nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-021-01253-y

When seeking help, women and racial/ethnic minorities benefit from explicitly stating their identity

Erika L. Kirgios[1]✉, Aneesh Rai[1], Edward H. Chang[2] and Katherine L. Milkman[1]

Receiving help can make or break a career, but women and racial/ethnic minorities do not always receive the support they seek. Across two audit experiments—one with politicians and another with students—as well as an online experiment (total *n* = 5,145), we test whether women and racial/ethnic minorities benefit from explicitly mentioning their demographic identity in requests for help, for example, by including statements like "As a Black woman…" in their communications. We propose that when a help seeker highlights their marginalized identity, it may activate prospective helpers' motivations to avoid prejudiced reactions and increase their willingness to provide support. Here we show that when women and racial/ethnic minorities explicitly mentioned their demographic identity in help-seeking emails, politicians and students responded 24.4% (7.42 percentage points) and 79.6% (2.73 percentage points) more often, respectively. These findings suggest that deliberately mentioning identity in requests for help can improve outcomes for women and racial/ethnic minorities.

In the United States, women and racial/ethnic minorities remain underrepresented in many organizational contexts, particularly in leadership positions[1,2]. One contributing factor may be that in-group favouritism and bias lead underrepresented group members to receive less instrumental help—advice, feedback, referrals or assistance on tasks—than White men[3–12]. Such instrumental help can be critical to career success, especially for members of historically marginalized groups[13–15]. Thus, increasing the rate at which assistance is offered to women and racial/ethnic minorities might be one way to reduce identity-based inequities.

When people from marginalized groups seek help or, more generally, pursue career advancement, past research suggests that they often face discrimination if decision-makers can infer their identity from cues like names, photographs or extracurricular activities[3,9,11,16–19]. For instance, Bertrand and Mullainathan randomly assigned White-sounding or Black-sounding names to otherwise identical resumes and used those resumes to apply for entry-level jobs[20]. They found that those with Black-sounding names received 50% fewer callbacks than those with White-sounding names[20]. People may be particularly likely to discriminate based on identity when deciding how to respond to requests for help. The process of deciding whether to help someone can be ambiguous and unstructured, and discrimination is more likely to arise in ambiguous contexts[9,21]. Together, these findings suggest that marginalized group members might be wise to downplay or even hide their demographic identity when seeking help[19].

We propose, however, that women and racial/ethnic minorities may benefit from explicitly stating their demographic identity in help requests. When help seekers highlight their marginalized demographic identity, prospective helpers may worry that a failure to respond could amount to discrimination. That is, explicitly mentioning identity makes it salient to prospective helpers that prejudice could affect their decisions. To avoid feeling or appearing prejudiced, prospective helpers may then be more likely to offer their assistance. Indeed, research shows that people have both internal and external motivations to reduce their expression of prejudice[22]. Specifically, people seek to avoid actions that they or others could interpret as discriminatory to (1) maintain a positive self-image (by behaving consistently with their personal values), and (2) escape social sanctioning (by conforming to norms of political correctness or egalitarianism)[22–27]. So, when someone asking for help calls attention to the potential for discrimination by explicitly highlighting their marginalized identity, we theorize that prospective helpers' internal and external motivations to respond without prejudice will be activated and will increase the likelihood that prospective helpers provide instrumental support.

Prior research suggests that when the potential for prejudice is more salient, people are less likely to behave in a biased manner[26,28,29]. For example, following media coverage of a study demonstrating that White National Basketball Association referees tended to be biased in favour of White players, this in-group bias declined significantly[28]. Making referees aware of the potential for bias may have helped them counteract it. Similarly, Sommers and Ellsworth found that when mock juries evaluated cases, White jurors were generally biased against Black defendants[29]. However, this bias was eliminated when the case involved a racially charged incident, suggesting that when racial prejudice was salient to decision-makers, they made less biased decisions. This evidence indicates that, at least in some cases, people make less prejudiced decisions when they are given cause for concern that prejudice might affect their choices.

In this paper, we examine whether women and racial/ethnic minorities are more likely to receive instrumental help when they explicitly mention their demographic identity in a request for support. For instance, a woman asking for a referral to a technology company might highlight her gender by saying, "As a woman in tech, I would be grateful for your referral." We propose and find that the inclusion of such statements in help requests increases the likelihood that women and racial/ethnic minorities receive the support they seek.

[1]Department of Operations, Information and Decisions, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Negotiation, Organizations & Markets, Harvard Business School, Harvard University, Boston, MA, USA. ✉e-mail: ekirgios@wharton.upenn.edu

We present results from two field experiments and one online experiment demonstrating this effect. First, in a preregistered audit experiment with 2,476 city council members from across the United States, we show that city councillors are a regression-estimated 7.42 percentage points (or 24.4%) more likely to respond to help-seeking emails from women and racial/ethnic minorities when the sender explicitly mentions their demographic identity. In a second audit experiment with 1,169 undergraduates at a large northeastern university, we replicate our key finding. Specifically, we demonstrate that undergraduates are a regression-estimated 2.73 percentage points (or 79.6%) more likely to volunteer to help a Black male graduate student when his request for help includes an explicit reference to his demographic identity. Finally, in a preregistered online experiment with 1,500 participants, we find that internal motivation to respond without prejudice is associated with prospective helpers' increased responsiveness to requests for assistance when help seekers mention their marginalized identities.

Our work suggests that when someone explicitly mentions their marginalized demographic identity in a request for help, it elicits a different reaction than inadvertently conveying the same demographic identity (for example, via a Black-sounding name). Past work indicates that whether information about an individual's identity is conveyed deliberately or inadvertently, it activates stereotypes, which can produce discrimination[3,9,19,20,30–32]. However, we propose that, unlike information about identity conveyed inadvertently, information divulged deliberately may also draw prospective helpers' attention to the possibility for prejudice to affect their decisions. This, in turn, can increase people's concern about internal or external censure, making them more likely to help members of marginalized groups.

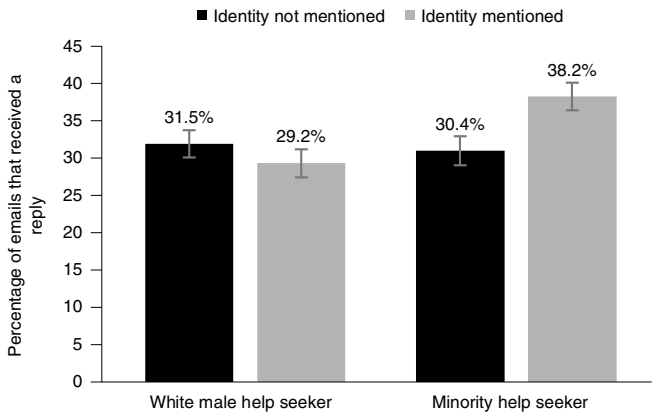

Fig. 1 | Reply rates to emails across conditions in Study 1. White male city councillors' ($n = 2,476$) response rates to help-seeking emails from fictitious students in Study 1. The two bars on the left display response rates to emails from help-seeking students whose names signalled that they were White men and the two bars on the right display response rates to emails from help-seeking students whose names signalled that they belonged to a marginalized identity group (that is, that they were White women, Black men, Black women, Latinos or Latinas). The black bars display response rates in the identity-not-mentioned condition and the grey bars display response rates in the identity-mentioned condition. Standard error bars are depicted around each proportion. Full regression results estimating the significance of these effects are provided in Table 1 and Supplementary Table 4 (using an OLS regression) and Supplementary Table 5 (using a logistic regression).

