BRINGING IDENTITY TO THE FOREFRONT: THE BENEFITS OF HIGHLIGHTING IDENTITY AND DIVERSITY

Erika L. Kirgios

A DISSERTATION

in

Operations, Information, and Decisions

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Katherine L. Milkman, Professor of Operations, Information, and Decisions

Graduate Group Chairperson

Nancy Zhang, Ge Li and Ning Zhao Professor, Professor of Statistics

Dissertation Committee

Maurice E. Schweitzer, Professor of Operations, Information, and Decisions

Angela Duckworth, Professor of Psychology

Modupe Akinola, Professor of Management at Columbia Business School

Sendhil Mullainathan, Professor of Computation and Behavioral Science at University of Chicago

This dissertation is dedicated to Teddy and Veronica, the two people I'd run to at the end of the world. Not because I think any of us would be particularly skilled at surviving the apocalypse, but because there's no one else I'd rather have by my side when life gets hard.

ACKNOWLEDGMENT

I am grateful beyond words for the support and mentorship of my advisor, Katy Milkman, who taught me approximately 79.3% of what I know about behavioral science and who introduced me to the people who taught me the other 20.7%. She is, in every respect, a role model to me: in her brilliance, her work ethic, her motivation to do good with her research, and her ability to somehow do everything at once. When I say I can't imagine a better advisor, I mean it. None of this would have been possible without her.

I am also deeply indebted to the members of my dissertation committee—Modupe Akinola, Angela Duckworth, Sendhil Mullainathan, and Maurice Schweitzer—whose guidance has been invaluable to me. You are all some of the best academics (and people) I have ever met, and I am profoundly lucky to have had the chance to learn from you all. Thank you for seeing me through not just this dissertation, but my academic journey to date.

A massive thank you to my collaborators on this and other work—Edward Chang, Aneesh Rai, Emma Levine, and Judd Kessler—and to the close friends I have made through the Ph.D.—Ike Silver, Sam Skowronek, Linda Chang, Brad Bitterly, Celia Gaertig, Josh Lewis, Katie Mehr, and Robert Mislavsky. I have benefited immensely from both your friendship and your intellect. I also want to thank the many faculty who helped me become the researcher I am today, especially Rebecca Schaumberg, Alice Moon, Sigal Barsade, Christophe van den Bulte, Uri Simonsohn, and Deb Small.

On a personal note, I am grateful for my family (Veronica Kirgios, Christos Kirgios, Claudia Polini, Teresa Belli, and Eftixia Kirgios), my fiancé (Teddy Terezis), and my incredible friends (Tiffani Chanroo, Maddie Ziegler, Camilla Damonte, Alice Damonte, Jessica Reed, Dora Chen, Sadiki Wiltshire, Charlotte Williams, Lulu Zhong, and Ann Lites). Your love kept me steady. Thank you.

Finally, thank you to the Wharton Behavioral Lab, the Wharton Risk Center Ackoff Doctoral Student Fellowship, the Wharton Leadership Center, the Mack Institute for Innovation Management, the National Science Foundation Graduate Research Fellowship, and the Operations, Information, and Decisions Department for providing financial support for this work.

ABSTRACT

BRINGING IDENTITY TO THE FOREFRONT

Erika L. Kirgios

Supervisor: Katherine L. Milkman

Prior research overwhelmingly shows that when information about an individual's marginalized identity is communicated inadvertently (via a name that signals gender or race, for example), that information tends to trigger prejudiced behavior. As a result, both conventional wisdom and extant research suggest that women and racial minorities should obscure or de-emphasize their minority status to reduce their likelihood of experiencing discrimination. In this work, I propose that women and racial minorities might instead benefit from strategically *emphasizing* their demographic identity. This approach has two potential benefits: when a person's marginalized identity is made more salient, (1) the potential for discrimination on the basis of that identity is also more salient, so decision-makers may be more likely to avoid prejudiced behavior and (2) it highlights an opportunity to support marginalized people, which may appeal to those who want to engage in pro-diversity behaviors. I also investigate whether and why marginalized people strategically choose teams to emphasize their identity, and how organizations might leverage these insights to motivate pro-diversity behavior in their employees. In Chapter 1, I share evidence from two audit experiments—one with politicians and another with students—as well as an online experiment showing that women and racial minorities benefit from explicitly mentioning their demographic identity in requests for help (e.g., by including statements like "As a Black woman. . . "). Politicians and students responded 24.4% and 79.6% more often, respectively, when help-seeking emails included an explicit mention of the sender's marginalized identity. In Chapter 2, I find that when women and racial minorities expect to compete for a job or promotion, they're more willing to be tokens because they think standing out based on their

demographic identity will be strategically beneficial, suggesting that they intuit the benefits of highlighting identity that I establish in Chapter 1. In Chapter 3, I build on these insights in an audit experiment exploring how feedback about either discriminatory or pro-diversity behaviors in one's professional ingroup influences subsequent prejudice. Returning to the population of city councilors in Study 1, I first deliver PSAs with negative feedback (evidence that city councilors discriminate against Black constituents) or positive feedback (evidence that city councilors support Black constituents who emphasize their identity) then measure subsequent response rates to Black vs. White male help-seekers. Receiving negative feedback does not influence responsiveness to Black men relative to receiving no feedback, but positive feedback induces a regression-estimated 36.3% increase in city councilors' response rates to Black men. Positive feedback seems to create new descriptive norms for pro-diversity behavior that councilors are motivated to maintain. Together, my dissertation illuminates the previously unexplored benefits of strategically highlighting marginalized identity, diversity, and bias, suggesting that women and racial minorities don't always need to obscure or hide their identity to succeed.

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

INTRODUCTION

Prior research has documented manifold challenges women and racial minorities face because of bias and discrimination. First, when someone's marginalized identity is communicated inadvertently (via a name or photo, for example), it can trigger prejudicial behavior, making it harder for women and racial minorities to acquire their dream job, buy a car for a fair price, rent a home, secure financial backing for a start-up idea, or even get career advice (Bertrand & Mullainathan, 2004; Ahmed & Hammarstedt, 2008; Hanson & Hawley, 2011; Ayres & Siegelman, 1995; Brooks, Huang, Kearney, & Murray, 2014; Milkman, Akinola, & Chugh, 2015). Identity-based stereotypes color the expectations teachers and managers have for marginalized group members, which can damage their performance and constrain their behavior (Glover, Pallais, & Pariente, 2017; Carlana, 2019; Correll, 2004; Steele & Aronson, 1995). On the job, marginalized group members are often excluded from networks of power and influence, particularly when they're numeric minorities in the office (Watkins, Simmons, & Umphress, 2019; Cullen & Perez-Truglia, 2020; Mehra, Kilduff, & Brass, 1998; Ibarra, 1992). Even when organizations attempt to employ identity-conscious practices, like organizing diversity initiatives or implementing promotion quotas, minoritized group members often face stigma and judgment as their colleagues categorize them as "diversity hires" (Heilman, 1994; Leslie, Mayer, & Kravitz, 2014). In sum, the picture painted by existing literature is bleak: Conveying your marginalized identity to others is limiting and serves as a barrier to success and happiness.

In my dissertation, I present a collection of work that challenges this narrative, proposing that women and racial minorities can and do strategically use their marginalized identities to their advantage. Importantly, I don't mean to suggest that in-group bias, prejudice, and stereotyping don't have a terrible impact. Discriminatory behaviors are harmful to marginalized group members, in work and in life. Instead, I propose that in certain situations, marginalized group members (and organizations) can use identity-based strategies to counteract prejudicial behavior.

My theorizing draws on the idea that people generally want to signal, to themselves and others, that they're fair, moral, and believe in equality—and even those who do not actually endorse egalitarianism may fear social sanction if others judge them to be sexist or racist (Plant & Devine, 1998; Apfelbaum, Sommers, & Norton, 2008; Plant & Devine, 2009). As a result, I propose that when concerns about diversity and discrimination are salient, people are more likely to promote the success of marginalized group members in order to avoid feeling or seeming prejudiced. Early evidence indicates that others' wish to appear anti-discriminatory can indeed advantage minoritized groups. For example, in firms that prioritize diversity, high-achieving women are paid more relative to similarly-achieving men because of a so-called "diversity premium" (Leslie, Manchester, & Dahm, 2017). And people hire more women when diversity concerns are made salient by asking people to hire multiple group members at once rather than one group member at a time (Chang, Kirgios, Rai, & Milkman, 2020). I suggest that marginalized group members can and do leverage this insight.

This theorizing makes predictions that the current literature does not. Conventional wisdom and existing scholarship suggest that women and racial minorities should obscure or de-emphasize their minority status to reduce their likelihood of experiencing discrimination (Kang, DeCelles, Tilscik, & Jun, 2016; Goldin & Rouse, 2000). And research on social categorization suggests that when identity is more salient, it heightens in-group biases and may increase discriminatory behavior (Brewer, 2007). So, these findings imply that marginalized group members would be wise to obscure their identity and avoid situations that emphasize it. Meanwhile, my work suggests that *explicitly highlighting* their demographic identity can improve outcomes for women and racial minorities. This approach has two potential benefits: When a person's marginalized identity is made more salient, (1) the potential for discrimination on the basis of that identity is also more salient so decision-makers may be more likely to avoid prejudiced behavior and (2) it highlights an opportunity to support a woman or racial minority,

which may appeal to those who want to engage in pro-diversity behaviors. Consistent with this theorizing, I demonstrate that women and racial minorities can (and expect to) reap strategic benefits from making their demographic identity more (rather than less) salient.

In Chapter 1, I provide evidence that when seeking help, women and racial minorities benefit from highlighting their marginalized identity. I propose that when women and racial minorities explicitly mention their demographic identity in requests for help (e.g., by including statements like "As a Black man…" or "As a woman…" in their communications), it activates prospective helpers' motivations to avoid prejudiced reactions, ultimately increasing the likelihood that they provide support. I find evidence in support of this theorizing in a preregistered audit experiment with politicians (N = 2,476) and an audit experiment with undergraduate students (N = 1,169). Specifically, when women and racial minorities mentioned their demographic identity in help-seeking emails, politicians and undergraduates responded 24.4% and 79.6% more often, respectively. I replicated this effect in a preregistered online experiment (N = 1,503) and found evidence that mentioning demographic identity activates prospective helpers' internal motivation to respond without prejudice, which in turn increases their willingness to help.

In Chapter 2, I explore whether women and racial minorities strategically seek to stand out based on their demographic identity. Typically, women and racial minorities prefer to affiliate with people who resemble them demographically—both due to similarity attraction and a desire to avoid being tokens. However, I propose when they expect intra-group competition, women will be more willing to join all-male groups and racial minorities will be more willing to join all-White groups. Across six preregistered experiments (N=2,738), I show that marginalized people are more willing to be in the numeric minority when choosing colleagues against whom they will compete for jobs, promotions, and bonuses. This seems to be driven by strategic decision-making:

Women and racial minorities expect that being distinct (on the basis of their identity) will lead their performance to stand out, increasing their likelihood of success.

In Chapter 3, I find that people are more willing to help racial minorities after being informed that their professional ingroup has engaged in pro-diversity behaviors, but not after being informed that their professional ingroup has engaged in discriminatory behaviors. Prior theorizing about prejudice reduction suggests that people are motivated to improve their behavior towards racial minorities when discrepancies between their egalitarian values and their actions are made salient (Monteith, 1993; Monteith, Ashburn-Nardo, Voils, & Czopp, 2002; Pope, Price, & Wolfers, 2018). However, those moments of discrepancy can also be deeply ego threatening and may lead people to dismiss the information rather than learn from it (Eskreis-Winkler & Fishbach, 2019; Levy & Maaravi, 2018). Instead, I propose that highlighting pro-diversity behavior may activate consistency and conformity motives and motivate continued support for racial minorities (Fishbach & Dhar, 2005; Mullen & Monin, 2016; Goldstein, Cialdini, & Griskevicius, 2008). I find evidence consistent with this theorizing in a two-stage audit experiment with city councilors (N = 3,981). Councilors first received emails from a research lab informing them about evidence that city councilors discriminate against Black men (negative feedback) or about evidence that city councilors support Black men who emphasize their identity (positive feedback). Councilors in a no feedback control condition received no initial email. Then, I measured their response rates to help-seeking (fictitious) Black and White male students in an audit experiment. Receiving negative feedback did not influence response rates to Black male help seekers relative to receiving no feedback. However, receiving positive feedback increased response rates to Black male help seekers by 36.3%. Heterogeneity analyses suggest these results may be driven by a desire to conform to group-level pro-diversity norms.

Together, my dissertation finds that women and racial minorities make strategic—and often effective—decisions about how and when to emphasize their demographic identity. Furthermore, my work adds to the literature on prejudice and discrimination by highlighting that people are able to avoid prejudicial behavior—if they are made aware that prejudice might influence their decision-making. However, I also demonstrate that it matters *how* people are made aware that identity might influence their decision-making. Emphasizing moments in which people have failed to avoid discrimination—suggesting they used information about identity to *perpetuate* bias—does not effectively motivate prejudice reduction efforts because ego threat leads people to discount the feedback. Instead, spotlighting moments in which people have successfully supported marginalized group members—suggesting they used information about identity to *mitigate* bias—can motivate prejudice reduction efforts as people seek to keep up with pro-diversity norms set by past good behavior.

By identifying the benefits of highlighting marginalized identities, I provide insights about potential interventions that can improve outcomes for women and racial minorities. In particular, rather than using interventions that obscure identity (i.e., identity-blind evaluation, "Whitened" resumes), I suggest that in certain situations, women and racial minorities should act in ways that emphasize their identity (Kang et al., 2016; Goldin & Rouse, 2000). Organizations, too, might be able to create hiring, promotion, and evaluation processes that make diversity and/or the potential for discrimination more salient and improve outcomes for women and racial minorities (Chang et al., 2020).

References

Ahmed, A. M., & Hammarstedt, M. (2008). Discrimination in the rental housing market: A field experiment on the Internet. *Journal of Urban Economics*, *64*(2), 362-372.

Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *Journal of personality and social psychology*, *95*(4), 918.

Ayres, I., & Siegelman, P. (1995). Race and gender discrimination in bargaining for a new car. *The American Economic Review*, 304-321.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, *94*(4), 991-1013.

Brewer, M. B. (2007). The importance of being we: Human nature and intergroup relations. *American psychologist*, *62*(8), 728.

Brooks, A. W., Huang, L., Kearney, S. W., & Murray, F. E. (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences*, *111*(12), 4427-4431.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, *134*(3), 1163-1224.

Chang, E. H., Kirgios, E. L., Rai, A., & Milkman, K. L. (2020). The isolated choice effect and its implications for gender diversity in organizations. *Management Science*, *66*(6), 2752-2761.

