002). The unintended effects of parental mediation of television on adolescents. *Media Psychology*, 3(*4*), 207–230. https://doi.org/10.1207/S1532785XMEP0403_01

Nosek, B. A., Graham, J., & Hawkins, C. B. (2010). *Implicit political cognition*. In Gawronski B. & Payne B. K. (Ed.), Handbook of implicit social cognition: Measurement, theory, and applications (pp. 548–564). The Guilford Press.

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in Western Europe. *European Journal of Social Psychology*, *25*(1), 57–75. https://doi.org/10.1002/ejsp.2420250106

Pritlove, C., Juando-Prats, C., Ala-leppilampi, K., & Parsons, J. A. (2019). The good, the bad, and the ugly of implicit bias. *The Lancet*, *393*(10171), 502–504. https://doi.org/10.1016/S0140-6736(18)32267-0

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *43*(4), 565–578. https://doi.org/10.1016/j.jesp.2006.05.011

Pronk, T. (2014). *JASMIN; a library for cognitive tasks in JavaScript (Version 2.1.11)*. https://github.com/tpronk/JASMIN

Pronk, T., Molenaar, D., Wiers, R., & Murre, J. M. J. (2020). *Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment*. https://doi.org/10.31234/osf.io/ywste

Pronk, T., Wiers, R. W., Molenkamp, B., & Murre, J. (2020). Mental chronometry in the pocket? Timing accuracy of web applications on touchscreen and keyboard devices. *Behavior Research Methods*, *52*(3), 1371–1382. https://doi.org/10.3758/s13428-019-01321-2

Ramasubramanian, S. (2007). Media-based strategies to reduce racial stereotypes activated by news stories. *Journalism & Mass Communication Quarterly*, *84*(2), 249–264. https://doi.org/10.1177/107769900708400204

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, *35*(3), 213–229. https://doi.org/10.1177/0267323120922066

Saucier, D. A. (2002). Self-reports of racist attitudes for oneself and for others. *Psychologica Belgica*, *42*(1–2), 99–105. https://doi.org/10.5334/pb.987

Scaife, R., Stafford, T., Bunge, A., & Holroyd, J. (2020). To blame? The effects of moralized feedback on implicit racial bias. *Collabra: Psychology*, *6*(1), 30. https://doi.org/10.1525/collabra.251.s1

Schemer, C. (2012). Reinforcing spirals of negative affects and selective attention to advertising in a political campaign. *Communication Research*, *39*(3), 413–434. https://doi.org/10.1177/0093650211427141

Schemer, C. (2012). The influence of news media on stereotypic attitudes toward immigrants in a political campaign. *Journal of Communication*, *62*(5), 739–757. https://doi.org/10.1111/j.1460-2466.2012.01672.x

Schieferdecker, D., & Wessler, H. (2017). Bridging segregation via media exposure? Ingroup identification, outgroup distance, and low direct contact reduce outgroup appearance in media repertoires. *Journal of Communication*, *67*, 993–1014. https://doi.org/10.1111/jcom.12338

Shoemaker, P., & Vos, T. (2009). *Gatekeeping theory*. New York, NY: Routledge.

Soroka, S. N., & McAdams, S. (2015). News, politics, and negativity. *Political Communication*, *32*(1), 1–22. https://doi.org/10.1080/10584609.2014.881942

Steinke, J., Lapinski, M. K., & Crocker, N. (2007). Assessing media influences on middle school-aged children's perceptions of women in science using the draw-a-scientist test (dast). *Science Communication*, *29*(1), 35–64. https://doi.org/10.1177/1075547007306508

Sukhera, J., Milne, A., Teunissen, P. W., Lingard, L., & Watling, C. (2018). The actual versus idealized self: Exploring responses to feedback about implicit bias in health professionals. *Academic Medicine*, *93*(4), 623–629. https://doi.org/10.1097/ACM.0000000000002006

Sukhera, J., Wodzinski, M., Milne, A., Teunissen, P. W., Lingard, L., & Watling, C. (2019). Implicit bias and the feedback paradox: Exploring how health professionals engage with feedback while questioning its credibility. *Academic Medicine: Journal of the Association of American Medical Colleges*, *94*(8), 1204–1210. https://doi.org/10.1097/ACM.0000000000002782

Tajfel, H., & Turner, J. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The Social Psychology of Intergroup Relations* (pp. 33–47). Monterey, CA: Brooks/Cole

Tully, M., Vraga, E. K., & Bode, L. (2020). Designing and testing news literacy messages for social media. *Mass Communication and Society*, *23*(1), 22–46. https://doi.org/10.1080/15205436.2019.1604970

Van der Meer, T. G. L. A., & Hameleers, M. (2020). Fighting biased news diets: Using news media literacy interventions to stimulate online cross-cutting media exposure patterns. *New Media and Society*. https://doi.org/10.1177/1461444820946455

Vraga, E. K., & Tully, M. (2019a). Engaging with the other side: Using news media literacy messages to reduce selective exposure and avoidance. *Journal of Information Technology and Politics*, *16*(1), 77–86. https://doi.org/10.1080/19331681.2019.1572565

Vraga, E. K., & Tully, M. (2019b). News literacy, social media behaviors, and skepticism toward information on social media. *Information, Communication & Society*, *24*(2), 150–166. https://doi.org/10.1080/1369118x.2019.1637445

West, K., & Eaton, A. A. (2019). Prejudiced and unaware of it: Evidence for the Dunning-Kruger model in the domains of racism and sexism. *Personality and Individual Differences*, *146*, 111–119. https://doi.org/10.1016/j.paid.2019.03.047