Results

Study 1: Audit experiment with city councillors. Participants were 2,476 White male city councillors serving in cities across the United States. Each city councillor received an email from a fictitious student requesting career advice (following a design similar to that used in Kalla et al.[12]). The emails were identical across conditions except for two randomized elements: (1) whether the help-seeking student was a White male help seeker (hereafter referred to as the White male help-seeker condition; see Supplementary Table 1 for information about the help-seekers' names which were used to manipulate identity) or a minority help seeker (that is, a White female, Black male, Black female, Latino or Latina; hereafter referred to as the minority-help-seeker condition); and (2) whether the student explicitly mentioned their identity in the email (calling themselves a 'young man/woman/Black man/Black woman/Latino/Latina'; hereafter referred to as the identity-mentioned condition) or not (instead, calling themselves a 'young person'; hereafter referred to as the identity-not-mentioned condition). Supplementary Table 2 includes participant summary statistics, and balance checks presented in Supplementary Table 3 show that we did not find imbalances across experimental conditions on any observable participant characteristics.

Our preregistered dependent variable of interest was whether a city councillor replied to our email within 1 week. Following our preregistration, automatic replies and replies from aides or assistants—as opposed to the city councillor—were counted as non-responses. As Fig. 1 shows, city councillors replied to emails from White men requesting help 31.5% of the time when the help request did not mention the sender's identity. They replied to emails from White men requesting help 29.2% of the time when the help request mentioned the sender's identity. We found no evidence of a difference in response rates to White men across the identity-not-mentioned and identity-mentioned conditions (two-sample, two-tail proportions test: $z = 0.870$, $P = 0.384$, effect size $h = -0.050$, 95% confidence interval (CI): [−0.074, 0.029]). However, city councillors replied to emails from women and racial/ethnic minorities requesting help 30.4% of the time in the identity-not-mentioned condition and 38.2% of the time in the identity-mentioned condition, a difference that was statistically significant (two-sample, two-tail proportions test: $z = 2.89$, $P = 0.004$, effect size $h = 0.164$, 95% CI [0.025, 0.130]). Fig. 2 shows the breakdown of response rates across all sender minority groups studied (White women, Black women, Black men, Latinas and Latinos).

Our preregistered main analysis was an ordinary least squares (OLS) regression with robust standard errors predicting whether city councillors replied to an email containing a request for help with the following independent variables: an indicator for assignment to the identity-mentioned condition; an indicator for assignment to the minority-help-seeker condition; and an interaction between these two indicators, along with controls for which of several slightly different email templates requesting help was sent; the city councillor's region; the city's population size; the city councillor's political party; years until the city councillor's next re-election; and the city councillor's current position (whether or not they had recently been replaced or stepped down). Complete regression results for this analysis are included in Supplementary Table 4. Given that our outcome variable is binary, our data violate both normality and homoskedasticity assumptions. Despite these violations, we analysed our data using preregistered OLS regressions because interactions cannot be estimated without bias when using logistic regressions, and OLS regressions are the recommended method for estimating treatment effects on binary outcomes in experiments[33,34]. Moreover, in Supplementary Table 5, we present the results of our primary analysis with a logistic regression rather than an OLS regression (further robustness checks presented in Supplementary Tables 6 and 7 (a) removing any city councillors who had been replaced or stepped down and (b) including replies to our emails that arrived within 7 weeks rather than only replies received within 1 week). Our Supplementary Information also contains further

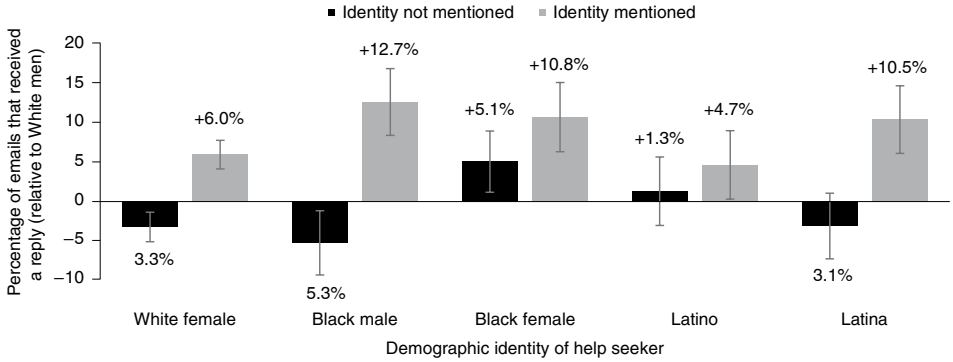

Fig. 2 | Reply rates to emails from women and/or racial/ethnic minorities (relative to White male help seekers) across conditions in Study 1. White male city councillors' ($n = 2,476$) response rates to emails from women and/or racial/ethnic minorities seeking help (relative to White men seeking help) in the identity-not-mentioned and identity-mentioned conditions. Response rates to White men were 31.5% in the identity-not-mentioned condition and 29.2% in the identity-mentioned condition. Standard error bars are depicted around each proportion. Full regression results estimating the significance of these effects are provided in Table 1 and Supplementary Table 4 (using an OLS regression) and Supplementary Table 5 (using a logistic regression).

Table 1 | Regression-estimated effects of explicitly stating your identity in a request for help in Study 1

	Model 1 outcome: responded (1 = Yes, 0 = No)			Model 2 outcome: log word count			Model 3 outcome: log character count		
	b	95% CI	*P*	*b*	95% CI	*P*	*b*	95% CI	*P*
Female and/or racial/ethnic minority help seeker	−0.010	[−0.062, 0.042]	0.705	−0.070	[−0.301, 0.162]	0.554	−0.101	[−0.416, 0.214]	0.528
Identity mentioned	−0.023	[−0.075, 0.029]	0.380	−0.114	[−0.346, 0.118]	0.327	−0.148	[−0.463, 0.168]	0.350
Female and/or racial/ethnic minority help seeker × identity mentioned	0.097	[0.024, 0.171]	0.010	0.390	[0.062, 0.717]	0.020	0.530	[0.085, 0.975]	0.020
Observations	2,476			2,476			2,476		
Adjusted R^2	0.007			0.006			0.007		

This table reports the results of six OLS regression models. The first regression model predicts whether a given city councillor in Study 1 responded to an email from a student requesting career advice (Model 1, preregistered). The final two regression models predict the length of the response a given city councillor in Study 1 provided, as measured by either the log word count of the response (Model 2) or the log character count of the response (Model 3). All models show the main effects of assignment to the minority-help-seeker condition, assignment to the identity-mentioned condition and the interaction between these two variables. The models also include the following controls: fixed effects for which email variant a city councillor received (we stimulus sampled by testing three similar emails requesting help), the log-transformed population size of the city councillor's city, a binary indicator for whether the city councillor is a Democrat, a binary indicator for whether the city councillor is a Republican, a continuous variable for the number of years until the city councillor faces re-election (0 if the participant has been replaced), and a binary indicator for whether the city councillor was replaced in 2020 just prior to our experiment. We also include fixed effects for the city councillor's region of the country, as determined by the US Census (northeast, midwest, south and west). Robust standard errors are reported in parentheses.

details about the covariates included in our primary regression (in Supplementary Methods for Study 1 and in Supplementary Table 2), as well as additional preregistered analyses examining senders' gender and race separately (in Supplementary Table 8).