Correll, S. J. (2004). Constraints into preferences: Gender, status, and emerging career aspirations. *American sociological review*, *69*(1), 93-113.

Cullen, Z. B., & Perez-Truglia, R. (2019). *The Old Boys' Club: Schmoozing and the Gender Gap* (No. w26530). National Bureau of Economic Research.

Eskreis-Winkler, L., & Fishbach, A. (2019). Not learning from failure—The greatest failure of all. Psychological science, 30(12), 1733-1744.

Fishbach, A., & Dhar, R. (2005). Goals as excuses or guides: The liberating effect of perceived goal progress on choice. Journal of Consumer Research, 32(3), 370-377.

Glover, D., Pallais, A., & Pariente, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *The Quarterly Journal of Economics*, *132*(3), 1219-1260.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American economic review*, *90*(4), 715-741.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. Journal of consumer Research, 35(3), 472-482.

Hanson, A., & Hawley, Z. (2011). Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in US cities. *Journal of urban Economics*, *70*(2-3), 99-114.

Heilman, M. (1994). Affirmative action: Some unintended consequences for working women. In *Research in organizational behavior* (pp. 125-169). JAI Press.

Ibarra, H. (1992). Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative science quarterly*, 422-447.

Kang, S. K., DeCelles, K. A., Tilcsik, A., & Jun, S. (2016). Whitened résumés: Race and self-presentation in the labor market. *Administrative Science Quarterly*, *61*(3), 469-502.

Leslie, L. M., Manchester, C. F., & Dahm, P. C. (2017). Why and when does the gender gap reverse? Diversity goals and the pay premium for high potential women. *Academy of Management Journal*, *60*(2), 402-432.

Leslie, L. M., Mayer, D. M., & Kravitz, D. A. (2014). The stigma of affirmative action: A stereotyping-based theory and meta-analytic test of the consequences for performance. *Academy of Management Journal*, *57*(4), 964-989.

Levy, A., & Maaravi, Y. (2018). The boomerang effect of psychological interventions. *Social Influence*, *13*(1), 39-51.

Mehra, A., Kilduff, M., & Brass, D. J. (1998). At the margins: A distinctiveness approach to the social identity and social networks of underrepresented groups. *Academy of Management Journal*, *41*(4), 441-452.

Milkman, K. L., Akinola, M., & Chugh, D. (2015). What happens before? A field experiment

exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology*, *100*(6), 1678.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. Journal of personality and social psychology, 65(3), 469.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: on the development and operation of cues for control. Journal of personality and social psychology, 83(5), 1029.

Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. Annual review of psychology, 67, 363-385.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of personality and social psychology*, *75*(3), 811.

Plant, E. A., & Devine, P. G. (2009). The active control of prejudice: Unpacking the intentions guiding control efforts. *Journal of personality and social psychology*, *96*(3), 640.

Pope, D. G., Price, J., & Wolfers, J. (2018). Awareness reduces racial bias. *Management Science*, *64*(11), 4988-4995.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology*, *69*(5), 797.

Watkins, M. B., Simmons, A., & Umphress, E. (2019). It's not black and white: Toward a contingency perspective on the consequences of being a token. *Academy of Management Perspectives*, *33*(3), 334-365.

CHAPTER 1. WHEN SEEKING HELP, WOMEN AND RACIAL MINORITIES BENEFIT FROM EXPLICITLY STATING THEIR IDENTITY

Erika L. Kirgios, Aneesh Rai, Edward H. Chang, Katherine L. Milkman

Published in *Nature Human Behaviour* in 2022

ABSTRACT:

Receiving help can make or break a career, but women and racial 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 minorities benefit from explicitly mentioning their demographic identity in requests for help (e.g., by including statements like "As a Black woman…" in their communications). We propose that when someone highlights their marginalized identity, it activates prospective helpers' motivations to avoid prejudiced reactions and increases prospective helpers' willingness to provide support. Here we show that, consistent with this theorizing, when marginalized group members 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 minorities.

Link to *Online Supplement*, data, and code*:* https://bit.ly/3zYDjBO.

Introduction

In the United States, women and racial minorities remain underrepresented in many organizational contexts, particularly in leadership positions (Coury et al., 2020; National Center for Education Statistics, 2021). One contributing factor may be that in-group favoritism and bias lead underrepresented group members to receive less instrumental help—advice, feedback, referrals, or assistance on tasks—than White men (Butler & Broockman, 2011; Giuliano, Levine, & Leonard, 2011; Keeves & Westphal, 2021; Lavy & Sand, 2018; McDonald, Keeves, & Westphal, 2018; Milkman, Akinola, & Chugh, 2012; Milkman, Akinola, & Chugh, 2015; Price & Wolfers, 2010; White, Nathan, & Faller, 2015; cf. Kalla, Rosenbluth, & Teele, 2018). Such instrumental help can be critical to career success, especially for members of historically marginalized groups (Eby, Allen, Evans, Ng, & DuBois, 2008; Kaas & Manger, 2012; Seibert, Kraimer, & Liden, 2001). Thus, increasing the rate at which assistance is offered to women and racial 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 (Butler & Broockman, 2011; Milkman, Akinola, & Chugh, 2015; White et al., 2015; Bohren, Imas, & Rosenberg, 2019; Doleac & Stein, 2013; Edelman, Luca, & Svirsky, 2017; Kang, DeCelles, Tilcsik, & Jun, 2016). For instance, Bertrand and Mullainathan (2004) randomly assigned White-sounding or Black-sounding names to otherwise identical resumes and used those resumes to apply for entry-level jobs (Bertrand & Mullainathan, 20014). They found that those with Black-sounding names received 50% fewer callbacks than those with White-sounding names (Bertrand & Mullainathan, 2004). 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 (Milkman, Akinola, & Chugh, 2015; Dovidio & Gaertner, 2000). Together, these findings suggest that marginalized group members might be wise to downplay or even hide their demographic identity when seeking help (Kang et al., 2016).

We propose, however, that marginalized group members may benefit from explicitly stating if they are a woman and/or a racial minority in help requests. When help seekers highlight their marginalized 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, past 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 in order to (1) maintain a positive self-image (by behaving consistently with their personal values), and (2) escape social sanction (by conforming to norms of political correctness or egalitarianism) (Plant & Devine, 1998; Apfelbaum, Sommers, & Norton, 2008; Bodner & Prelec, 2003; Paluck & Green, 2009; Plant & Devine, 2009; Rokeach, 1971). 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 motivation to respond without prejudice will be activated and will increase the likelihood that prospective helpers provide support.

Some prior research suggests that when the potential for prejudice is more salient, people are less likely to behave in a biased manner (Plant & Devine, 2009; Pope, Price, & Wolfers, 2018; Sommers & Ellsworth, 2001). For example, following media coverage of a study demonstrating that White National Basketball Association referees tended to be biased in favor of White players, this in-group bias declined significantly (Pope, Price, & Wolfers, 2018). Making

referees aware of the potential for bias may have helped them counteract it. Similarly, Sommers and Ellsworth (2001) found that when mock juries evaluated cases, White jurors were generally biased against Black defendants (Sommers & Ellsworth, 2001). 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 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 minorities receive the support they seek.

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 councilors are a regression-estimated 7.42 percentage-points (or 24.4%) more likely to respond to help-seeking emails from women and racial 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, consistent with our theorizing, internal

motivation to respond without prejudice mediates the effects of mentioning demographic identity on a prospective helper's responsiveness to requests for assistance.

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 (e.g., 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 (Butler & Broockman, 2011; Milkman, Akinola, & Chugh, 2015; Kang et al., 2016; Bertrand & Mullainathan, 2004; Banaji & Hardin, 1996; Devine, 1989; Taylor, Fiske, Etcoff, & Ruderman, 1978). 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.

Results

Study 1: Audit Experiment with City Councilors. Participants were 2,476 White male city councilors serving in cities across the U.S. Each city councilor received an email from a (fictitious) student requesting career advice (following a design similar to that used in Kalla, Rosenbluth, & Teele, 2018). The emails were identical across conditions except for two randomized elements: (1) whether the help-seeking student was a White male help seeker (we'll hereafter refer to this as the White male help seeker condition; see Supplement Table 1 for information about the help seeker's names, which were used to manipulate identity) or a minority help seeker (i.e., a White female, Black male, Black female, Latino, or Latina; we'll hereafter refer to this 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"; we'll hereafter refer to this as the identity mentioned condition) or not (instead calling themselves a "young person"; we'll hereafter refer to this as the identity not mentioned condition). Supplement Table 2 includes participant summary statistics, and balance checks presented in Supplement Table 3 confirm that there were no significant imbalances across experimental conditions on any observables.

Our preregistered dependent variable of interest was whether a city councilor replied to our email within one week. Following our preregistration, automatic replies and replies from aides or assistants–as opposed to the city councilor–were counted as non-responses. As Figure 1 shows, city councilors 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. The difference in response rates to White men across the identity not mentioned and identity mentioned conditions was not statistically significant (two-sample, two-tail proportions test: $z = 0.870$, $p = .384$, effect size $h = -0.050$, 95% CI: [-0.074, 0.029]). However, city councilors replied to emails from women and racial 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: $z = 2.89$, $p = .004$, effect size $h = 0.164$, 95% CI: [0.025, 0.130]). Figure 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 councilors 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 councilor's region, the city's population size, the city councilor's political party, years until the city councilor's next re-election, and the city councilor's current position (whether or not they had recently been replaced or stepped down). Complete regression results for this analysis are included in Supplement Table 4. Given that our outcome variable is binary, our data violates both normality and homoskedasticity assumptions. Despite these violations, we analyze 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 (Ai & Norton, 2003; Gomila, 2020). Moreover, in Supplement Table 5 we present the results of our primary analysis with a logistic regression rather than an OLS regression (further robustness checks presented in Supplement Tables 6-7 (a) remove any city councilors who had been replaced or stepped down and (b) include replies to our emails that arrived within 7 weeks rather than only replies received within 1 week). Our Supplement also contains further details about the covariates included in our primary regression (in Sections 1a and 1c 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, SE = 0.038, 95% CI [0.024, 0.171]; p = .010; see Table 1, Model 1 for full regression results). This result is robust to the removal of our preregistered covariates (b = 0.100, SE = 0.038, 95% CI [0.027, 0.174]; p = .007) and to analyzing our data using a logistic regression instead of an OLS regression (b = 0.441, SE = 0.173, 95% CI [0.101, 0.782]; p = 0.011; see Supplement Tables 4 and 5 for full regression results). There was no statistically significant main effect of assignment to the identity mentioned condition (b = -0.023, SE = 0.026, 95% CI: [-0.075, 0.029]; p = .380) and no statistically significant main effect of assignment to the minority help seeker condition (b

= -0.010, SE = 0.026, 95% CI: [-0.062, 0.042]; p = .705). In sum, these results show that White male city councilors 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 minorities when the emails city councilors received mentioned the help seeker's demographic identity. The lack of statistically significant discrimination against women and racial minorities overall may be due to our audit study's context: past work finds mixed evidence as to whether local politicians discriminate against women and racial minorities when responding to help requests (Butler & Broockman, 2011; White et al., 2015; Kalla et al., 2018; Butler & Crabtree, 2017; Einstein & Glick, 2017).

None of the exploratory moderators we preregistered significantly moderated our key interaction. These included (1) the city councilor'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 Supplement Tables 9 -13 for full regression results).

In exploratory analyses that were not pre-registered, we examined the quality of help city councilors 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 Supplement Section 1f for details). Specifically, we examined (1) whether the city councilor provided specific advice to the student (15.8% did; interrater ICC(3,3) = 0.96); (2) whether the city councilor suggested scheduling a call or a meeting (18.4% did; interrater ICC(3,3) = 0.76); and (3) whether the city councilor offered a work or volunteer opportunity (5.1% did; interrater ICC(3,3) = 0.76). We also examined the length of each city councilor's response message. Following Kalla, Rosenbluth, & Teele, 2018, we operationalized length of response both by calculating (1) the log word count of the city councilor's reply (mean = 1.371; S.D. = 2.080) and

(2) the log character count of the city councilor's reply (mean = 1.900; S.D. = 2.829). 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 2-3 and Supplement Table 14 for full regression results).

We find that city councilors wrote significantly more words (a regression-estimated 31.8% more) and significantly more characters (a regression-estimated 46.6% more) in response to women and racial minorities when their emails mentioned their demographic identity (word count regression: interaction b = 0.390, SE = 0.167, 95% CI [0.062, 0.717]; p = .020; see Table 1, Model 2; character count regression: interaction b = 0.530, SE = 0.227, 95% CI [0.085, 0.975]; p = .020; see Table 1, Model 3). No other measures of response quality differed significantly for women and racial minorities who mentioned their identity and those who did not: the interaction between assignment to the minority help seeker condition and the identity mentioned condition was not statistically significant for regressions predicting the likelihood that city councilors offered specific advice (interaction b = 0.040, SE = 0.029, 95% CI [-0.018, 0.097]; p = .177; see Supplement Table 14, Model 1), suggested scheduling a meeting (interaction b = 0.057, SE = 0.031, 95% CI [-0.004, 0.118]; p = .067; see Supplement Table 14, Model 2), and offered work or volunteer opportunities (interaction b = 0.012, SE = 0.018, 95% CI [-0.023, 0.046]; p = .506; see Supplement Table 14, Model 3).

Taken together, this indicates that when women and racial minorities mentioned their demographic identity in requests for help, they (a) received more and longer replies and (b) the quality of those replies did not change significantly.

Study 2: Audit Experiment with Undergraduate Students. Study 2 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 behavioral lab participant pool at an East Coast university.

All Study 2 participants received an email from the behavioral 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 Supplement Section 2b). 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 Supplement Table 15, and balance checks presented in Supplement Table 16 confirm that there was no significant 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: $z = 2.17$, $p = .030$, 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 (Barron & Yechiam, 2002). There is also a norm of paying behavioral lab participants for

their participation in research and Demarcus did not offer compensation for help, whereas there is no norm of paying city councilors to respond to constituent emails.

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, and political ideology (measured on a seven-point Likert scale from "Very liberal" to "Very conservative"). These control variables were provided by the behavioral lab and were collected when each undergraduate in our study first signed up to participate in behavioral lab research (see Supplement 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 primarily analysis is an OLS regression (following Study 1) because it is the recommended method for estimating treatment effects on binary outcomes in experiments (Gomila, 2020).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 ($b = 0.027$, $SE = 0.013$, 95% CI [0.003, 0.052]; $p = .029$; see Supplement Table 17 for complete regression results and Supplement Table 18 for regression results relying on a logistic regression rather than an ordinary least squares regression model). This means students volunteered to assist Demarcus 79.6% more when he mentioned his demographic identity in his request for help (see Figure 3). Furthermore, this result is robust to the removal of our covariates ($b = 0.027$, $SE = 0.012$, 95% CI [0.003, 0.052]; $p = 0.030$).

The effect of the identity mentioned condition again did not vary significantly as a function of participant characteristics, including their gender, race, political ideology, or age (see Supplement 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 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 (#1) to the one they were least likely to refer (#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 & Devine, 1998; see Supplement Section 3b 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 significantly 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: $t(1498) = 5.06$, $p < .001$, Cohen's $d = 0.261$, 95% CI: [0.157, 0.357]). The ranking data was 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 ($z(1498) = -5.12$, $p < 0.001$, Cliff's delta=-0.936, 95% CI for the delta estimate = [0.926, 0.944]). Participants in the identity mentioned condition were also significantly more likely to choose to refer the Black male student (by ranking him 1st) than participants in the identity not mentioned condition (15.1% vs. 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 Supplement Figure 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 provide correlational evidence of potential mechanisms that might be responsible for the effects documented. However, we also note that there are inherent weaknesses to mediation analysis with measured rather than manipulated mediators (Bullock, Gren, & Ha, 2010). Specifically, causal mediation analysis relies on the Sequential Ignorability Assumption, which states, in part, that no omitted pre-treatment covariates are correlated with both the mediator and the outcome (Imai, Keele, & Tingley, 2010). To address this issue, we conduct sensitivity analyses developed by Imai, Keele, & Yamamoto (2010) to assess the robustness of our findings to deviations from the Sequential Ignorability Assumption (Imai, Keele, & Yamamoto, 2010). The sensitivity parameter is rho, 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 (rho 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 confidence interval 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 significant ($b = -0.079$, $SE = 0.016$, $p < .001$). In particular, Imai, Keele, & Yamamoto's (2010) 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. 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 confidence interval 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 the reduction in effect size was not statistically significant ($b = -0.010$, $SE = 0.006$, $p = .083$). 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 (Imai, Keele, & Yamamoto, 2010). 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 ($r = 0.612$; $p < .001$). To address this multicollinearity issue, we ran a preregistered multiple mediation model. In this model, the 95% confidence interval 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 predictor 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 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 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 resumes with Black versus White names identified in past research was 3.2 percentage-points (Butler & Broockman, 2011; Bertand & Mullainathan, 2004).

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 suggestive evidence that mentioning demographic identity activates prospective helpers' internal motivation to respond without prejudice, which, in turn, is correlated with an increased willingness to help marginalized identity group members. These results suggest that women and racial minorities stand to reap important benefits if they mention their demographic identity in help requests.

Making the potential for poor judgment 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 (Fishbane, Ouss, & Shah, 2020; Tiefenbeck et al., 2018; Zhang, Fletcher, Gino, & Bazerman, 2015). Drawing timely attention to the risk of exhibiting prejudice may have an important and underappreciated impact on decision making that is worthy of further study and theorizing.

Our work also suggests that features of a decision-making environment can increase 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 (Plant & Devine, 1998; Butz & Plant, 2009). That is, people become more motivated to overcome their prejudice when women and minorities highlight their identity in requests for help. 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 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 behavioral 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 (e.g., Bertrand & Mullainathan, 2004). The fact that we find consistent results across Studies 1 and 2 suggests that our findings are not dependent on prospective helpers' expectation that a help seeker would observe their response.

Moreover, our findings offer suggestive evidence that decision makers' internal motivation to control prejudice—and not external motivation—is related to improved outcomes for women and racial minorities who mention their demographic identity. External motivation to control prejudice may have been less influential 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-signaling 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 impact 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 (e.g., mentorship), or seek other outcomes (e.g., 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 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 focus 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 minority help seekers who highlight their identity and don't 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 minority requesting help that explicitly mentions the sender's identity, is that recipient more likely to help other women and racial minorities who reach out subsequently?

Women and racial minorities have long been left out of positions of power, held back by negative stereotypes, prejudice, tokenism, and in-group favoritism (Price & Wolfers, 2010; Glover, Pallais, & Pariente, 2017; Heilman, 2001; Ibarra, 1992; Rosette, Leonardelli, & Phillips, 2008; Watkins, Simmons, & Umphress, 2019). Time and again, evidence has shown that when information about an individual's marginalized identity is communicated inadvertently, it limits women and racial minorities' opportunities (Milkman, Akinola, & Chugh, 2015; Edelman, Luca, & Svirsky, 2017; Kang et al., 2016; Bertrand & Mullainathan, 2004). In this work, however, we demonstrate that when women and racial 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 ($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 (OSF) at https://bit.ly/3zYDjBO.

Studies 1, 2, and 3 were preregistered on July 8th, 2020; September 18th, 2020; and November 10th, 2020, respectively. The OSF folder also includes our Supplement, 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 Councilors. Study 1 tested our hypothesis in a preregistered email audit experiment. Participants were 2,476 White male city councilors from 701 of the largest cities in the United States (by population, based on 2019 Census data; see Supplement Table 2 for participant summary statistics).[50] No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications.[3,11,12,35,36] A team of research assistants inferred councilors' gender and race from publicly available information, including their names, photographs, personal bios, and news articles. Each city councilor's gender and race were classified by two research assistants, and any disagreements were resolved by the first author of this manuscript. City councilors' 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 councilor in our study received an email from a (fictitious) student on the morning of July 14, 2020. The email stated that the student had dreams of a career in politics and asked the city councilor 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 councilors to conditions was stratified by their city to ensure balance on this dimension. 621 city councilors were assigned to the minority help seeker x identity mentioned condition, 620 city councilors were assigned to the White male help seeker x identity mentioned condition, 625 city councilors were assigned to the minority help seeker x identity not mentioned condition, and 610 city councilors 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 signaled the student's gender and race.[9,12,20] Names were chosen to signal one of six demographic identities: White male, White female, Black male, Black female, Latino, or Latina. We selected four names for each race-gender combination of interest, or 24 names total (all names can be found in Supplement Table 1, along with information about how they were selected in Supplement Section 1b). City councilors 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 (i.e., 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 councilor to share advice with "a young [person]/[man/woman/Black man/Black woman/Latino/Latina] hoping to become a city councilor." In the identity not mentioned condition, the student made no mention of their identity

and asked the city councilor to share advice with "a young person." By contrast, in the identity mentioned condition, the help seeker asked the city councilor 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 (i.e., by asking city councilors to share advice with a "young White man/woman") because qualitative data suggested that by labeling themselves explicitly as "White", senders might signal White nationalist political attitudes.

Complete study stimuli and further details about our methods and results are available in our Supplement Section 1.

Study 2: Audit Experiment with Undergraduate Students. Our participants were 1,169 undergraduate members of the behavioral 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 = 19.8 years old). We used G*Power to calculate the sample size we would need to detect an effect size similar to that of the identity mentioned condition for women and racial minorities in Study 1 ($h = 0.164$) with 80% power. The result was 1,162. In order to fulfill this required sample size, we contacted all undergraduate members of the behavioral lab's participant pool.

The behavioral lab sent an email to active members of its undergraduate participant pool on September 23, 2020 with the subject line "Request for Research Help." The email explained that the behavioral lab was forwarding a request for free research help from a Ph.D. student named Demarcus Rivers (a fictitious student whose name was selected to signal a Black male identity; see Supplement Section 2b 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 Behavioral Lab) to ensure balance on these dimensions. 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. 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 our Supplement in Tables 23 and 24.

Complete study stimuli and further details about our methods and results are in our Supplement in Sections 2 and 7.

Study 3: Online Experiment. We recruited 1,500 participants (48.4% female; 73.3% White) through Prolific to participate in a preregistered 7-minute study in exchange for $0.80. We

did not collect data on participants' age for this study because our IRB recommends 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. 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. In order to comply with our preregistration, we excluded the three participants who completed the study last. All of our results are consistent when we include 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 signaled their gender and race. All participants read two emails from White men (Brad Miller and Todd Anderson; see Supplement Section 3a for details on how names were selected for this study), 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 (#1) to the one they were least likely to refer

(#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 7-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 (1998) internal motivation to respond without prejudice scale (e.g., "Given my personal values and beliefs, an important factor in my decision was my desire to promote the success of racial 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 (1998) external motivation to respond without prejudice scale (e.g., "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 (e.g., White women, White men, Black women, Black men, etc.) they recalled requesting a referral (as a manipulation check). Finally, participants reported their own gender and race. Further study details are included in Supplement Section 3 and all study stimuli and scale items are included in Supplement Section 8.

References

Ai, C., & Norton, E. C. Interaction terms in logit and probit models. *Economics letters*, *80*(1), 123-129 (2003).

Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *Journal of personality and social psychology*, *95*(4), 918 (2008).

Banaji, M. R., & Hardin, C. D. Automatic stereotyping. *Psychological science*, *7*(3), 136-141 (1996).

Barron, G., & Yechiam, E. Private e-mail requests and the diffusion of responsibility. *Computers in Human Behavior*, *18*(5), 507-520 (2002).

Bertrand, M., & Mullainathan, S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, *94*(4), 991-1013 (2004).

Bodner, R., & Prelec, D. Self-signaling and diagnostic utility in everyday decision making. *The psychology of economic decisions*, *1*(105), 26 (2003).

Bohren, J. A., Imas, A., & Rosenberg, M. The dynamics of discrimination: Theory and evidence. *American economic review*, *109*(10), 3395-3436 (2019).

Bullock, J. G., Green, D. P., & Ha, S. E. Yes, but what's the mechanism?(don't expect an easy answer). *Journal of personality and social psychology*, *98*(4), 550 (2010).

Butler, D. M., & Broockman, D. E. Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science*, *55*(3), 463-477 (2011).

Butler, D. M., & Crabtree, C. Moving beyond measurement: Adapting audit studies to test bias-reducing interventions. *Journal of Experimental Political Science*, *4*(1), 57 (2017).

Butz, D. A., & Plant, E. A. Prejudice control and interracial relations: The role of

motivation to respond without prejudice. *Journal of Personality*, *77*(5), 1311-1342 (2009).

Coury, S., Huang, J., Kumar, A., Prince, S., Krivkovich, A., & Yee, L. Women in the Workplace 2020. https://www.mckinsey.com/featured-insights/diversity-and-inclusion/women-in-the-workplace (2020, October 08).

Devine, P. G. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, *56*(1), 5 (1989).

Doleac, J. L., & Stein, L. C. The visible hand: Race and online market outcomes. *The Economic Journal*, *123*(572), F469-F492 (2013).

Dovidio, J. F., & Gaertner, S. L. Aversive racism and selection decisions: 1989 and 1999. *Psychological science*, *11*(4), 315-319 (2000).

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. *Journal of vocational behavior*, *72*(2), 254-267 (2008).

Edelman, B., Luca, M., & Svirsky, D. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, *9*(2), 1-22 (2017).

Einstein, K. L., & Glick, D. M. Does race affect access to government services? An experiment exploring street-level bureaucrats and access to public housing. *American Journal of Political Science*, *61*(1), 100-116 (2017).

Fishbane, A., Ouss, A., & Shah, A. K. Behavioral nudges reduce failure to appear for court. *Science*, *370*(6517) (2020).

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. *Journal of Human Resources*, *46*(1), 26-52 (2011).

Glover, D., Pallais, A., & Pariente, W. Discrimination as a self-fulfilling prophecy:

Evidence from French grocery stores. *The Quarterly Journal of Economics*, *132*(3), 1219-1260 (2017).

Gomila, R. Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General* (2020).

Heilman, M. E. Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of social issues*, *57*(4), 657-674 (2001).

Ibarra, H. Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative science quarterly*, 422-447 (1992).

Imai, K., Keele, L., & Tingley, D. A general approach to causal mediation analysis. *Psychological methods*, *15*(4), 309 (2010).

Imai, K., Keele, L., & Yamamoto, T. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, *25*(1), 51-71 (2010).

Kaas, L., & Manger, C. Ethnic discrimination in Germany's labour market: A field experiment. *German economic review*, *13*(1), 1-20 (2012).

Kalla, J., Rosenbluth, F., & Teele, D. L. Are you my mentor? A field experiment on gender, ethnicity, and political self-starters. *The Journal of Politics*, *80*(1), 337-341 (2018).

Kang, S. K., DeCelles, K. A., Tilcsik, A., & Jun, S. Whitened résumés: Race and self-presentation in the labor market. *Administrative Science Quarterly*, *61*(3), 469-502 (2016).

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 Science* (2021).

Lavy, V., & Sand, E. On the origins of gender gaps in human capital: Short-and long-term consequences of teachers' biases. *Journal of Public Economics*, *167*, 263-279 (2018).

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. *Academy of Management Journal*, *61*(2), 405-439 (2018).

Milkman, K. L., Akinola, M., & Chugh, D. Temporal distance and discrimination: An audit study in academia. *Psychological science*, *23*(7), 710-717 (2012).

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. *Journal of Applied Psychology*, *100*(6), 1678 (2015).

National Center for Education Statistics. The NCES Fast Facts Tool provides quick answers to many education questions. Retrieved January 29, 2021, from https://nces.ed.gov/fastfacts/display.asp?id=61 (2020).

Paluck, E. L., & Green, D. P. Prejudice reduction: What works? A review and assessment of research and practice. *Annual review of psychology*, *60*, 339-367 (2009).

Plant, E. A., & Devine, P. G. Internal and external motivation to respond without prejudice. *Journal of personality and social psychology*, *75*(3), 811 (1998).

Plant, E. A., & Devine, P. G. The active control of prejudice: Unpacking the intentions guiding control efforts. *Journal of personality and social psychology*, *96*(3), 640 (2009).

Pope, D. G., Price, J., & Wolfers, J. Awareness reduces racial bias. *Management Science*, *64*(11), 4988-4995 (2018).

Price, J., & Wolfers, J. Racial discrimination among NBA referees. *The Quarterly Journal of Economics*, *125*(4), 1859-1887 (2010).

Rokeach, M. Long-range experimental modification of values, attitudes, and behavior. *American psychologist*, *26*(5), 453 (1971).

Rosette, A. S., Leonardelli, G. J., & Phillips, K. W. The White standard: racial bias in

leader categorization. *Journal of Applied Psychology*, *93*(4), 758 (2008).

Seibert, S. E., Kraimer, M. L., & Liden, R. C. A social capital theory of career success. *Academy of management journal*, *44*(2), 219-237 (2001).

Sommers, S. R., & Ellsworth, P. C. White juror bias: An investigation of prejudice against Black defendants in the American courtroom. *Psychology, Public Policy, and Law*, *7*(1), 201 (2001).

Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. J. Categorical and contextual bases of person memory and stereotyping. *Journal of personality and social psychology*, *36*(7), 778 (1978).

Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., & Staake, T. Overcoming salience bias: How real-time feedback fosters resource conservation. *Management science*, *64*(3), 1458-1476 (2018).