We find the expected, significant positive interaction between assignment to the identity-mentioned condition and assignment to the minority-help-seeker condition ($b = 0.097$, standard error (s.e.) = 0.038, 95% CI [0.024, 0.171]; $P = 0.010$; see Table 1, Model 1 for full regression results). This result is robust to the removal of our preregistered covariates ($b = 0.100$, s.e. = 0.038, 95% CI [0.027, 0.174]; $P = .007$) and to analysing our data using a logistic regression instead of an OLS regression ($b = 0.441$, s.e. = 0.173, 95% CI [0.101, 0.782]; $P = 0.011$; see Supplementary Tables 4 and 5 for full regression results). We found no evidence of an effect of assignment to the identity-mentioned condition on response rates ($b = -0.023$, s.e. = 0.026, 95% CI [0.075, 0.029]; $P = 0.380$) and no evidence of an effect of assignment to the minority-help-seeker condition on response rates ($b = -0.010$, s.e. = 0.026, 95% CI [−0.062, 0.042]; $P = 0.705$). In summary, these results show that White male city councillors in our audit study were a regression-estimated 7.42 percentage points (or 24.4%) more likely to respond to help-seeking emails from women and racial/ethnic minorities when the emails city councillors received mentioned the help seeker's demographic identity. We may not have detected evidence of discrimination against women and racial/ethnic minorities overall because of our audit study's context: past work finds mixed evidence as to whether local politicians discriminate against women and racial/ethnic minorities when responding to help requests[3,11,12,35,36].

We did not find evidence that our key interaction was attenuated or strengthened by any of the exploratory variables we preregistered as potential moderators. These included (1) the city councillor's political party, (2) the county's log-transformed median household income, (3) the log-transformed city population, (4) the county's Republican vote share in the 2016 presidential election and (5) the percentage of the population that was White in the county as of 2016 (see Supplementary Tables 9–13 for full regression results).

In exploratory analyses that were not preregistered, we examined the quality of help city councillors offered by considering five different outcomes. The first three outcomes were hand-coded by a team of three research assistants who were unaware of our hypotheses (see Supplementary Methods for Study 1 for details). Specifically, we examined (1) whether the city councillor provided specific advice to

the student (15.8% did; interrater intraclass correlation coefficient ICC(3,3) = 0.96); (2) whether the city councillor suggested scheduling a call or a meeting (18.4% did; interrater ICC(3,3) = 0.76); and (3) whether the city councillor offered a work or volunteer opportunity (5.1% did; interrater ICC(3,3) = 0.76). We also examined the length of each city councillor's response message. Following Kalla et al., we operationalized length of response both by calculating (1) the log word count of the city councillor's reply (mean = 1.371; s.d. = 2.080) and (2) the log character count of the city councillor's reply (mean = 1.900; s.d. = 2.829)[12]. To predict each of these five measures of response quality, we relied on our primary preregistered OLS regression specification where the effect of interest was the interaction between assignment to the identity-mentioned condition and assignment to the minority-help-seeker condition (see Table 1, Models 2 and 3 and Supplementary Table 14 for full regression results).

We found that city councillors wrote more words (a regression-estimated 31.8% more) and more characters (a regression-estimated 46.6% more) in response to women and racial/ethnic minorities when their emails mentioned their demographic identity (word count regression: interaction $b = 0.390$, s.e. = 0.167, 95% CI [0.062, 0.717]; $P = 0.020$; see Table 1, Model 2; character count regression: interaction $b = 0.530$, s.e. = 0.227, 95% CI [0.085, 0.975]; $P = 0.020$; see Table 1, Model 3). We were not able to detect evidence of a difference across conditions for any other measures of response quality: we did not find an effect of the interaction between assignment to the minority-help-seeker condition and the identity-mentioned condition in regressions predicting the likelihood that city councillors offered specific advice (interaction $b = 0.040$, s.e. = 0.029, 95% CI [−0.018, 0.097]; $P = 0.177$; see Supplementary Table 14, Model 1), suggested scheduling a meeting (interaction $b = 0.057$, s.e. = 0.031, 95% CI [−0.004, 0.118]; $P = .067$; see Supplementary Table 14, Model 2) and offered work or volunteer opportunities (interaction $b = 0.012$, s.e. = 0.018, 95% CI [−0.023, 0.046]; $P = .506$; see Supplementary Table 14, Model 3).

Taken together, this indicates that when women and racial/ethnic minorities mentioned their demographic identity in requests for help, they received more and longer replies, although we found no evidence that the quality of those replies differed.

Study 2: Audit experiment with undergraduate students. In Study 2 we aimed to establish the generalizability of our findings by replicating the key results from Study 1 in a different field context with a different population and a different type of help request. While participants in Study 1 were all White men, Study 2 participants were a demographically diverse group of 1,169 undergraduate members (69.5% non-White, 65.7% female) of the behavioural lab participant pool at an East Coast university.

All Study 2 participants received an email from the behavioural lab containing a forwarded request for research help from a fictitious graduate student named Demarcus Rivers (a name chosen to signal a Black male demographic identity; see Supplementary Methods for Study 2). The email was identical across conditions except for one randomized element: in the identity-mentioned condition, Demarcus's request included an explicit mention of his demographic identity ("As a Black man…"), while in the identity-not-mentioned condition, his email did not mention his demographic identity ("As someone…"). Summary statistics describing participant characteristics are included in Supplementary Table 15, and balance checks presented in Supplementary Table 16 show that we did not detect an imbalance across experimental conditions on any observable participant characteristics.

Our dependent variable of interest was whether undergraduates volunteered to help Demarcus by providing their contact information. Anyone who provided their email address was counted as volunteering. Consistent with our hypothesis, significantly more undergraduates in the identity-mentioned condition shared their contact information with Demarcus (6.14%) than in the identity-not-mentioned condition (3.43%; two-sample, two-tail proportions test: $z = 2.17$, $P = 0.030$, effect size $h = 0.128$, 95% CI [0.003, 0.052]). The fact that Study 1 emails were sent to an individual recipient while Study 2 emails were sent to a group of recipients may partially account for the much lower email response rate in Study 2, as prior work has demonstrated that sending emails to multiple recipients leads to a diffusion of responsibility and, ultimately, lower response rates[37]. There is also a norm of paying behavioural lab participants for their participation in research, and Demarcus did not offer compensation for help, whereas there is no norm of paying city councillors to respond to constituent emails.

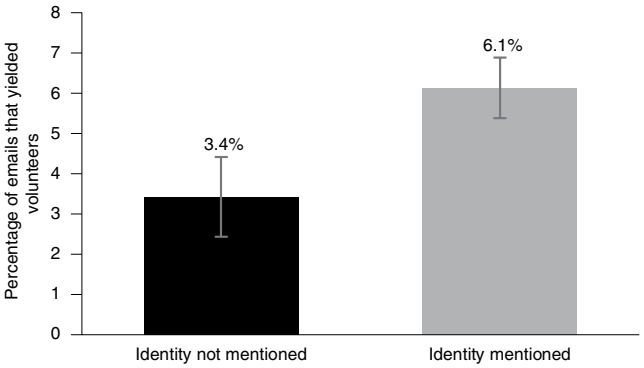

Fig. 3 | Percentage of emails that yielded volunteers across conditions in Study 2. The percentage of undergraduates ($n = 1{,}169$) who volunteered to help a fictitious Black male graduate student with his dissertation research in response to a help-seeking email in Study 2 by experimental condition. The black bar displays the percentage of undergraduates who volunteered in the identity-not-mentioned condition and the grey bar displays the percentage of undergraduates who volunteered in the identity-mentioned condition. Standard error bars are depicted around each proportion. Full regression results estimating the significance of these effects are provided in Supplementary Table 17 (using an OLS regression) and Table 18 (using a logistic regression).

As in Study 1, we again conducted an OLS regression with robust standard errors to predict whether each undergraduate participant in our study volunteered to help Demarcus. The primary predictor in this regression was an indicator for whether the undergraduate participant was assigned to the identity-mentioned condition. We controlled for participant gender, race/ethnicity and political ideology (measured on a seven-point Likert scale from 'Very liberal' to 'Very conservative'). These control variables were provided by the behavioural lab and were collected when each undergraduate in our study first signed up to participate in behavioural lab research (see Supplementary Table 15 for more details about these covariates). The volunteer data violated both normality and homoskedasticity assumptions because the outcome measured was binary, but our primary analysis is an OLS regression (following Study 1) because it is the recommended method for estimating treatment effects on binary outcomes in experiments[34]. We present logistic regression results as robustness checks.