Watkins, M. B., Simmons, A., & Umphress, E. It's not black and white: Toward a contingency perspective on the consequences of being a token. *Academy of Management Perspectives*, *33*(3), 334-365 (2019).

White, A. R., Nathan, N. L., & Faller, J. K. What do I need to vote? Bureaucratic discretion and discrimination by local election officials. *American Political Science Review*, 129-142 (2015).

Zhang, T., Fletcher, P. O., Gino, F., & Bazerman, M. H. Reducing bounded ethicality: How to help individuals notice and avoid unethical behavior. *Organizational Dynamics* (2015).

Tables

Table 1, Chapter 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 Minority Help Seeker	-0.010	[-0.062, 0.042]	.705	-0.070	[-0.301, 0.162]	.554	-0.101	[-0.416, 0.214]	.528
Identity Mentioned	-0.023	[-0.075, 0.029]	.380	-0.114	[-0.346, 0.118]	.327	-0.148	[-0.463, 0.168]	.350
Female and/or Racial Minority Help Seeker * Identity Mentioned	0.097	[0.024, 0.171]	.010	0.390	[0.062, 0.717]	.020	0.530	[0.085, 0.975]	.020
Observations	2476			2476			2476		
Adjusted R^2	0.007			0.006			0.007		

Note. This table reports the results of six ordinary least squares (OLS) regression models. The first regression model predicts whether a given city councilor 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 councilor 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 councilor received (we stimulus sampled by testing three similar emails requesting help), the log-transformed population size of the city councilor's city, a binary indicator for whether the city councilor is a Democrat, a binary indicator for whether the city councilor is a Republican, a continuous variable for the number of years until the city councilor faces re-election (0 if the participant has been replaced), and a binary indicator for whether the city councilor was replaced in 2020 just prior to our experiment. We also include fixed effects for the city councilor's region of the country, as determined by the U.S. Census (Northeast, Midwest, South, and West). Robust standard errors are reported in parentheses.

†, *, **, and *** denote significance at the 10%, 5%, 1%, and 0.1% levels, respectively.

Figures

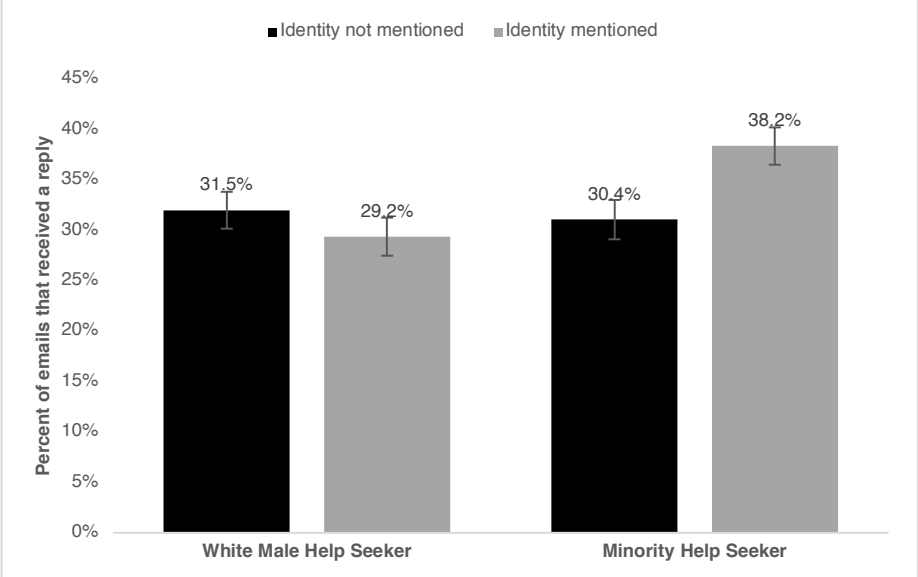


Figure 1, Chapter 1 | Reply rates to emails across conditions in Study 1. This figure depicts White male city councilors' (N = 2,476) response rates to help-seeking emails from fictitious students in Study 1. The two bars on the left display response rates to emails from help-seeking students whose names signaled that they were White men and the two bars on the right display response rates to emails from help-seeking students whose names signaled that they belonged to a marginalized identity group (i.e., 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 ordinary least squares regression) and Supplementary Table 5 (using a logistic regression).

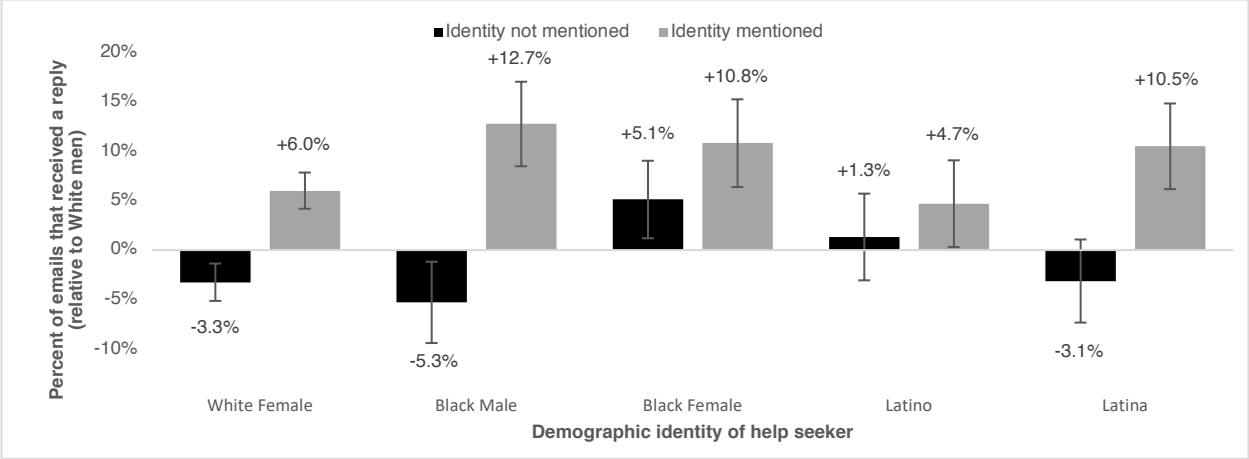


Figure 2, Chapter 1 | Reply rates to emails from women and/or racial minorities (relative to White male help seekers) across conditions in Study 1. This figure displays White male city councilors' (N = 2,476) response rates to emails from women and/or racial minorities seeking help (relative to White 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 ordinary least squares regression) and Supplementary Table 5 (using a logistic regression).

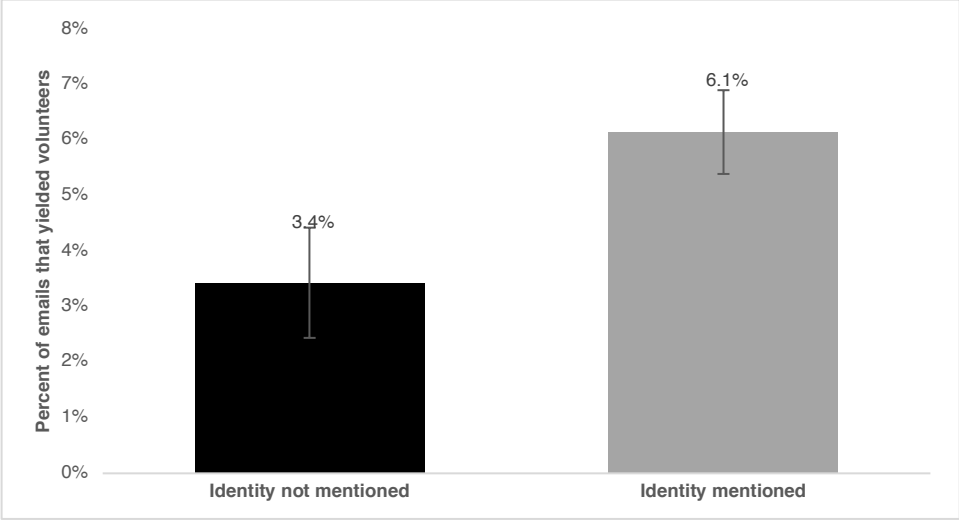


Figure 3, Chapter 1 | Percent of emails that yielded volunteers across conditions in Study 2. This figure displays the percentage of undergraduates (N = 1,169) who volunteered to help a fictitious Black male graduate student with his dissertation research in response to a help-seeking email in Study 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 ordinary least squares regression) and Table 18 (using a logistic regression).

CHAPTER 2: GOING IT ALONE: COMPETITION INCREASES THE ATTRACTIVENESS OF MINORITY STATUS

Erika L. Kirgios, Edward H. Chang, and Katherine L. Milkman

Published in *Organizational Behavior and Human Processes* in 2020

ABSTRACT:

Past research demonstrates that people prefer to affiliate with others who resemble them demographically. However, we posit that when competing for scarce opportunities, strategic considerations moderate the strength of this tendency toward homophily. Across six experiments, we find that anticipated competition weakens people's desire to join groups that include similar others. When expecting to compete against fellow group members, women are more willing to join all-male groups and Black participants are more willing to join all-White groups than in the absence of competition. We show that this effect is mediated by the belief that being distinct will lead your performance to stand out. Our findings offer a new perspective to enrich past research on homophily, shedding light on the instances when minorities are more likely to join groups in which they will be underrepresented.

Link to *Online Supplement*, data, and code:

https://osf.io/j8wnt/?view_only=9d83b69e252b4391b1a6faa01d8c58c7

Introduction

People often have the opportunity to select the groups they'll join at work and beyond. For example, some organizations have internal talent markets or rotational programs such that employees can sample several teams before choosing one to join. In academic environments, students choose between classes, majors, and research groups at their college or university. More commonly, such choices are inter-organizational: for instance, many people choose between job offers, which typically means selecting between work groups.

While research on organizational attractiveness often focuses on how organizational features and individual attitudes interact to shape people's preferences between jobs (Cable & Judge, 1996; Lievens, Decaesteker, Coetsier, & Geirnaert, 2001; Martins & Parsons, 2007; Turban & Greening, 1997; Turban & Keon, 1993), in this paper, we explore how people choose between groups or teams based on their anticipated coworkers. Specifically, we examine how members of historically underrepresented populations choose between work groups based on both organizational context and work group composition, and we offer a theory challenging the idea that underrepresented group members are universally opposed to being tokens[1] (cf. Duguid, 2011; Umphress, Smith-Crowe, Brief, Dietz, & Watkins, 2007). By more closely examining the preferences and choices of members of historically underrepresented populations (namely women and racial minorities), our work contributes to a richer understanding of diversity in organizations.

Most theory and scholarship about why prospective group members are attracted to one group over another is grounded in research on homophily. Homophily is a term that describes our tendency to join groups composed of people whose beliefs, attitudes, and demographic traits

[1] As per Kanter's (1977) definition, we consider tokens to be individuals who constitute less than 15% of their group.

resemble our own (see McPherson, Smith-Lovin, & Cook, 2001 for a review). There is particularly strong evidence of homophily among members of underrepresented populations (Baugher, Varanelli, & Weisbord, 2000; Mehra, Kilduff, & Brass, 1998; cf. Umphress et al., 2007), in part due to the aversive consequences that women and racial minorities face when they are tokens (Cohen & Swim, 1995; Kanter, 1977).

We posit that past research may have overlooked an important moderator of the strength of homophily. Specifically, we focus on the consequences of intra-group competition, or competition against fellow work group members, which is a common feature of organizational life (Scheiber, 2015; Steinhage, Cable, & Wardley, 2017). Work group members frequently compete amongst themselves for promotions, recognition, and bonuses. Any organization with limited opportunities for advancement involves some form of competition against peers, but intra-group competition is particularly common at elite companies, where large numbers of entry-level employees are culled down through consistent cuts until a small number reach senior positions within the firm (Scheiber, 2015).

We theorize that intra-group competition affects which groups women and racial minorities prefer to join by reducing their desire to work with similar others. Competition for scarce recognition gives rise to desires for individuation and differentiation from fellow competitors (Maslach, 1974). Because race and gender are highly salient identities for social categorization (Stangor, Lynch, Duan, & Glas, 1992), the desire to appear different and set oneself apart from competitors may increase the rate at which historically underrepresented minorities in organizations (e.g., female employees, Black employees) prefer to join groups of dissimilar others. In addition, prior work suggests that implicit quotas, which are norms or unstated rules for the number of underrepresented minorities offered jobs or promotions, may dictate whom managers attempt to attract and retain (Chang, Milkman, Chugh, & Akinola, 2019;

Dezső, Ross, & Uribe, 2016). If women and racial minorities expect their managers' decisions to be influenced by implicit quotas, they may strategically choose to be tokens in order to increase their chances of success when facing intra-group competition. Finally, across many domains, competition has been shown to increase people's strategic thinking and focus on social comparisons, and it has been shown to reduce their focus on maintaining relationships (Camerer, 2003; Halevy, Cohen, Chou, Katz, & Panter, 2014; Kilduff, 2014). If people anticipate that intra-group competition will damage social relationships, they may prefer to compete against peers they do not expect to befriend (e.g., demographically dissimilar others; Byrne, 1997).

Thus, when competing, women and racial minorities may be more willing to join groups in which they will be tokens for three primary reasons: (1) they believe that, by virtue of being a demographic minority, their performance and point of view will stand out relative to majority group members; (2) they believe that organizations have implicit quotas for demographic minorities and hope to benefit from these quotas; and (3) they want to avoid competition against demographically similar others.

Across a series of six experiments, we show that anticipated intra-group competition influences the groups women and racial minorities choose to join, as predicted. Specifically, we find that competition for scarce opportunities weakens women's and racial minorities' desire to join groups that include similar others, and we present evidence that sheds light on the mechanism responsible for this effect. Our key contributions are to highlight a previously unappreciated moderator of the well-studied preference for homophily — intra-group competition — that is also a common feature of organizational life (Scheiber, 2015; Steinhage et al., 2017) and to explain this phenomenon.

The Desire for Similar Others in Groups

Homophily, defined as the tendency to affiliate with others who have similar beliefs, attitudes, and personal traits (McPherson et al., 2001), is a powerful phenomenon that has been documented across a wide range of contexts and types of relationships (see McPherson et al., 2001 for a review; McPherson & Smith-Lovin, 1987). Past research on homophily suggests that, all else being equal, people are more likely to join groups composed of others who are similar to them than groups composed of dissimilar others.

There is particularly ample evidence that people exhibit homophily when deciding which groups to join or which people to affiliate with in professional settings. For example, studying the decisions of undergraduates tasked with choosing a group to work with on a semester-long project, Baugher et al. (2000) found that self-selected groups were much more similar–or less diverse–with regard to race, gender, and cultural background than would be expected by chance. Similarly, Hinds, Carley, Krackhardt, and Wholey (2000) found that work groups chosen for a four-month software engineering project were also more similar demographically than would be expected by chance. These patterns have also been identified in non-work decisions: McPherson and Smith-Lovin (1987) found people are driven toward homophily in their choice of social organizations and in their choice of friends.

One force behind homophily is the tendency to like people who resemble us (McPherson et al., 2001). Similarity-attraction theory posits that people prefer to affiliate with those who share their attitudes and beliefs (Byrne, 1969; Byrne, London, & Reeves, 1968) or demographic traits (Byrne, 1997; Montoya, Horton, & Kirchner, 2008; Turban, Dougherty, & Lee, 2002). Not only do we have positive affective responses to those who are similar to us, but we expect increased comfort and trust when interacting with them (Baskett, 1974; Byrne, 1969, 1997). People's attitudes toward their work groups are also often consistent with the predictions of similarity-attraction theory. In a survey of employees in a large company, Riordan and Shore (1997) found

employees had more positive attitudes toward their work groups when other members of those groups were more demographically similar to them. Both homophily and similarity-attraction theory suggest that, if given the choice, people will be more likely to join groups that include demographically similar others than groups that do not.

While the aforementioned findings and theorizing apply to all people, racial minorities and women have particular reasons to exhibit homophily. For members of these groups, homophily may also be propelled by an aversion to being in the numeric minority. For example, there is evidence that members of historically underrepresented populations feel isolated, hyper-visible, and pressured to conform to stereotypical roles or behaviors when they are in the minority in groups (Chatman, Boisnier, Spataro, Anderson, & Berdahl, 2008; Yoder, 1991). Furthermore, being severely underrepresented in a work group can harm an individual's performance (Thompson & Sekaquaptewa, 2002) and reduce their job satisfaction (Niemann & Dovidio, 1998). Together, these findings suggest that the experience of being a token in a group can be particularly unpleasant and taxing for historically underrepresented minorities.

The Effects of Competition on Group Preferences

Competition has been linked to increased motivation and a focus on winning in past research (Berger & Pope, 2011; Kilduff, 2014; Plass et al., 2013). For example, Berger and Pope (2011) found in laboratory studies that participants who were told they were competing against others persisted longer on tedious tasks. Further, past research has shown that when people in organizations face competition for scarce resources, they are more likely to engage in strategic thinking (Camerer, 2003; Halevy et al., 2014; Ray, King-Casas, Montague, & Dayan, 2009) and to make comparative social judgments in order to evaluate their position and status (Ashmore, Jussim, & Wilder, 2001). Thus, the prospect of intra-group competition (i.e., competing against

fellow group members) is likely to encourage people to think strategically and engage in social comparison processes as they consider the best ways to achieve success.

One promising strategy for people to deploy in the face of competition for scarce opportunities may be to attempt to stand out from their peers. Differentiating oneself from others prompts attention and increases perceptions of status, both of which can be beneficial in competitions (Maslach, Stapp, & Santee, 1985; Snyder & Lopez, 2001). For example, when competing for rewards, people generally engage in more self-differentiating behaviors (Maslach, 1974). In addition, job candidates often attempt to set themselves apart from others by giving unique answers to traditional interview questions, a strategy that leads to more positive outcomes (Roulin, Bangerter, & Yerly, 2011).

We propose that to stand out from peers, people may elect to join groups where their beliefs, attitudes, and personal traits make them distinct. When competing, people are more likely to compare themselves to those who resemble them because they perceive similar others to be more appropriate targets for comparison than dissimilar others (Brewer & Gardner, 1996; Duffy, Scott, Shaw, Tepper, & Aquino, 2012; Hoffman, Festinger, & Lawrence, 1954). Shared attributes are even more likely to be a basis for social comparison when these attributes are relatively rare (Kilduff, Elfenbein, & Staw, 2010; Mehra et al., 1998). If people facing competitive pressure believe that evaluators are likely to make comparisons within social categories, they may prefer to surround themselves with dissimilar others to stand out. This may be a wise strategy for members of certain groups: past research has found that women and racial minorities tend to stand out in groups, especially when they are numerically underrepresented (Dovidio, Gaertner, & Saguy, 2008).

We propose that being demographically rare in a group can provide those in the numeric minority with three primary benefits. First, people who are tokens may expect their work and behavior to be more visible to colleagues and evaluators (Kanter, 1977; Watkins, Simmons, & Umphress, 2019), and this increased attention to their work could be seen as beneficial in a competitive context. In an experimental study where women were randomly assigned to task-oriented groups such that they would either be the only female in the group (a "solo") or not, female solos were significantly more likely than female non-solos to expect to stand out in their group (Cohen & Swim, 1995). Furthermore, people expect their perspectives, background, and ideas to be more similar to those who resemble them demographically than those who do not (Dipboye & Colella, 2013; Tajfel & Turner, 1979), so they may expect their performance to be more distinctive and salient to evaluators in work groups in which their social identity is also distinctive and salient. Indeed, in a study of women in state legislatures, token women were found to produce work that was more distinct from that of their coworkers than were non-token women (Bratton, 2005). Thus, women and racial minorities may expect their performance and perspective to be more likely to be noticed when they are tokens in a group.

Second, being a token can be beneficial if managers' decision-making is affected by implicit quotas. Prior research suggests that some organizations have implicit quotas that affect their demographic composition (Chang et al., 2019; Dezső et al., 2016). This means that standing out as one of the only underrepresented minorities in a group could actually improve an individual's access to opportunities, particularly when advancement is competitive. Consider, for example, a woman in a male-dominated, competitive, up-or-out organization who is faced with a choice between joining a work group of all men or a gender-diverse group. If she believes that her organization has an implicit quota for the number of women who will be promoted from each group, she may anticipate that her superiors will be reluctant to promote only men. Thus, it would

be strategically beneficial to join an all-male work group, where her token female status increases her chances of earning a promotion. If people believe that managers may be guided by implicit (or explicit) quotas when deciding whom to support or promote, then standing out as one of a few minorities in the running for limited opportunities could be strategically beneficial.

Finally, being a token in a group also means avoiding direct competition with similar peers. Past research has shown that the relationally damaging effects of competition and rivalry tend to be strongest when competing against similar others, and this is especially true for women (Kilduff, 2014; Lee, Kesebir, & Pillutla, 2016). If women and racial minorities expect to get along better with similar others in their organizations, as similarity-attraction theory would predict, they may want to preserve potential relationships with other women or racial minorities, respectively, by avoiding the damaging effects of competition (Lee et al., 2016; Singleton & Vacca, 2007). Instead, they may prefer to compete against people who differ from them demographically (e.g., men, White people), whom they may be more comfortable beating in a competition for a job or promotion. Further, because similar others are more frequent targets for social comparisons (Hoffman et al., 1954) and resources for members of underrepresented populations may feel more limited (Ely, 1994), women and racial minorities may expect demographically similar others to be bigger competitive threats. In fact, such threat responses to potential competition with similar peers have been shown to lead female solos to reject female applicants to preserve their token status and avoid competition with fellow women (Duguid, 2011). They have also been shown to lead women in male-dominated workplaces to avoid relationships with other women to avoid competitive comparisons (Ely, 1994). Thus, women and racial minorities may prefer to compete against men and White people, respectively, because they find competition against similar others more relationally and strategically aversive.

We expect that the effects of intra-group competition on willingness to be in the minority would not extend to men and White people. Due to their frequent majority status in the workplace, dominant group members are less likely to categorize themselves based on their dominant group membership or to consider their dominant demographic characteristic to define their primary identity (McGuire, McGuire, Child, & Fujioka, 1978; McGuire & Padawer-Singer, 1976; Nelson & Miller, 1995). Thus, they may be less likely to consider the demographic identity that has traditionally put them in the majority as a source of distinctiveness that they could leverage in a competitive environment. Furthermore, even when they are in the minority in a group, they may not expect implicit quotas to favor them in a competition given their frequent majority status. Indeed, when dominant group members are in the minority, they tend to be treated differently than non-dominant group members who are in the numeric minority due to their social status and relevant identity-based stereotypes (Crocker & McGraw, 1984; Floge, College, & Merrill, 1986). This suggests that they may not expect managers to evaluate their performance based on implicit quotas, and they are likely to find token status more appealing than minority group members even in non-competitive contexts. Finally, because they are often in the majority in workplace environments, dominant group members are more likely to be comfortable competing against one another and to expect fewer relational costs from competition against demographically similar others (Lee, Kesebir, & Pillutla, 2016).

In sum, we theorize that women and racial minorities will anticipate benefits from being tokens in a group and that they will find it more attractive to distance themselves from others who share salient identity characteristics when competing for scarce opportunities. We propose that this stems from a belief that being in the minority on a salient identity dimension could help them attain scarce opportunities. This belief — whether due to a perception that it will be easier to differentiate themselves, a sense that they could benefit from implicit quotas, or a belief that

competition against same-identity peers will be more relationally damaging — should increase the attractiveness of choosing to be a token or numeric minority in a competitive work group. Taken together, we hypothesize that competition will decrease the tendency for members of historically underrepresented populations to join groups composed of people who share their demographic traits. Further, we predict that this effect will be mediated by (1) a belief that being distinct will allow one's work or performance to stand out from that of competitors; (2) a belief that being one of the only women or racial minorities in a group will allow one to benefit from implicit quotas; and (3) a desire to avoid competition against demographically similar peers.

Overview of Studies

We present six experiments that test our hypotheses about the influence of competition for scarce resources on group preferences. In all of our experiments, we randomly assigned participants to anticipate either competing against other group members for scarce resources (e.g., promotions, bonuses) or not. Then, we let participants choose between joining one of two work groups: a group where they would be underrepresented or a group where they would be surrounded by similar others. In Study 1, we found that female (Study 1A) and Black participants (Study 1B) were more likely to join an all-male group or all-White group, respectively, when competing for scarce resources than in the absence of competition. In Study 2, we disentangled the effects of competition and scarcity to demonstrate that competition drives the preference shift we document. In Study 3, we investigated the mechanisms underlying this phenomenon. We found that a belief that your contributions would stand out more if you were demographically underrepresented mediated this shift in preferences. In Studies 4 and 5, we extended our findings from scenario studies to incentive-compatible studies in which participants made choices between real groups. Notably, across all of our studies, we found evidence that women and minorities preferred working with similar others regardless of their experimental condition. However, we

documented a significant and reliable shift in preferences, such that women and racial minorities facing intra-group competition were more willing to be tokens than those who were not facing competition.

Study 1

Study 1A. In Study 1A, we tested our hypothesis that women would be more willing to join an all-male group when facing the prospect of intra-group competition. Women were asked to choose between joining one of two groups for a summer internship, and the groups differed only in their proportion of female members. Competition was experimentally manipulated by altering the percentage of the interns in each group who could expect to receive a full-time job offer at the end of the summer.

Methods

Participants. 900 U.S. participants were recruited through Amazon Mechanical Turk to participate in a 5-6-minute research study for $0.60. Per our preregistration, we excluded participants who indicated in our survey that they were not women, leaving us with a final sample size of 491 women.

Procedure. This experiment was a two-condition (*competitive* vs. *control*) scenario study preregistered on AsPredicted.org (http://aspredicted.org/blind.php?x=rt44qm).

Participants in our experiment were told to imagine they had been offered a summer internship at an organization and they had to choose which of two different departments to join. They were told that their roles and access to senior colleagues would be the same across departments, so the only difference between the two departments would be their fellow interns.

To confirm that all participants were women, participants were then asked to report their gender identity ("Woman", "Man", or "Another identity not listed").

All participants were randomly assigned to one of two experimental conditions: a *competitive* condition or a *control* condition. In the *competitive* condition, participants were told that only 25% of interns would be offered full-time jobs at the end of the summer, so they would be competing intensely against the other interns in the department they chose for a full-time job offer. In the *control* condition, participants were told that almost all interns would be offered full-time jobs at the end of the summer, so they would not be competing against the other interns in the department they selected for a full-time job offer. Participants were then asked to choose between the two departments.

Dependent variable. The dependent variable of interest was the proportion of women in each condition who chose to join the all-male group. The information displayed about each department included the photos, names, and college majors of the other summer interns who would be working in the department (see Figure 4 for an example of our stimuli). One department was composed of seven men. The other department was composed of four men and three women; thus, the composition of this group would be 50% female if the female participant joined that department. The photos of interns displayed were gathered from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015), and college majors and race were matched across groups, such that the racial composition of the groups was the same and the majors were similar (though not identical, in order to reduce suspicion) in both groups. We stimulus-sampled both the photographs and the college majors associated with each group, creating a total of six stimuli sets. After choosing a group, participants were asked to answer a free-response question explaining why they had chosen their preferred group. All study materials are available in our Online Supplement.

Manipulation check. As a manipulation check, at the end of our study, participants indicated to what extent they anticipated competing against the other interns in their department for a full-time job on a scale from 1 (Not competing at all) to 5 (Competing very intensely).

Results

Our manipulation appeared to work as intended: on a scale from 1 (Not competing at all) to 5 (Competing very intensely), participants expected to compete against the other interns significantly more in the *competitive* condition ($M_{competitive} = 4.67$, $SD_{competitive} = 0.61$) than in the *control* condition ($M_{control} = 1.57$, $SD_{control} = 0.97$; $t(489) = 42.23$, $p < .001$).

As predicted, women in the *competitive* condition were significantly more likely to choose to join the all-male group (46.1%) than were women in the *control* condition (17.5%), $z = 6.72$, $p < .001$. These results suggest that women's willingness to join all-male groups increased significantly when they expected to face intra-group competition.[2]

Study 1B. In Study 1B, we extended the results of Study 1A by examining whether they replicated with Black participants instead of women. Specifically, we examined whether Black participants were more willing to join a group whose members were all White when they anticipated competing against other group members for scarce opportunities.

Methods

Participants. To recruit enough Black participants in this experiment to reach our preregistered sample size target, we recruited participants on both Prolific and Amazon Mechanical Turk. In total, 278 Black participants were recruited via these sites to participate in a

[2] Women in Study 1A chose to join the all-male group significantly less than chance in the *control* condition (17.5%), $z = 7.53$, $p < .001$. The rate at which women in Study 1A chose to join the all-male group in the *competitive* condition did not differ significantly from chance (46.1%), $z = 0.77$, $p = .44$.

5-6 minute study. Prolific participants (N=104) were paid $0.70, while Mechanical Turk participants (N=174) were paid $0.60 due to the different pricing thresholds on the two services.

Procedure. This study was a two condition (*competitive* vs. *control*) scenario study preregistered on AsPredicted.org (http://aspredicted.org/blind.php?x=g3cs9e).