Our OLS regression indicates that undergraduates were an estimated 2.73 percentage points more likely to help the Black male graduate student when his request for help highlighted his demographic identity than when it did not ($b = 0.027$, s.e. = 0.013, 95% CI [0.003, 0.052]; $P = 0.029$; see Supplementary Table 17 for complete regression results and Supplementary Table 18 for regression results relying on a logistic regression rather than an OLS regression model). This means students volunteered to assist Demarcus 79.6% more when he mentioned his demographic identity in his

request for help (see Fig. 3). Furthermore, this result is robust to the removal of our covariates ($b = 0.027$, s.e. = 0.012, 95% CI [0.003, 0.052]; $P = 0.030$).

We found no evidence that the effect of the identity-mentioned condition varied as a function of participant characteristics, including their gender, race, political ideology or age (see Supplementary Tables 19–22 for details).

Study 3: Online experiment. Studies 1 and 2 provided evidence from the field that prospective helpers are more willing to assist women and racial/ethnic minorities when they explicitly mention their demographic identity in requests for help. In Study 3, we relied on an online scenario paradigm to explore whether this result may be correlated with people's increased internal and external motivations to respond without prejudice when a help seeker explicitly mentions their demographic identity.

Study 3 participants were 1,500 adults recruited through Prolific. Participants were asked to imagine being a Computer Science instructor tasked with choosing one (out of four) former students to refer to a prestigious conference. They read emails from each of the four candidates requesting a referral before choosing one to assist. One of the four students was a Black male. Participants were randomly assigned to either the identity-mentioned condition, in which the Black male student explicitly highlighted his demographic identity in his email, or the identity-not-mentioned condition, in which he did not. Participants ranked the four candidates from the one they were most likely to refer (no. 1) to the one they were least likely to refer (no. 4). After making their decisions, participants responded to items from two scales: one intended to measure the extent to which internal motivation to respond without prejudice influenced their decision, and one intended to measure the extent to which external motivation to respond without prejudice influenced their decision (both adapted from Plant and Devine[22]; see Supplementary Methods for Study 3 for details).

Our preregistered dependent variable of interest was the ranking participants assigned to the Black student. This ranking could vary from 1 (if the participant indicated they were most likely to refer the Black student) to 4 (if the participant indicated they were least likely to refer the Black student). Smaller numbers indicate a greater willingness to help the Black student. Consistent with our findings from Studies 1 and 2, we find that, on average, participants in the identity-mentioned condition ranked the Black male student higher (2.68 out of 4; s.d. = 1.00) than participants in the identity-not-mentioned condition (2.94 out of 4; s.d. = 0.97; two-tail t-test: $t(1498) = 5.06$, $P < 0.001$, Cohen's $d = 0.261$, 95% CI [0.157, 0.357]). The ranking data were not normally distributed but did meet the equal variance assumption, so we confirmed that this result was robust to using a non-parametric Mann–Whitney U-test instead of a t-test ($z(1,498) = -5.12$, $P < 0.001$, Cliff's delta = −0.936, 95% CI for the delta estimate = [0.926, 0.944]). Participants in the identity-mentioned condition were also more likely to choose to refer the Black male student (by ranking him first) than participants in the identity-not-mentioned condition (15.1% versus 10.4%; two-sample, two-tail proportions test: $z = 2.65$, $P = .008$, effect size $h = 0.141$, 95% CI [0.012, 0.082]; see Supplementary Fig. 1).

Next, we tested whether our hypothesized mechanisms mediated the effect of the identity-mentioned condition on willingness to help the Black male student. We present the results of mediation analyses to explore the relationships between our treatment, dependent variable of interest and hypothesized mechanisms, but we also note that there are inherent weaknesses to mediation analysis with measured rather than manipulated mediators[38]. Specifically, causal mediation analysis relies on the Sequential Ignorability Assumption, which states, in part, that no omitted pretreatment covariates are correlated with both the mediator and the outcome[39]. To address this issue, we conducted sensitivity analyses developed by Imai et al. to assess the robustness of our findings to deviations from the Sequential Ignorability Assumption[40]. The sensitivity parameter is ρ, which varies between −1 and 1 and indicates the magnitude of the correlation between the errors of the mediation and outcome models necessary for our mediation results to be null, or to reverse in direction (ρ is 0 when the Sequential Ignorability Hypothesis holds). We preregistered our mediation analyses, but the sensitivity analyses are exploratory and were not preregistered.

Following our preregistration, we tested each proposed mediator independently using a 10,000-sample bootstrapped mediation model and a Sobel test, and we tested both proposed mediators together with a 10,000-sample bootstrapped multiple mediation model. We find that the 95% bias-corrected CI for the size of the indirect effect of the salience of internal motivation to respond without prejudice excluded zero (95% CI [−0.111, −0.050]). A Sobel test confirmed that the reduction in effect size was statistically significant ($b = -0.079$, s.e. = 0.016, $P < 0.001$). In particular, Imai et al.'s average causal mediation effect approach suggests that 30.7% of the effect of mentioning identity on willingness to refer the Black male student occurs through mediation by internal motivation to control prejudice[40]. Furthermore, a 1,000-sample bootstrapped sensitivity analysis concluded that this effect is robust to sizable deviations from the Sequential Ignorability Assumption, as the indirect effect of the salience of internal motivation to control prejudice is negative and non-zero for any $\rho > -0.28$. Meanwhile, the 95% bias-corrected CI for the size of the indirect effect of the salience of external motivation to respond without prejudice included zero (95% CI [−0.024, −0.000]), and a Sobel test confirmed that we did not find a reduction in the size of the treatment effect when controlling for external motivation to control prejudice ($b = -0.010$, s.e. = 0.006, $P = 0.083$). The average causal mediation effect approach suggests that only 4.1% of the effect of mentioning identity occurs through mediation by external motivation to control prejudice[40]. Furthermore, this effect is not robust to deviations from the Sequential Ignorability Assumption: the indirect effect of the salience of external motivation to control prejudice is null, even when $\rho = 0$.

Notably, internal and external motivation to control prejudice were highly correlated in our study ($r = 0.612$; $P < 0.001$). To address this multicollinearity issue, we ran a preregistered multiple mediation model. In this model, the 95% CI for the indirect effect of internal motivation to respond without prejudice once again excluded zero (95% CI [−0.154, −0.069]). However, the multiple mediation model suggested that, conditional on the inclusion of internal motivation to control prejudice in the mediation model, higher external motivation to respond without prejudice was related to a lower willingness to help the Black male student when he mentioned his demographic identity (95% CI [0.018, 0.063]). Thus, internal motivation to respond without prejudice was the only positive correlate of the benefits of mentioning demographic identity in our multiple mediation model.

Discussion

Across two field experiments and one online experiment, we find evidence that women and racial/ethnic minorities are more likely to receive instrumental help when their requests for assistance explicitly highlight their demographic identity. City council members in Study 1 were 24.4% (7.42 percentage points) more likely to respond to help-seeking emails when women and racial/ethnic minorities mentioned their identity. Notably, this 7.4 percentage-point boost in response rates is larger than the discriminatory gaps identified in prior audit experiments. The discriminatory gap in responses from state legislators to Black versus White men identified in past research was 5.1 percentage points, while the discriminatory gap in callbacks for résumés with Black versus White names identified in past research was 3.2 percentage points[3,20].