The study design was nearly identical to the design of Study 1A. Participants again were randomly assigned to either a *competitive* or *control* condition and invited to choose which department they would prefer to join at a company where they had been offered a summer internship. However, in this experiment, the racial (rather than gender) composition of the other interns was the primary difference between the two departments. To confirm that all participants were Black, participants were asked about their racial identity (i.e., "White," "Black," "Asian," etc.) instead of their gender identity. As in Study 1A, participants in the *competitive* condition learned that only 25% of interns would be offered full-time jobs at the end of the summer, while those in the *control* condition were told that almost all interns would be offered full-time jobs.

Dependent variable. The dependent variable of interest was the proportion of participants choosing to join the all-White group. When choosing which group to join, participants again were shown the photos, names, and college majors of the other summer interns in each group. Both intern groups included four men and three women. In one group, all interns were White; in the other group, three were Black and four were White, such that the more diverse group would be 50% Black if a participant chose to join it. All study materials are available in our Online Supplement.

Manipulation check. At the end of the study, as a manipulation check, participants indicated to what extent they anticipated competing against the other interns in their department for a full-time job on a scale from 1 (Not competing at all) to 5 (Competing very intensely).

Results

A manipulation check confirmed that our manipulation of intra-group competition was successful: on a scale from 1 (Not competing at all) to 5 (Competing very intensely), participants expected to compete against their fellow interns for jobs significantly more in the *competitive* condition ($M_{competitive} = 4.49$, $SD_{competitive} = 0.82$) than in the *control* condition ($M_{control} = 1.50$, $SD_{control} = 0.87$; $t(276) = 29.34$, $p < .001$).

Lending additional support to our primary hypothesis, a significantly higher proportion of Black participants chose to join the all-White group in the *competitive* condition (36.6%) than in the *control* condition (19.9%), $z = 2.97$, $p = .003$.[3]

Discussion. In Study 1, we found that female (Study 1A) and Black participants (Study 1B) were more likely to choose to join a group in which they would be the only person of their gender or race when they expected to compete against other group members for scarce resources than when they did not expect to compete. Of note, neither experiment documented a reversal in preferences: across all conditions in all experiments, we found that participants preferred to join work groups that included similar others. However, we identified a reliable and statistically significant shift in preferences such that when intra-group competition was introduced, people found it more attractive to join groups where they would be in the numeric minority.

Study 1A and Study 1B demonstrate that the effects of competition on group choice generalize to those with different historically underrepresented demographic identities. Importantly, Black Americans and females have different levels of representation in the US workforce and the population at large. In particular, although women are roughly 50% of the US

[3] Black participants in Study 1B chose to join the all-White group significantly less than chance in both the *control* condition (19.9%), $z = 5.09$, $p < .001$ and the *competitive* condition (36.6%), $z = 2.16$, $p = .03$.

population, Black Americans make up only a little more than 13% of the population. Thus, in our studies, seeing a group with near gender parity might have produced a different reaction than seeing a group with near racial parity. Study 1 demonstrates that in spite of this, our findings generalize. They apply not only across distinct identity groups, but also across identity groups with very different levels of representation in the US population and workforce.

While participants in both Studies 1A and 1B decided whether to be a lone representative of their identity group, in a conceptual replication of Study 1A, we found the same pattern of results when the group with zero women was replaced with a group including one woman (see Study S1 in the Online Supplement). This suggests our phenomenon extends beyond situations in which women and racial minorities expect to be a lone representative of their identity group to situations in which they merely expect to be underrepresented.

One potential concern about Studies 1A and 1B is that they conflated competition with scarcity. That is, the *competitive* condition differed from the *control* condition in two ways: (1) participants were told that their group would be competitive, and (2) they were told that only 25% of their group members (rather than almost all group members) would receive a reward or job. In Study 2, we sought to disentangle the effects of competition for scarce resources from the effects of scarcity alone.

Study 2

In Study 2, we sought to isolate the effects of reward scarcity from the effects of competition to determine whether our effect is driven by intra-group competition, as we hypothesize, or mere scarcity of rewards.

Methods

Participants. Five hundred and ninety-two women were recruited for this experiment via Amazon Mechanical Turk.

Procedure. This experiment was a three condition (*competitive* vs. *lottery* vs. *control*) scenario study and was preregistered on AsPredicted.org (http://aspredicted.org/blind.php?x=a647tk).

Participants were asked to imagine that they were working at an organization poised to launch two new products, and special teams had been created to supervise each of the two product launches. They were then asked to make a hypothetical choice between joining one of the two product launch teams at the company. The teams were essentially indistinguishable, except that one was all-male and the other was mixed-gender. All participants were told that regardless of how their team performed as a whole, the organization would conduct an individual performance evaluation at the end of the project.

To confirm that all participants were women, participants were asked to report their gender identity ("Woman", "Man", or "Another identity not listed"). Participants were then randomly assigned to one of three experimental conditions. Participants randomly assigned to the *competitive* condition were told that only 25% of the employees from each team would be chosen based on performance to earn a cash bonus and company recognition, so they would be competing against their teammates for a reward. In the *lottery* condition, participants were told that only 25% of the employees from each team would be chosen based on pure luck of the draw to earn a cash bonus and company recognition, and they would not be competing against their teammates. Thus, the scarcity of rewards was held constant between the *competitive* and *lottery* conditions – 25% of employees from each team would earn a bonus – but the presence of competition was varied. Finally, in the *control* condition, which mirrored the control conditions in

prior studies, we eliminated both competition and scarcity by telling participants that after the performance evaluation, almost all employees from each team would earn a cash bonus and company recognition, so they would not be competing against their teammates nor would rewards be scarce.

Dependent variable. As in our past studies, our dependent variable of interest was the proportion of female participants in each condition choosing to join the all-male team. Participants were asked to choose between the two product launch teams. The information about each team included a set of professional headshots that were matched on apparent age as well as the names and job positions of the employees on each team. We stimulus sampled by creating three distinct sets of all-male teams and three distinct sets of gender-mixed teams. All study materials are available in our Online Supplement.

Manipulation check. At the end of the study, as a manipulation check, participants indicated to what extent they anticipated competing against the other employees on their team for a bonus on a scale from 1 (Not competing at all) to 5 (Competing very intensely).

Results and Discussion

First, we confirmed that our manipulation was successful: on a scale from 1 (Not competing at all) to 5 (Competing very intensely), participants reported that they expected to compete against their fellow interns for a full-time offer significantly more in the *competitive* condition ($M_{competitive} = 4.52$, $SD_{competitive} = 0.74$) than in the *control* condition ($M_{control} = 1.53$, $SD_{control} = 0.88$; $t(393) = 36.45$, $p < .001$) or the *lottery* condition ($M_{lottery} = 1.54$, $SD_{lottery} = .98$; $t(392) = 33.97$, $p < .001$), while expectations of competition in the *control* and *lottery* conditions did not differ ($t(393) = 0.06$, $p = .95$).

As in our prior studies, participants in the *competitive* condition chose to join the all-male group significantly more (22.8%) than participants in the *control* condition (9.1%), $z = 3.59$, $p < .001$. Furthermore, participants in the *competitive* condition were also more willing to join the all-male group than were participants in the *lottery* condition (12.7%), $z = 2.50$, $p = .012$. Finally, the rate of choosing the all-male team did not differ significantly between the *lottery* and *control* conditions, $z = 0.99$, $p = .32$.[4]

These findings suggest that scarcity alone is not enough to produce our effect. Rather, intra-group competition is necessary to increase women's desire to join an all-male team. However, Study 2 does not help us understand why intra-group competition leads women to be more willing to be tokens. In Study 3, we sought to identify the mechanism responsible for the effect of intra-group competition on the preferences of women and racial minorities.

Study 3

In Study 3 we extended our past studies by delving into the mechanisms responsible for women's and racial minorities' increased willingness to be tokens in competitive contexts. Specifically, we explored the extent to which this effect was driven by (1) a belief that being a token would make an individual's work more unique and more likely to be noticed, (2) a belief that being a token would allow them to benefit from implicit quotas, and (3) a desire to avoid competing against similar others due to the relationally damaging effects of competition.

Methods

Participants. Three hundred and ninety-six women were recruited for this study via Amazon Mechanical Turk.

[4] Women in Study 2 chose to join the all-male group significantly less than chance in the *competitive* (22.8%), $z = 5.50$, $p < .001$; *lottery* (12.7%), $z = 7.87$, $p < .001$; and *control* (9.1%), $z = 8.81$, $p < .001$ conditions.

Procedure. This study had two experimental conditions (*competitive* vs. *control*) and was preregistered on AsPredicted.org (http://aspredicted.org/blind.php?x=qc52u8).

Study 3 relied on the same paradigm as Study 1A, and again, participants were randomly assigned to either a *competitive* or a *control* condition. Again, they were told that they had to choose between two different departments within the same organization for a summer internship and that the only difference between the two departments would be their fellow interns. To confirm that all participants were women, they were then asked to report their gender identity ("Woman", "Man", or "Another identity not listed").

As in Study 1A, participants in the *competitive* condition were told that only 25% of interns would be offered a full-time job at the end of their summer internship, whereas participants in the *control* condition were told that almost all interns would be offered a full-time job. However, unlike Study 1A, after women selected which internship group they would prefer to join (an all-male group or a mixed gender group), we presented them with six questions designed to measure our three hypothesized mediators (with two questions for each mediator). Participants were asked to indicate their agreement with each of the six statements, presented in randomized order, on a scale from 1 (Strongly disagree) to 7 (Strongly agree). For each set of items, we report the Spearman-Brown coefficient.

Mediators. To measure whether participants thought being a different gender from other group members would make their performance stand out, we asked participants to rate their agreement with the statements, "I think my work or performance will be distinct from that of other interns in my department" and "I think I bring a unique perspective to my department" (Spearman-Brown coefficient = 0.57, $p < .001$). As per our preregistration, we averaged these two items to create a measure of participants' performance differentiation considerations.

To measure whether participants thought they might benefit from implicit gender quotas, we asked them to rate their agreement with the statements, "I think managers will want to ensure that at least one woman receives a full-time job from each department" and "I think managers will be reluctant to give a full-time job only to men in each department" (Spearman-Brown coefficient = 0.42, $p < .001$). As per our preregistration, we averaged these two items to create a single measure of participants' implicit quota motives.

Finally, to measure whether participants expected competition against women to be more relationally damaging than competition against men, we asked them to rate their agreement with the statements, "I feel tense competing against women" and "I don't feel as comfortable competing against women as I do competing against men" (Spearman-Brown coefficient = 0.65, $p < .001$). As per our preregistration, we averaged these two items to create a single measure of participants' aversion to competition against similar others.

Dependent variable. As in our past studies, our dependent variable of interest was the proportion of female participants in each condition choosing to join the all-male group. Participants were asked to choose between the two groups. The information about each group included photos, names, and college majors of the interns in each group. We stimulus-sampled both the photographs and the college majors of the group members, creating a total of six stimuli sets. All study materials are available in our Online Supplement.

Manipulation check. At the end of our study, as a manipulation check, participants indicated to what extent they anticipated competing against the other interns in their department for a full-time job on a scale from 1 (Not competing at all) to 5 (Competing very intensely).

Results

As in previous studies, our manipulation was successful: on a scale from 1 (Not competing at all) to 5 (Competing very intensely), participants expected to compete against their fellow interns for a full-time job offer significantly more in the *competitive* condition ($M_{competitive}$ = 4.62, $SD_{competitive}$ = 0.70) than in the *control* condition ($M_{control}$ = 1.54, $SD_{control}$ = 0.93; $t(394) = 37.30$, $p <$ .001). In addition, we replicated our findings from Study 1A: women were more willing to join the all-male work group in the *competitive* condition (37.4%) than in the *control* condition (19.2%), $z = 3.91$, $p < .001$.[5] There was also a significant, positive effect of assignment to the *competitive* condition on participants' belief that their performance or perspective would be distinct from that of fellow group members ($p < .001$). However, assignment to the *competitive* condition had no effect on implicit quota motives or aversion to competing against similar others ($p = .486$ and $p = .133$, respectively).

As per our preregistration, we first tested whether each proposed mechanism independently mediated the relationship between intra-group competition and choice of the all-male group. We found that only participants' belief that their performance or perspective would be distinct from that of fellow group members mediated the effect of intra-group competition on willingness to choose the all-male group. First, we documented a significant main effect of assignment to the *competitive* condition on performance differentiation considerations ($b = 0.361$, $SE = .107$, $p < .001$). Furthermore, the relationship between performance differentiation considerations and the choice of the all-male group was also significant ($b = 0.088$, $SE = .019$, $p <$ .001). Consistent with our mediation hypothesis, the effect of assignment to the *competitive* condition on study participants' choice to join the all-male group ($b = 0.180$, $SE = .045$, $p < .001$) was reduced when controlling for participants' expectation that they would bring a distinct perspective to their chosen group ($b = 0.148$, $SE = .044$, $p < .001$). A Sobel test confirmed that

[5] Women in Study 3 chose to join the all-male group significantly less than chance in both the *control* condition (19.2%), $z = 6.34$, $p < .001$ and the *competitive* condition (37.4%), $z = 2.43$, $p = .015$.

this reduction in effect size was significant ($b = 0.032$, $SE = .012$, $p = .008$), and a 5,000-sample bootstrap analysis (MacKinnon, Fairchild, & Fritz, 2007; Shrout & Bolger, 2002) also produced a 95% bias-corrected confidence interval for the size of the indirect effect that excluded zero (95% CI: [0.013, 0.058]). Neither implicit quota motives (indirect effect $b = -0.009$, $SE = 0.010$, $p = .408$) nor an aversion to competition against similar others (indirect effect $b = 0.020$, $SE = 0.015$, $p = .176$) significantly mediated the effect of intra-group competition on group choice.

Again following our pre-registration, we then tested all three mechanisms simultaneously as mediators of our effect with a 1,000 bootstrap sample multiple mediator model (Preacher & Hayes, 2008). When we include all three potential mediators in the bootstrapped mediation model, the results confirm that performance differentiation considerations significantly mediated the effect of intra-group competition ($b = 0.031$, $SE = 0.012$, $p = .007$; 95% CI: [0.008, 0.054]). And again, neither implicit quota motives ($b = -0.003$, $SE = 0.005$, $p = .508$; 95% CI: [-0.012, 0.006]) nor an aversion to competing against similar others ($b = 0.020$, $SE = 0.014$, $p = .143$; 95% CI: [-0.007, 0.047]) significantly mediated the relationship between assignment to the *competitive condition* and choosing to be a token woman (the results of this mediation model are depicted in Figure 5, and a full correlation table between variables is shown in Table 2).

Discussion

Study 3 provides evidence that one reason why women and underrepresented minorities may be more willing to join groups in which they will be tokens when facing competition is that they believe doing so will increase the odds that their work is differentiable from the work of others. Specifically, they believe that being demographically distinct from other group members will allow them to bring a unique perspective to their work, helping them stand out.