We also found that undergraduates in Study 2 were 79.6% (2.73 percentage points) more likely to volunteer to help a Black male graduate student when he highlighted his identity in his request. The benefits of mentioning demographic identity were robust, regardless of the political affiliation or demographic identity of the individual receiving a request for help. Study 3 provides evidence that prospective helpers' internal motivation to respond without prejudice is tied to an increased willingness to help marginalized identity group members when they explicitly mention their demographic identity. These results suggest that women and racial/ethnic minorities stand to reap important benefits if they mention their demographic identity in help requests.

Making the potential for poor judgement more salient has been shown to improve decision-making in many domains, but this insight has seldom been applied to issues of diversity and inclusion[41–43]. Drawing timely attention to the risk of exhibiting prejudice may have an important and underappreciated impact on decision-making that is worthy of further theorizing and study.

Our work also suggests that features of a decision-making environment are related to people's motivations to respond without prejudice. Past research has characterized internal motivation to respond without prejudice as a static trait, but we demonstrate that it is dynamic and context-dependent[22,44]. That is, when women and minorities highlight their identity in requests for help, prospective helpers indicate being more motivated to overcome their prejudice. It would be worthwhile for future work to provide causal evidence that motivation to control prejudice can, indeed, be harnessed to improve outcomes for women and racial/ethnic minorities and, if so, to explore other ways to capitalize on this motivation.

We find that the benefits of mentioning identity generalize across contexts where those asked for help are both anonymous and identifiable to the requestor. In Study 1, students directly emailed city council members to request help, and in Study 2, requests for help were sent via a third party (a university behavioural lab) to a large mailing list. As a result, the prospective helper was afforded a degree of anonymity in Study 2, making Study 2 more similar to past résumé audit studies in which the evaluator had information about the person being evaluated, but the converse was not true (for example, in the work by Bertrand and Mullainathan)[20]. The fact that we found consistent results across Studies 1 and 2 suggests that our findings are not dependent on a prospective helper's expectation that a help seeker would observe their response.

Moreover, our findings suggest that decision-makers' internal motivation to control prejudice—and not external motivation—is related to improved outcomes for women and racial/ethnic minorities who mention their demographic identity. Associations between willingness to help women and racial/ethnic minorities and external motivation to control prejudice may have been weaker in our studies because decision-makers were not making public decisions. Even when they were not anonymous to the help seeker, no one else was privy to the decision participants in our studies faced. Thus, reputational concerns may have been less salient than self-signalling concerns. Future research might explore whether external motivation to control prejudice plays a stronger role in driving decisions made publicly or in groups. Future research might also explore other potential mechanisms for the effect of mentioning identity, such as increases in the perceived impact of help provided or in prospective helpers' desire to behave altruistically.

Although we replicate our findings in two audit studies with different populations, future research replicating and extending our work would be valuable. Because our experiments focus on emailed requests to strangers for informal help, we cannot determine how mentions of demographic identity might affect other decisions. Mentioning your demographic identity may have a different effect when you make more formal requests, interact with people you already know, make face-to-face requests, ask for long-term help (for example, mentorship) or seek other outcomes (for example, a job, promotion or feedback). Exploring these variations on our paradigm would be useful. Similarly, we do not know if our findings would extend to directly disclosing identity dimensions beyond race/ethnicity and gender, such as socioeconomic status, sexuality, disability, veteran status, ideological identity or religious identity, and further research exploring this would therefore be valuable.

Our studies also primarily focused on one outcome measure: whether a request for help elicits a response. Future studies might explore other outcomes, such as the psychological consequences help seekers experience after mentioning their demographic identity. Women and racial/ethnic minority help seekers who highlight their identity and do not receive help might be more discouraged, as they may be more likely to attribute undesirable outcomes to prejudice.

It would also be valuable for future work to explore whether help seekers who mention their identity produce positive spillover effects for other, future help seekers from marginalized groups. In other words, if someone receives an email from a woman or racial/ethnic minority requesting help that explicitly mentions the sender's identity, is that recipient more likely to help other women and racial/ethnic minorities who reach out subsequently?

Women and racial/ethnic minorities have long been left out of positions of power, held back by negative stereotypes, prejudice, tokenism and in-group favouritism[10,45–49]. Time and again, evidence has shown that when information about an individual's marginalized identity is communicated inadvertently, it limits women and racial/ethnic minorities' opportunities[9,18–20]. In this work, however, we demonstrated that when women and racial/ethnic minorities deliberately reveal their identity in a request for help, it can be to their advantage.

Methods

This research was approved by the Institutional Review Board at the University of Pennsylvania and complies with all relevant ethical regulations. We received a waiver of informed consent for Studies 1 and 2, and informed consent was obtained from all study participants in Study 3. Participants in Study 3 were compensated for their time with a flat fee (US$0.80) while participants in Studies 1 and 2 were not compensated. The reference number for Study 1 is 833579, for Study 2 is 843870 and for Study 3 is 855057. All study preregistrations, anonymized data and analysis code can be found on Open Science Framework (https://doi.org/10.17605/OSF.IO/5DHBE).

Studies 1, 2 and 3 were preregistered on 8 July 2020, 18 September 2020 and 10 November 2020, respectively. The OSF folder also includes our Supplementary Information, which contains further details about the methods and results for each study. Data collection and analysis were not performed blind to the conditions of the experiments.

Study 1: Audit experiment with city councillors. Study 1 tested our hypothesis in a preregistered email audit experiment. Participants were 2,476 White male city councillors from 701 of the largest cities in the United States (by population, based on 2019 Census data; see Supplementary Table 2 for participant summary statistics)[50,51]. No statistical methods were used to pre-determine sample sizes but our sample sizes were similar to those reported in previous publications[3,11,12,35,36]. A team of research assistants inferred councillors' gender and race/ethnicity from publicly available information. Specifically, research assistants used names, photographs and personal biographies to glean information about councillors' gender and racial/ethnic characteristics. When a city councillor's demographic identity could not be classified based on this information, they searched news articles and social networking sites (for example, Facebook and LinkedIn) to see if the city councillor self-reported their demographic identity publicly. Each city councillor's gender and race/ethnicity were classified by two research assistants, and any disagreements were resolved by the first author of this manuscript, who again used phenotypic judgements, public news sources and social media profiles to classify city councillors' demographic identities. If the first author's classification corresponded with that of one of the two research assistants, that classification was applied; otherwise, the city councillor was considered unclassifiable. City councillors who could not be classified as White men were not included in our sample. A limitation of this approach was that if any city councillors presented phenotypically as White males and did not publicly self-identify otherwise, they may have been misclassified and included in our sample. City councillors' ages were not available online, so we were unable to collect data on age. Mayors were not included in this study, even if they served on their city council.

Each city councillor in our study received an email from a fictitious student on the morning of 14 July 2020. The email stated that the student had dreams of a career in politics and asked the city councillor to write back with career advice. All emails were identical except for two randomized features: (1) the help seeker's demographic identity and (2) whether the help seeker explicitly mentioned their demographic identity in the email. Randomization of city councillors to conditions was stratified by their city to ensure balance on this dimension. A total of 621 city councillors were assigned to the minority help seeker x identity-mentioned condition, 620 city councillors were assigned to the White male help seeker x identity-mentioned condition, 625 city councillors were assigned to the minority help seeker x identity-not-mentioned condition and 610 city councillors were assigned to the White male help seeker x identity-not-mentioned condition.

Following past research, our audit experiment varied the identity of the help seeker by selecting names that signalled the student's gender and race/ethnicity[9,12,20]. Names were chosen to signal one of six demographic identities: White male, White female, Black male, Black female, Latino or Latina. Specifically, we used 2010 US Census data to identify common surnames typically belonging to White, Black and Latinx individuals, and used online baby name lists to identify popular first names for different gender and racial/ethnic groups (see Supplementary Methods for Study 1 for more details). We then combined these to create full names and asked an online sample to infer the gender and racial/ethnic identity of each name. We selected the four names with the highest demographic recognizability for each race/ethnicity-gender combination of interest, or 24 names total (all names can be found in Supplementary Table 1, along with further information about how they were created and selected in Supplementary Methods for Study 1). City councillors were randomly assigned to receive an email from either a help seeker with a White male-sounding name in the White male help seeker condition or a help seeker with a non-White male-sounding name (that is, a sender with a female and/or Black or Latinx-sounding name) in the minority-help-seeker condition.

Our experiment included an additional variable component that appeared in the opening sentence of the email. In this sentence, the help seeker either did or did not explicitly mention their demographic identity, asking the city councillor to share advice with "a young [person]/[man/woman/Black man/Black woman/Latino/Latina] hoping to become a city councillor". In the identity-not-mentioned condition, the student made no mention of their identity and asked the city councillor to share advice with "a young person". By contrast, in the identity-mentioned condition, the help seeker asked the city councillor if they would be willing to share advice with "a young man/woman" (for White senders), "a young Black man/woman" (for Black senders) or "a young Latino/a" (for Latinx senders), thereby explicitly mentioning their identity. We did not explicitly reference the White senders' race in the identity-mentioned condition (that is, by asking city councillors to share advice with a "young White man/woman") because qualitative data suggested that by labelling themselves explicitly as 'White', senders might signal White nationalist political attitudes.

Complete study stimuli and further details about our methods are available in Supplementary Methods for Study 1.

Study 2: Audit experiment with undergraduate students. Our participants were 1,169 undergraduate members of the behavioural lab participant pool at a large East Coast university (65.7% female; 30.5% White, 35.8% Asian, 15.7% Black, 10.2% Latinx and 7.8% other; average age was 19.8 years old). We used G*Power to calculate the sample size needed to detect an effect size similar to that of the identity-mentioned condition for women and racial/ethnic minorities in Study 1 ($h = 0.164$) with 80% power. The result was 1,162. To fulfil this required sample size, we contacted all undergraduate members of the behavioural lab's participant pool.

The behavioural lab sent an email to active members of its undergraduate participant pool on 23 September 2020, with the subject line "Request for Research Help". The email explained that the behavioural lab was forwarding a request for free research help from a PhD student named Demarcus Rivers (a fictitious student whose name was selected to signal a Black male identity; see Supplementary Methods for Study 2 for more details about the name selection procedure).

Demarcus's forwarded message was identical across experimental conditions except for one randomized element: whether his demographic identity was explicitly mentioned in the email's opening lines or not (it was mentioned in the identity-mentioned condition and omitted in the identity-not-mentioned condition). Specifically, the opening lines of the email read: "Hi, I'm Demarcus Rivers. As [a Black man]/[someone] working towards a PhD during this difficult time, I could really use your help." Demarcus went on to ask undergraduates for their contact details if they were willing to volunteer, without pay, to complete a 15-minute phone interview for his dissertation research. We stratified randomization to conditions within our sample by participant gender and race ('Asian', 'Black', 'Hispanic', 'Native American', 'White' and 'declined to answer': these categories were provided by the behavioural lab) to ensure balance on these dimensions. A total of 586 undergraduates were assigned to the identity-mentioned condition and 583 undergraduates were assigned to the identity-not-mentioned condition.

After our study launched, one professor at the East Coast university in question offered their students extra class credit for volunteering to help the (fictional) Black PhD student in our audit experiment. Because our intention was to test participants' willingness to offer help to a minority student with no external incentive, we excluded the 272 students who we learned had been offered this extra credit from our analyses. We were informed about the extra class credit because the professor in question asked the behavioural lab to confirm which students had volunteered to help Demarcus. The behavioural lab confirmed that no other professor had offered an external incentive. This led to a final sample size of 1,169, rather than the sample size of 1,441 that we originally preregistered, so this study is not formally preregistered. We otherwise followed our preregistered analysis plan in full. We include analyses with our full dataset in Supplementary Tables 23 and 24.

Complete study stimuli and further details about our methods are in Supplementary Methods for Study 2 and Supplementary Information screenshots of the volunteer survey linked in the behavioural lab email in Study 2.

Study 3: Online experiment. We recruited 1,500 participants (48.4% female; 73.3% White) through Prolific on 11 November 2020 to participate in a preregistered 7-minute study in exchange for US$0.80. We did not collect data on participants' age for this study because our Institutional Review Board recommended that we collect only demographic information deemed relevant to our experiment's focus, and in this case, we decided to collect only participant gender and race/ethnicity. We used G*Power to calculate the sample size we would need to detect an effect size of 0.19 with 95% power and ultimately preregistered a sample size of 1,500. When collecting our data, the Prolific platform allowed three extra participants to complete the experiment. To comply with our preregistration, we excluded the three participants who completed the study last. All our results were consistent when we included these participants.

Participants were asked to imagine that they were Computer Science instructors at a university tasked with selecting one former student to refer to a prestigious conference. Participants who passed a three-question attention check then read four emails, presented in random order, from students requesting a referral to this conference. The students' names signalled their gender and race/ethnicity (see Supplementary Methods for Study 3 for details about how the names were selected for this study). All participants read two emails from White men (Brad Miller and Todd Anderson), one email from a White woman (Emma Nelson) and one email from a Black man (Hakeem Mosley). Everyone in the study was randomly assigned to one of two different conditions, which determined the content of the email they reviewed from Hakeem Mosley (a Black man). In the identity-mentioned condition ($n = 753$), the email from Hakeem Mosley highlighted his demographic identity (the second sentence began with the statement "As a Black student..."). In the identity-not-mentioned condition ($n = 747$), the email did not explicitly mention Hakeem's race (the second sentence began with the statement "As a student..."). These emails were otherwise identical, and all other emails were identical across conditions.

After reviewing the four student emails, participants were asked to rank the students in order from the one they were most likely to refer (no. 1) to the one they were least likely to refer (no. 4). Participants then answered a series of questions designed to measure the thought processes underlying their rankings. For each question, participants indicated their agreement with a statement on a seven-point Likert scale ranging from '1: Strongly disagree' to '7: Strongly agree'.

To measure the extent to which participants were motivated to act consistently with their values when deciding which student to refer to the conference, we adapted four items from Plant and Devine's internal motivation to respond without prejudice scale (for example, "Given my personal values and beliefs, an important factor in my decision was my desire to promote the success of racial/ethnic minorities"; Cronbach's $\alpha = 0.87$)[22]. We standardized each item, then averaged them to create a scale.

To measure the extent to which participants considered impression management motives when deciding which students to refer to the conference, we adapted three items from Plant and Devine's external motivation to respond without prejudice scale (for example, "Given today's PC (political correctness) standards, a factor in my decision was that I should do my best not to act racist"; Cronbach's $\alpha = 0.85$)[22]. We standardized each item, then averaged them to create a scale.

The questions on each of the two scales described above were presented in randomized order. After participants responded to these scale items, we asked them how many students from different identity groups (for example, White women, White men, Black women, Black men, and so on) they recalled requesting a referral (as a manipulation check). Finally, participants reported their own gender and race/ethnicity. Further study details are included in Supplementary Methods for Study 3 and all study stimuli and scale items are included in Supplementary Information in the screenshots of Study 3 survey.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

De-identified participant data from Studies 1, 2 and 3 are permanently and publicly available in an OSF folder at https://doi.org/10.17605/OSF.IO/5DHBE. There are no restrictions on data availability. These data include information about

participant responses, condition assignment and any preregistered control variables for each of our studies. The OSF folder also includes preregistrations for Studies 1, 2 and 3 as well as a copy of our Supplementary Information, which contains further details about our experimental methods and results (for example, stimulus language, screenshots of surveys shared with participants in Study 3 and results of robustness checks). Figures that have associated raw data include Figs. 1, 2 and 3. The raw data for these figures are also included in the OSF folder. Source data are provided with this paper.