Study 3 also shows that implicit quota considerations and the desire to avoid competition against similar others do not mediate women's choice to join groups devoid of other women at a higher rate when they expect to compete with fellow group members. Empirically, this may be because there were no significant differences across conditions in the degree to which women expected managerial decisions to be affected by implicit quotas, and there were no significant differences across conditions in the degree to which women expected competition against fellow women to be more aversive (see Figure 5).

While Studies 1–3 established the robustness of our findings and delved into the mechanism responsible for them, they all involved hypothetical scenarios. In our remaining studies, we asked participants to make real, incentive-compatible decisions to replicate our effects and show their generalizability to other settings.

Study 4

In Study 4, we extended our findings to participants in an incentive-compatible experiment who expected to interact with their chosen group members on an in-person task. Participants in a laboratory experiment chose which of two groups to join for an in-person brainstorming session, and we randomly assigned them to either anticipate competing with others in their group of choice for public recognition and a cash bonus, or not.

Methods

Participants. Participants (145 women and 57 men) were recruited at a U.S. university to participate in a one-hour research session that included our experiment. Participants were paid $10 to participate in the session and were told that they could earn a bonus of up to $10 by participating in a follow-up brainstorming session. Unlike past studies, we included both male

and female participants in this experiment because both were present in the lab session. However, as in prior studies, our analyses focused on the behavior of female participants.

Procedure. This experiment had two conditions (*competitive* vs. *control*). Prior to the research session, participants were asked to fill out a pre-survey that asked for their name, year in college, a hobby, and a photograph of themselves. They were told that these photos would be used during our laboratory session. Before entering the lab, any participants who had not completed the pre-survey were pulled aside and asked for their name, year in college, and a hobby. If consent was granted, their photo was also taken for use in the laboratory session.

During the experiment, participants were provided with a brief overview of the body positivity movement, a social movement rooted in the belief that all bodies are good bodies and that everyone should be able to achieve a positive body image. They were truthfully told that we were seeking ideas to use in a body positivity campaign at their university and that we would be hosting an in-person brainstorming session at a separate time and place to generate these ideas. Participants were informed that they would work in a group with fellow lab participants at the brainstorming session to develop ideas for a body positivity campaign, but that all group members would submit an independent write-up of their favorite idea. We told participants that the brainstorming session would occur after the lab experiment was over, and they would earn $5 for showing up plus a potential bonus depending on the quality of ideas they submitted individually at the end of the brainstorming session. In other words, while the group choice happened on a computer in the lab, the group task was in-person and outside of the lab. All participants, regardless of condition, learned that a real panel of judges would evaluate the ideas each individual submitted during the brainstorming session to choose several that would be posted on a real university website, earning the authors of the selected ideas public recognition and a $5 bonus on top of their show-up fee.

After answering several questions about their demographics, participants were assigned to either a *competitive* or *control* condition in this experiment. Participants in the *competitive* condition learned that only 25% of the ideas from each brainstorming group would be selected, so they would be competing against fellow group members for rewards and recognition. Participants in the *control* condition learned that nearly all of the ideas from each brainstorming group would be selected, so they would not be competing against fellow group members. We held these brainstorming sessions as promised and assigned bonuses as described.

Dependent variable. The primary dependent variable of interest was the proportion of women choosing to join the all-male group across conditions. After reading the instructions, participants were asked to choose between two seven-person groups to join for the brainstorming session and were shown photographs and background information (name, year in college, and a hobby) about the other seven people in the two available groups.[6] Participants who indicated that they were women were presented with a choice between a group of only men and another equal-sized group of three women and four men. Participants who indicated that they were men chose between one group of only women and another equal-sized group of three men and four women. As in our prior experiments, we stimulus-sampled the photographs, names, class years, and hobby in each group, creating a total of six stimuli sets (three for men and three for women). Complete study stimuli are available in our Online Supplement.

Manipulation check. At the end of the study, after selecting their group for the brainstorming session, participants completed a manipulation check in which they were asked to answer the following question: "To what extent do you feel like you'll be competing against the

[6] In order to ensure that participant behavior would not be affected by seeing photos of their friends or acquaintances, the stimuli included the names, years in college, hobbies, and photos of college students or recent graduates from other institutions rather than other members of their study session. In other words, this study involved deception, which was approved by our IRB.

other participants in your group for a bonus and recognition?" They were asked to answer this question on a scale from 1 (Not competing at all) to 5 (Competing very intensely).

Results and Discussion

As in previous studies, our manipulation was successful: on a scale from 1 (Not competing at all) to 5 (Competing very intensely), participants expected to engage in significantly more intra-group competition in the *competitive* condition ($M_{competitive} = 2.82, SD_{competitive} = 1.08$) than in the *control* condition ($M_{control} = 1.26, SD_{control} = 0.58$; $t(143) = 10.80$, $p < .001$).

We were primarily interested in whether women in the *competitive* condition would be more likely to choose to join the all-male group for the brainstorming session than women in the *control* condition. Thus, we compared the proportion of women choosing the all-male group of students across conditions. Consistent with the results of our scenario studies, women in the *competitive* condition were significantly more likely to join the all-male brainstorming group (23.3%) than were women in the *control* condition (9.7%); $z = 1.97, p = .048$.[7]

Although we were primarily interested in the behaviors of women and were underpowered to test the parallel effect among men (N=57 men), we also explored the impact of competition on men's choices. As noted in our introduction, our theory predicts that men, being dominant group members, should be less prone to show our effect. Indeed, we found no significant differences in the rate at which men in the *competitive* condition chose to join an all-female group (24.1%) as compared to men in the *control* condition (28.6%); $z = 0.08, p = .94$. These results provide suggestive evidence that men's decisions to be tokens in groups are relatively unaffected by the presence of intra-group competition.

[7] Women in Study 4 chose to join the all-male group significantly less than chance in both the *competitive* (23.3%), $z = 3.18$, $p = .001$ and *control* (9.7%), $z = 5.10$, $p < .001$ conditions.

The results of Study 4 confirm that women in an incentive-compatible context choosing a group for an in-person interaction are still more willing to choose all-male groups when they expect to compete against their fellow group members than when they do not expect to compete. To ensure that these results were not due to the context and population being studied or our use of deception, we next ran an incentive-compatible, non-deceptive study in a different context.

Study 5

In Study 5, we sought to replicate the results of Study 4 in a preregistered, non-deceptive experiment in another setting involving real decisions. Workers on Amazon Mechanical Turk were invited to choose one of two real, digital work groups to join, knowing that they either would or would not compete against their fellow group members for a bonus.

Methods

Participants. Five hundred and eighty-three women were recruited through Amazon Mechanical Turk to participate in an eight-minute research study in exchange for $0.90 and a potential $0.50 bonus.[8]

Procedure. This was a two condition (*competitive* vs. *control)* experiment preregistered on AsPredicted.org (http://aspredicted.org/blind.php?x=j8vm2h).

Participants in our experiment began by indicating their gender and telling us their preferred nickname and hometown. Participants were then told they would be writing a review for a website along with a group of other MTurk workers and that they would be choosing which of two groups of reviewers to join. The two groups would review different (but very similar) websites and were also composed of different people. Participants were informed that after

[8] We collected 630 female participants on MTurk, aiming for 600 participants after exclusions. Ultimately, we ended up with 583 participants after our preregistered exclusions.

writing their website review, they would interact with other members of their group. Finally, participants were truthfully told that their review would actually be used to describe the website to a diverse group of consumers and that their reviews would be published along with those of other MTurkers in the group.[9]

Participants were randomly assigned to one of two experimental conditions: a *competitive* condition or a *control* condition. In the *competitive* condition, participants were told that we would select the three best reviews from each reviewer group and that only the participants who wrote those reviews would earn a $0.50 bonus. Thus, they would be competing against the other MTurkers in their group. In the *control* condition, participants were told that we would use all the reviews from each group and that everyone would earn a $0.50 bonus. Therefore, they would not be competing against their fellow group members for a bonus.

Dependent variable. The dependent variable of interest was the proportion of participants who chose to join the all-male group. After reading the task description, participants were asked to choose which of two website-evaluation groups to join. As mentioned previously, the groups would evaluate different (but similar) websites (either Buzzfeed.com, HuffingtonPost.com, Vice.com or Vox.com), and membership in the two groups would not overlap.[10] To facilitate their group selection, participants were shown avatars of other group members (revealing their genders) as well as the nicknames and hometowns of each group member (see Figure 6 for an example). Both groups included nine people, and each participant chose between a group

[9] This study did not involve deception; we followed through on all promises made to MTurk workers and they were paired with the group of their choice.

[10] We stimulus-sampled in this study, and the two websites up for review were randomly selected from a set of four sites: Buzzfeed.com, HuffingtonPost.com, Vice.com, and Vox.com. All groups displayed were composed entirely of prior participants who had reviewed each of the four websites and provided us with their gender, a nickname, and their hometown. In total, there were three different pairs of group stimuli sampled in this study.

composed exclusively of men and a group composed of five men and four women. Complete study stimuli are available in our Online Supplement.

After selecting their group, participants were asked to write a short review of the website associated with their group of choice. They then read a website review written by a fellow group member and provided feedback.

Manipulation check. Finally, at the end of the study, as a manipulation check, participants indicated on a scale from 1 (Not at all) to 5 (Very much) to what extent they felt they would be competing against their fellow group members for a bonus.

Results and Discussion

Our manipulation was again successful: on a scale from 1 (Not at all) to 5 (Very much), participants expected to engage in significantly more intra-group competition in the *competitive* condition ($M_{competitive} = 3.61$, $SD_{competitive} = 1.27$) than in the *control* condition ($M_{control} = 2.16$, $SD_{control} = 1.37$; $t(581) = 13.22$, $p < .001$).

To test our primary hypothesis, we compared the proportion of women in each condition who chose to join the all-male review group. Consistent with our other studies, we found that significantly more women in the *competitive* condition chose to join the all-male review group (41.6%) than in the *control* condition (32.1%); $z = 2.27$, $p = .023$.[11] In other words, when women expected to compete against fellow group members for a monetary bonus, they were more likely to join an all-male group (in which they would be the sole female) than in the absence of competition.

[11] Women in Study 5 chose to join the all-male group significantly less than chance in both the *competitive* (41.6%), $z = 2.00$, $p = .046$ and *control* (32.1%), $z = 4.21$, $p < .001$ conditions. In other words, women chose homophily (the diverse group containing similar others) significantly more than chance in all conditions.

General Discussion and Conclusion

Across six experiments, we show that competition for scarce resources increases the rate at which people from historically underrepresented populations choose to join groups in which they will be tokens. In short, competition serves as a partial counterweight to the well-established tendency toward homophily. We find this pattern for female and Black participants, and it arises in both hypothetical scenario studies and studies involving real, incentivized choices[12]. Our findings suggest that intra-group competition leads to a greater desire to join groups where people believe their work output and ideas will be differentiated from those of their peers, and women and racial minorities anticipate that joining a group where they will have token status makes this more likely.

Our findings add to the relatively limited literature examining how women and racial minorities select their teams and groups at work (cf. Avery & McKay, 2006; Duguid, 2011; McKay et al., 2007; Umphress et al., 2007). We find that competition can shape the willingness of women and racial minorities to work with dissimilar others. Of note, across all of our studies, we see that people prefer to join groups in which they will not be tokens: we demonstrate that the preference for homophily is weakened–but not reversed–when people expect to compete against fellow group members for scarce resources.

In our theorizing, we suggested three potential reasons for women's and racial minorities' increased willingness to be tokens when anticipating intra-group competition. Namely, we hypothesized that this effect might be driven by (1) a belief that being a token would make your perspective and work more unique and therefore more likely to get noticed by decision-makers; (2) a belief that being a token would allow you to benefit from implicit quotas; and (3) an

[12] For a full summary of our results across studies, see Table 2.

aversion to competition against similar others because of the relationally damaging effects of competition. In Study 3, we found evidence for only the first of these hypothesized mechanisms.

Our work does not examine whether the effects of intra-group competition can actually enhance demographic diversity in organizations. In homogeneous organizations, the preferences we document may encourage more women and racial minorities to join when intra-group competition is emphasized; however, in organizations that are already diverse, competitive work groups may be unattractive to minorities. It would be valuable for future work to explore this question and determine the effects of emphasizing intra-group competition on a firm's ability to diversify its workforce.

The results of Study 4 also suggest that majority group members do not show our effect: men are just as likely to choose to be solos when they expect to compete against their fellow group members as when they do not. Given that this finding involved one small (N=57) subgroup in one study (which was originally not intended for analysis), it would be valuable for future work to more thoroughly examine the effects of intra-group competition on dominant group members in well-powered studies. Despite being underpowered, however, these results are consistent with our theorizing. Because dominant group members are frequently in the majority, they are less likely to spontaneously categorize themselves based on their dominant identity (McGuire et al., 1978; McGuire & Padawer-Singer, 1976; Nelson & Miller, 1995) and may be less likely to expect to be distinctive or stand out to evaluators due to their identity. Thus, the strategic considerations that Study 3 suggests drive our effect for those used to being underrepresented may not be as salient for those used to being well-represented. These findings add to the literature on the different impacts of competitive environments on majority and minority group members (Flory, Leibbrandt, & List, 2015; Niederle & Vesterlund, 2007).

An important limitation of our studies is that they relied exclusively on data collected in the laboratory and online. As a result, even in our incentive-compatible studies, the groups participants joined only interacted briefly, and the incentives provided were relatively small. Past research suggests that people may behave differently in one-shot and repeated interactions (Bó, 2005; Bornstein, Winter, & Goren, 1996). Thus, future tests of our theories in workplaces or other settings where groups interact repeatedly over extended time intervals and where the incentives available for individual performance are larger would be valuable. Finally, assessing whether the rate at which people opt in to being tokens varies systematically based on their social identity and why would add richness to our understanding of this phenomenon.

An important question raised by this research is whether women and racial minorities are wise to choose to join all-male and all-White groups, respectively, in competitive environments given the potential negative long-term consequences of being a token. Past research has shown that when women and racial minorities are tokens, their performance tends to suffer (Thompson & Sekaquaptewa, 2002), as does their organizational commitment (Niemann & Dovidio, 1998). Furthermore, being a token can harm long-term psychological well-being and feelings of belonging in the workplace (Kanter, 1977; Yoder & Sinnett, 1985). Over time, the perceived strategic value of standing out may be dwarfed by the damaging effects of hyper-visibility and isolation (Cohen & Swim, 1995; Kanter, 1977).

Future studies might test whether demographic minorities anticipate this tension by measuring which groups they believe will lead them to be happiest at work and where they predict having the longest tenure. Employees may strategically choose to join groups in which they will be in the minority when facing the prospect of competition, despite anticipating being happier and remaining longer in groups composed of similar others. Future research could also explore whether an increased desire to be in the numeric minority when competing affects

affiliative or collaborative behavior or social cognition after women and racial minorities choose a team.