Code availability

The code to replicate the analyses in the manuscript and our Supplementary Information is available permanently and publicly in an OSF folder at https://doi.org/10.17605/OSF.IO/5DHBE.

Received: 27 May 2021; Accepted: 10 November 2021;
Published online: 20 January 2022

References

1. Coury, S. et al. Women in the workplace 2020. *McKinsey & Company* https://www.mckinsey.com/featured-insights/diversity-and-inclusion/women-in-the-workplace (2020).
2. The NCES Fast Facts Tool provides quick answers to many education questions. *National Center for Education Statistics* https://nces.ed.gov/fastfacts/display.asp?id=61 (2020).
3. Butler, D. M. & Broockman, D. E. Do politicians racially discriminate against constituents? A field experiment on state legislators. *Am. J. Pol. Sci.* **55**, 463–477 (2011).
4. Giuliano, L., Levine, D. I. & Leonard, J. Racial bias in the manager-employee relationship an analysis of quits, dismissals, and promotions at a large retail firm. *J. Hum. Resour.* **46**, 26–52 (2011).
5. Keeves, G. D. & Westphal, J. D. From help to harm: increases in status, perceived underreciprocation, and the consequences for access to strategic help and social undermining among female, racial minority, and white male top managers. *Organization Sci.* **32**, 909–1148 (2021).
6. Lavy, V. & Sand, E. On the origins of gender gaps in human capital: short-and long-term consequences of teachers' biases. *J. Public. Econ.* **167**, 263–279 (2018).
7. McDonald, M. L., Keeves, G. D. & Westphal, J. D. One step forward, one step back: White male top manager organizational identification and helping behavior toward other executives following the appointment of a female or racial minority CEO. *Acad. Manag. J.* **61**, 405–439 (2018).
8. Milkman, K. L., Akinola, M. & Chugh, D. Temporal distance and discrimination: an audit study in academia. *Psychol. Sci.* **23**, 710–717 (2012).
9. Milkman, K. L., Akinola, M. & Chugh, D. What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *J. Appl. Psychol.* **100**, 1678–1712 (2015).
10. Price, J. & Wolfers, J. Racial discrimination among NBA referees. *Q. J. Econ.* **125**, 1859–1887 (2010).
11. White, A. R., Nathan, N. L. & Faller, J. K. What do I need to vote? Bureaucratic discretion and discrimination by local election officials. *Am. Polit. Sci. Rev.* **109**, 129–142 (2015).
12. Kalla, J., Rosenbluth, F. & Teele, D. L. Are you my mentor? A field experiment on gender, ethnicity, and political self-starters. *J. Politics* **80**, 337–341 (2018).
13. Eby, L. T., Allen, T. D., Evans, S. C., Ng, T. & DuBois, D. L. Does mentoring matter? A multidisciplinary meta-analysis comparing mentored and non-mentored individuals. *J. Vocat. Behav.* **72**, 254–267 (2008).
14. Kaas, L. & Manger, C. Ethnic discrimination in Germany's labour market: a field experiment. *Ger. Econ. Rev.* **13**, 1–20 (2012).
15. Seibert, S. E., Kraimer, M. L. & Liden, R. C. A social capital theory of career success. *Acad. Manag. J.* **44**, 219–237 (2001).
16. Bohren, J. A., Imas, A. & Rosenberg, M. The dynamics of discrimination: theory and evidence. *Am. Econ. Rev.* **109**, 3395–3436 (2019).
17. Doleac, J. L. & Stein, L. C. The visible hand: race and online market outcomes. *Econ. J.* **123**, F469–F492 (2013).
18. Edelman, B., Luca, M. & Svirsky, D. Racial discrimination in the sharing economy: evidence from a field experiment. *Am. Econ. J. Appl. Econ.* **9**, 1–22 (2017).
19. Kang, S. K., DeCelles, K. A., Tilcsik, A. & Jun, S. Whitened résumés: race and self-presentation in the labor market. *Adm. Sci. Q.* **61**, 469–502 (2016).
20. Bertrand, M. & Mullainathan, S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* **94**, 991–1013 (2004).
21. Dovidio, J. F. & Gaertner, S. L. Aversive racism and selection decisions: 1989 and 1999. *Psychol. Sci.* **11**, 315–319 (2000).
22. Plant, E. A. & Devine, P. G. Internal and external motivation to respond without prejudice. *J. Pers. Soc. Psychol.* **75**, 811–832 (1998).
23. Apfelbaum, E. P., Sommers, S. R. & Norton, M. I. Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *J. Pers. Soc. Psychol.* **95**, 918 (2008).
24. Bodner, R. & Prelec, D. in *The Psychology of Economic Decisions* Vol. 1 (eds Brocas, I. & Carillo, J.) (Oxford Univ. Press, 2003).
25. Paluck, E. L. & Green, D. P. Prejudice reduction: what works? A review and assessment of research and practice. *Annu. Rev. Psychol.* **60**, 339–367 (2009).
26. Plant, E. A. & Devine, P. G. The active control of prejudice: unpacking the intentions guiding control efforts. *J. Pers. Soc. Psychol.* **96**, 640–652 (2009).
27. Rokeach, M. Long-range experimental modification of values, attitudes, and behavior. *Am. Psychol.* **26**, 453–459 (1971).
28. Pope, D. G., Price, J. & Wolfers, J. Awareness reduces racial bias. *Manag. Sci.* **64**, 4988–4995 (2018).
29. Sommers, S. R. & Ellsworth, P. C. White juror bias: an investigation of prejudice against Black defendants in the American courtroom. *Psychol. Public. Policy. Law.* **7**, 201–229 (2001).
30. Banaji, M. R. & Hardin, C. D. Automatic stereotyping. *Psychol. Sci.* **7**, 136–141 (1996).
31. Devine, P. G. Stereotypes and prejudice: their automatic and controlled components. *J. Pers. Soc. Psychol.* **56**, 5–18 (1989).
32. Taylor, S. E., Fiske, S. T., Etcoff, N. L. & Ruderman, A. J. Categorical and contextual bases of person memory and stereotyping. *J. Pers. Soc. Psychol.* **36**, 778–793 (1978).
33. Ai, C. & Norton, E. C. Interaction terms in logit and probit models. *Econ. Lett.* **80**, 123–129 (2003).
34. Gomila, R. Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *J. Exp. Psychol. Gen.* **150**, 700–709 (2021).
35. Butler, D. M. & Crabtree, C. Moving beyond measurement: adapting audit studies to test bias-reducing interventions. *J. Exp. Pol. Sci.* **4**, 57 (2017).
36. Einstein, K. L. & Glick, D. M. Does race affect access to government services? An experiment exploring street-level bureaucrats and access to public housing. *Am. J. Pol. Sci.* **61**, 100–116 (2017).
37. Barron, G. & Yechiam, E. Private e-mail requests and the diffusion of responsibility. *Comput. Hum. Behav.* **18**, 507–520 (2002).
38. Bullock, J. G., Green, D. P. & Ha, S. E. Yes, but what's the mechanism? (Don't expect an easy answer). *J. Pers. Soc. Psychol.* **98**, 550–558 (2010).
39. Imai, K., Keele, L. & Tingley, D. A general approach to causal mediation analysis. *Psychol. Methods.* **15**, 309–334 (2010).
40. Imai, K., Keele, L. & Yamamoto, T. Identification, inference and sensitivity analysis for causal mediation effects. *Stat. Sci.* **25**, 51–71 (2010).
41. Fishbane, A., Ouss, A. & Shah, A. K. Behavioral nudges reduce failure to appear for court. *Science* **370**, eabb6591 (2020).
42. Tiefenbeck, V. et al. Overcoming salience bias: How real-time feedback fosters resource conservation. *Manag. Sci.* **64**, 1458–1476 (2018).
43. Zhang, T., Fletcher, P. O., Gino, F. & Bazerman, M. H. Reducing bounded ethicality: how to help individuals notice and avoid unethical behavior. *Organ. Dyn.* **44**, 310–317 (2015).
44. Butz, D. A. & Plant, E. A. Prejudice control and interracial relations: the role of motivation to respond without prejudice. *J. Pers.* **77**, 1311–1342 (2009).
45. Glover, D., Pallais, A. & Pariente, W. Discrimination as a self-fulfilling prophecy: evidence from French grocery stores. *Q. J. Econ.* **132**, 1219–1260 (2017).
46. Heilman, M. E. Description and prescription: how gender stereotypes prevent women's ascent up the organizational ladder. *J. Soc. Issues.* **57**, 657–674 (2001).
47. Ibarra, H. Homophily and differential returns: sex differences in network structure and access in an advertising firm. *Adm. Sci. Q.* **37**, 422–447 (1992).
48. Rosette, A. S., Leonardelli, G. J. & Phillips, K. W. The White standard: racial bias in leader categorization. *J. Appl. Psychol.* **93**, 758–777 (2008).
49. Watkins, M. B., Simmons, A. & Umphress, E. It's not black and white: toward a contingency perspective on the consequences of being a token. *Acad. Manag. Perspect.* **33**, 334–365 (2019).
50. City and town population totals: 2010–2019. *U.S. Census Bureau* https://www.census.gov/data/datasets/time-series/demo/popest/2010s-total-cities-and-towns.html (2020).
51. Kirgios, E. L., Rai, A., Chang, E. H. & Milkman, K. L. When seeking help, women and racial/ethnic minorities benefit from explicitly stating their demographic identity. OSF https://doi.org/10.17605/OSF.IO/5DHBE (2021).

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under grant no. DGE-1845298 (awarded to E.L.K.) and the Wharton School. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We are grateful to A. Duckworth, J. Kessler, A. Rees-Jones, I. Silver and attendees at the Society for Judgment and Decision Making conference for their feedback. We also thank K. Shonk and K. Brabaw for providing editorial input on this manuscript and the Wharton Behavioural

Lab for their assistance in gathering data for this research. Finally, we are grateful to the many research assistants who helped us make this work possible, particularly K. Herrera, M. Chung, M. Huang, C. Kornicker and G.M. Waldman.

Author contributions

E.L.K., A.R., E.H.C. and K.L.M. designed and performed the research. E.L.K. and A.R. analysed the data. E.L.K. wrote the paper. A.R., E.H.C. and K.L.M. provided critical feedback on the paper. E.L.K. prepared the supplementary materials.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-021-01253-y.

Correspondence and requests for materials should be addressed to Erika L. Kirgios.

Peer review information *Nature Human Behaviour* thanks the anonymous reviewers for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Erika Kirgios (ekirgios@wharton.upenn.edu)

Last updated by author(s): Nov 5, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software (i.e., custom code) was used for data collection. Data was collected via email servers (Study 1) and Qualtrics (Studies 2 and 3). Data in Study 1 was double-checked by research assistants.
Data analysis	We used R version 3.5.4 for data analysis. All analysis code is available at https://doi.org/10.17605/OSF.IO/5DHBE.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

De-identified data is publicly and permanently available at https://doi.org/10.17605/OSF.IO/5DHBE. Figures 1, 2, and 3 have associated raw data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	All three studies are quantitative experimental.
Research sample	Experiment 1 included 2,476 White male city council members from 701 of the 777 largest cities the U.S. This sample was chosen because we could access the email addresses and demographic information of the city councilors online. We could not find age information online. All participant characteristics we could find online (i.e., geographic location, ideology) are described in our Supplement. Experiment 2 included 1,169 undergraduates (65.7% female; 30.5% White, 35.8% Asian, 15.7% Black, 10.2% Latinx, and 7.8% Other; average age = 19.8 years old). These are undergraduates who signed up to complete studies through the behavioral lab at the University of Pennsylvania. All participant information shared with us through the behavioral lab is summarized in our Supplement. Experiment 3 included 1,500 online participants (48.4% female, 73.3% White). No further information about these participants was collected.
Sampling strategy	These were samples of convenience. For experiment 1, we used GPower to estimate the sample size we would need to detect an interaction given baseline response rates of 30% (per our pilot) and a 10% increase in the treatment condition with 80% power. For experiment 2, we asked the behavioral lab to email their entire participant pool to maximize power, so our sample size was limited by the size of the pool. Based on the treatment effect in experiment 1 and the available sample size, we opted to do a two-condition study rather than a four-condition study. For experiment 3, we chose a similar sample size to experiment 2 (since it was another two-cell design) but added participants to ensure we had power for our mediation analysis. All sample sizes were preregistered.
Data collection	For experiments 2 and 3, we used Qualtrics. In experiment 2, the DV was signing up to volunteer to help the (fictitious) graduate student on a Qualtrics survey, so we used the survey and information provided by the behavioral lab for our analyses. For experiment 3, all data was collected on Qualtrics. For experiment 1, a tech team at our university created email servers to automatically send out emails and to track responses. Research assistants blind to our hypotheses also tracked response rates manually to confirm that our data was accurate. All data collection was electronic (i.e., no one was present).
Timing	Experiment 1: July 14, 2020 - July 21, 2020. Experiment 2: September 23, 2020- September 30, 2020. Experiment 3: November 11, 2020.
Data exclusions	In experiment 1, we preregistered excluding responses that did not come from the city councilors directly (i.e., emails from assistants or auto-replies). In experiment 2, some data were excluded: After our study launched, one professor at the East Coast university in question offered their students extra class credit for volunteering to help the (fictional) Black PhD student in our audit experiment. Because our intention was to test participants' willingness to offer help to a minority student with no external incentive, we excluded the 272 students who we learned had been offered this extra credit from our analyses. This led to a final sample size of 1,169 rather than the 1,441 sample size that we originally preregistered. Given this significant deviation from our preregistration, we do not consider the study to be formally preregistered although we otherwise followed our preregistered analysis plan in full. We include analyses with our full dataset in our Supplement in Section 5a. In experiment 3, we preregistered including 1500 participants but Prolific allowed 3 extra participants to take our study. In order to remain consistent with our preregistration, we excluded the data from the 3 participants who completed our study the latest (i.e., the final 3 participants), leaving our final sample size at 1500. Our results remain consistent if we include these 3 participants.
Non-participation	No participants dropped out.
Randomization	In experiment 1, participants were randomly assigned to one of four conditions using stratified random assignment by city. In experiment 2, participants were randomly assigned to one of two conditions using stratified random assignment by gender and race. In experiment 3, participants were randomly assigned to one of two conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	In experiment 1, participants were not formally recruited (nor did they know they were participants in a research experiment, as approved by our IRB). They were identified by looking up the city councilors in the largest cities in the U.S. on city council websites (supplemented by Google searches). In study 2, participants again did not know they were part of a research study (again, as approved by our IRB). They were identified by our university's behavioral lab: they were the population of undergraduates who had opted in to take studies and receive emails from the behavioral lab. In experiment 3, participants were recruited via Prolific. This is a sample of people who have chosen to complete online experiments for pay.
Ethics oversight	This research was approved by the Institutional Review Board at the University of Pennsylvania and complies with all relevant ethical regulations. We received a waiver of informed consent for Studies 1 and 2, and informed consent was obtained from all study participants in Study 3. Participants in Study 3 were compensated for their time with a flat fee ($0.80). The reference number for Study 1 is 833579, for Study 2 is 843870, and for Study 3 is 855057.

Note that full information on the approval of the study protocol must also be provided in the manuscript.