Furthermore, much of the past literature on the consequences of being a token focuses on situations in which individuals did not actively *choose* to be tokens. It would be valuable for future work to examine whether women and racial minorities who make the active decision to be tokens experience diminished negative effects on their performance and organizational commitment in the long run.

It is also an open question as to whether choosing to be a token is a wise strategic decision for career advancement. Prior work suggests that tokens feel they have to work harder for promotions and that women who anticipate being tokens perform worse on ability tests than women who anticipate working with other women (Archbold & Schulz, 2008; Keller & Sekaquaptewa, 2008). However, there is some evidence that being one of few underrepresented minorities in a group *does* have the kinds of strategic benefits that participants in our studies appeared to anticipate when they chose which groups to join, particularly in firms that care about diversity. For example, past research has shown that some companies appear to have implicit quotas for the levels of diversity they aim to achieve on top management teams (Chang et al., 2019; Dezső et al., 2016). If there are indeed a fixed number of opportunities for women and racial minorities to advance, then it may in fact be advantageous for them to join groups in which they will have a better chance of "standing out." Furthermore, Leslie, Manchester, and Dahm (2017) have shown that high-potential women receive larger rewards in the workplace than high-potential men precisely because they are in short supply in many firms. Future research that directly explores whether the kinds of decisions made by women and racial minorities in our studies are optimal or sub-optimal would be valuable.

References

Archbold, C. A., & Schulz, D. M. (2008). Making Rank: The Lingering Effects of Tokenism on Female Police Officers' Promotion Aspirations. *Police Quarterly*, *11*(1), 50–73. https://doi.org/10.1177/1098611107309628

Ashmore, R. D., Jussim, L. J., & Wilder, D. (2001). *Social Identity, Intergroup Conflict, and Conflict Reduction*. Oxford University Press.

Avery, D. R., & McKay, P. F. (2006). Target Practice: An Organizational Impression Management Approach to Attracting Minority and Female Job Applicants. *Personnel Psychology*, *59*(1), 157–187. https://doi.org/10.1111/j.1744-6570.2006.00807.x

Baskett, G. D. (1974). Interview decisions as determined by competency and attitude similarity. *Journal of Applied Psychology*, *57*(3), 343. https://doi.org/10.1037/h0034707

Baugher, D., Varanelli, A., & Weisbord, E. (2000). Gender And Culture Diversity Occurring In Self-formed Work Groups. *Journal of Managerial Issues*, *12*(4), 391–407.

Berger, J., & Pope, D. (2011). Can Losing Lead to Winning? *Management Science*, *57*(5), 817–827. https://doi.org/10.1287/mnsc.1110.1328

Bó, P. D. (2005). Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games. *American Economic Review*, *95*(5), 1591–1604. https://doi.org/10.1257/000282805775014434

Bornstein, G., Winter, E., & Goren, H. (1996). Experimental study of repeated team-games.

European Journal of Political Economy, *12*(4), 629–639. https://doi.org/10.1016/S0176-2680(96)00020-1

Bratton, K. A. (2005). Critical Mass Theory Revisited: The Behavior and Success of Token Women in State Legislatures. *Politics & Gender*, *1*(1), 97–125. https://doi.org/10.1017/S1743923X0505004X

Brewer, M. B., & Gardner, W. (1996). *Who Is This "We"? Levels of Collective Identity and Self Representations*. 11.

Byrne, D. (1969). Attitudes and Attraction. In *Advances in Experimental Social Psychology* (Vol. 4, pp. 35–89). https://doi.org/10.1016/S0065-2601(08)60076-3

Byrne, D. (1997). An Overview (and Underview) of Research and Theory within the Attraction Paradigm. *Journal of Social and Personal Relationships*, *14*(3), 417–431. https://doi.org/10.1177/0265407597143008

Byrne, D., London, O., & Reeves, K. (1968). The effects of physical attractiveness, sex, and attitude similarity on interpersonal attraction1. *Journal of Personality*, *36*(2), 259–271. https://doi.org/10.1111/j.1467-6494.1968.tb01473.x

Cable, D. M., & Judge, T. A. (1996). Person–Organization Fit, Job Choice Decisions, and Organizational Entry. *Organizational Behavior and Human Decision Processes*, *67*(3), 294–311. https://doi.org/10.1006/obhd.1996.0081

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in Cognitive*

Sciences, *7*(5), 225–231. https://doi.org/10.1016/S1364-6613(03)00094-9

Chang, E. H., Milkman, K. L., Chugh, D., & Akinola, M. (2019). Diversity Thresholds: How Social Norms, Visibility, and Scrutiny Relate to Group Composition. *Academy of Management Journal*, *62*(1), 144–171. https://doi.org/10.5465/amj.2017.0440

Chatman, J. A., Boisnier, A. D., Spataro, S. E., Anderson, C., & Berdahl, J. L. (2008). Being distinctive versus being conspicuous: The effects of numeric status and sex-stereotyped tasks on individual performance in groups. *Organizational Behavior and Human Decision Processes*, *107*(2), 141–160. https://doi.org/10.1016/j.obhdp.2008.02.006

Cohen, L. L., & Swim, J. K. (1995). The Differential Impact of Gender Ratios on Women and Men: Tokenism, Self-Confidence, and Expectations. *Personality and Social Psychology Bulletin*, *21*(9), 876–884. https://doi.org/10.1177/0146167295219001

Crocker, J., & McGraw, K. M. (1984). What's Good for the Goose Is Not Good for the Gander: Solo Status as an Obstacle to Occupational Achievement for Males and Females. *American Behavioral Scientist*, *27*(3), 357–369. https://doi.org/10.1177/000276484027003007

DeHaas, D., Akutagawa, L., Spriggs, S., & Deloitte. (2019, February 5). Missing Pieces Report: The 2018 Board Diversity Census of Women and Minorities on Fortune 500 Boards. Retrieved December 17, 2019, from https://corpgov.law.harvard.edu/2019/02/05/missing-pieces-report-the-2018-board-diversity-census-of-women-and-minorities-on-fortune-500-boards/

Dezső, C. L., Ross, D. G., & Uribe, J. (2016). Is there an implicit quota on women in top management? A large-sample statistical analysis. *Strategic Management Journal*, *37*(1), 98–115. https://doi.org/10.1002/smj.2461

Dipboye, R. L., & Colella, A. (2013). *Discrimination at Work: The Psychological and Organizational Bases*. Psychology Press.

Dovidio, J. F., Gaertner, S. L., & Saguy, T. (2008). Another view of "we": Majority and minority group perspectives on a common ingroup identity. *European Review of Social Psychology*, *18*(1), 296–330. https://doi.org/10.1080/10463280701726132

Duffy, M. K., Scott, K. L., Shaw, J. D., Tepper, B. J., & Aquino, K. (2012). A Social Context Model of Envy and Social Undermining. *Academy of Management Journal*, *55*(3), 643–666. https://doi.org/10.5465/amj.2009.0804

Duguid, M. (2011). Female tokens in high-prestige work groups: Catalysts or inhibitors of group diversification? *Organizational Behavior and Human Decision Processes*, *116*(1), 104–115. https://doi.org/10.1016/j.obhdp.2011.05.009

Ely, R. J. (1994). The Effects of Organizational Demographics and Social Identity on Relationships among Professional Women. *Administrative Science Quarterly*, *39*(2), 203–238. https://doi.org/10.2307/2393234

Floge, L., College, B., & Merrill, D. M. (1986). *Tokenism Reconsidered: Male Nurses and Female Physicians in a Hospital Setting*." *Social Forces 64:925*.

Flory, J. A., Leibbrandt, A., & List, J. A. (2015). Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions. *The Review of Economic Studies*, *82*(1), 122–155. https://doi.org/10.1093/restud/rdu030

Halevy, N., Cohen, T. R., Chou, E. Y., Katz, J. J., & Panter, A. T. (2014). Mental Models at Work: Cognitive Causes and Consequences of Conflict in Organizations. *Personality and Social Psychology Bulletin*, *40*(1), 92–110. https://doi.org/10.1177/0146167213506468

Hoffman, P. J., Festinger, L., & Lawrence, D. H. (1954). Tendencies toward Group Comparability in Competitive Bargaining. *Human Relations*, *7*(2), 141–159. https://doi.org/10.1177/001872675400700203

Kanter, R. M. (1977). Some Effects of Proportions on Group Life: Skewed Sex Ratios and Responses to Token Women. *American Journal of Sociology*, *82*(5), 965–990. https://doi.org/10.1086/226425

Keller, J., & Sekaquaptewa, D. (2008). Solo status and women's spatial test performance: The role of individuation tendencies. *European Journal of Social Psychology*, *38*(6), 1044–1053. https://doi.org/10.1002/ejsp.490

Kilduff, G. J. (2014). Driven to Win: Rivalry, Motivation, and Performance. *Social Psychological and Personality Science*, *5*(8), 944–952. https://doi.org/10.1177/1948550614539770

Kilduff, G. J., Elfenbein, H. A., & Staw, B. M. (2010). The Psychology of Rivalry: A

Relationally Dependent Analysis of Competition. *Academy of Management Journal*, *53*(5), 943–969. https://doi.org/10.5465/amj.2010.54533171

Lee, S. Y., Kesebir, S., & Pillutla, M. M. (2016). Gender differences in response to competition with same-gender coworkers: A relational perspective. *Journal of Personality and Social Psychology*, *110*(6), 869–886. https://doi.org/10.1037/pspi0000051

Lievens, F., Decaesteker, C., Coetsier, P., & Geirnaert, J. (2001). Organizational Attractiveness for Prospective Applicants: A Person–Organisation Fit Perspective. *Applied Psychology*, *50*(1), 30–51. https://doi.org/10.1111/1464-0597.00047

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation Analysis. *Annual Review of Psychology*, *58*(1), 593–614. https://doi.org/10.1146/annurev.psych.58.110405.085542

Martins, L. L., & Parsons, C. K. (2007). Effects of gender diversity management on perceptions of organizational attractiveness: The role of individual differences in attitudes and beliefs. *Journal of Applied Psychology*, *92*(3), 865. https://doi.org/10.1037/0021-9010.92.3.865

Maslach, C. (1974). Social and personal bases of individuation. *Journal of Personality and Social Psychology*, *29*(3), 411–425. https://doi.org/10.1037/h0036031

Maslach, C., Stapp, J., & Santee, R. T. (1985). Individuation: Conceptual analysis and

assessment. *Journal of Personality and Social Psychology*, *49*(3), 729–738. https://doi.org/10.1037/0022-3514.49.3.729

McGuire, W. J., McGuire, C. V., Child, P., & Fujioka, T. (1978). Salience of ethnicity in the spontaneous self-concept as a function of one's ethnic distinctiveness in the social environment. *Journal of Personality and Social Psychology*, *36*(5), 511–520. https://doi.org/10.1037/0022-3514.36.5.511

McGuire, W. J., & Padawer-Singer, A. (1976). Trait salience in the spontaneous self-concept. *Journal of Personality and Social Psychology*, *33*(6), 743–754. https://doi.org/10.1037/0022-3514.33.6.743

McKay, P. F., Avery, D. R., Tonidandel, S., Morris, M. A., Hernandez, M., & Hebl, M. R. (2007). Racial Differences in Employee Retention: Are Diversity Climate Perceptions the Key? *Personnel Psychology*, *60*(1), 35–62. https://doi.org/10.1111/j.1744-6570.2007.00064.x

McPherson, M., & Smith-Lovin, L. (1987). Sex Segregation in Voluntary Associations. *American Sociological Review*, *51*(1), 61–79. https://doi.org/10.2307/2095478

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Mehra, A., Kilduff, M., & Brass, D. J. (1998). At the Margins: A Distinctiveness Approach to the

Social Identity and Social Networks of Underrepresented Groups. *Academy of Management Journal*, *41*(4), 441–452. https://doi.org/10.5465/257083

Montoya, R. M., Horton, R. S., & Kirchner, J. (2008). Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity. *Journal of Social and Personal Relationships*, *25*(6), 889–922. https://doi.org/10.1177/0265407508096700

Nelson, L. J., & Miller, D. T. (1995). The Distinctiveness Effect in Social Categorization: You Are What Makes You Unusual. *Psychological Science*, *6*(4), 246–249. https://doi.org/10.1111/j.1467-9280.1995.tb00600.x

Niederle, M., & Vesterlund, L. (2007). Do Women Shy Away From Competition? Do Men Compete Too Much? *Quarterly Journal of Economics*, 35.

Niemann, Y. F., & Dovidio, J. F. (1998). Relationship of solo status, academic rank, and perceived distinctiveness to job satisfaction of racial/ethnic minorities. *The Journal of Applied Psychology*, *83*(1), 55–71.

Plass, J. L., O'Keefe, P. A., Homer, B. D., Case, J., Hayward, E. O., Stein, M., & Perlin, K. (2013). The impact of individual, competitive, and collaborative mathematics game play on learning, performance, and motivation. *Journal of Educational Psychology*, *105*(4), 1050–1066. https://doi.org/10.1037/a0032688

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 717–731. https://doi.org/10.3758/BF03206553

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

Ray, D., King-Casas, B., Montague, P. R., & Dayan, P. (2009). Bayesian Model of Behaviour in Economic Games. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21* (pp. 1345–1352). Retrieved from http://papers.nips.cc/paper/3589-bayesian-model-of-behaviour-in-economic-games.pdf

Roulin, N., Bangerter, A., & Yerly, E. (2011). The Uniqueness Effect in Selection Interviews. *Journal of Personnel Psychology*, *10*(1), 43–47. https://doi.org/10.1027/1866-5888/a000024

Scheiber, N. (2015). Work Policies May Be Kinder, but Brutal Competition Isn't. Retrieved June 26, 2019, from The New York Times website: https://www.nytimes.com/2015/08/18/business/work-policies-may-be-kinder-but-brutal-competition-isnt.html

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and non-experimental studies: New procedures and recommendations. *Psychological Methods*, 422–445.

Singleton, R. A., & Vacca, J. (2007). Interpersonal Competition in Friendships. *Sex Roles*, *57*(9–10), 617–627. https://doi.org/10.1007/s11199-007-9298-x

Snyder, C. R., & Lopez, S. J. (2001). *Handbook of Positive Psychology*. Oxford University Press.

Stangor, C., Lynch, L., Duan, C., & Glas, B. (1992). Categorization of individuals on the basis of multiple social features. *Journal of Personality and Social Psychology*, *62*(2), 207. https://doi.org/10.1037/0022-3514.62.2.207

Steinhage, A., Cable, D., & Wardley, D. (2017). The pros and cons of competition among employees. *Harvard Business Review*, 2–5.

Tajfel, H., & Turner, J. (1979). *An Integrative Theory of Intergroup Conflict*.

Thompson, M., & Sekaquaptewa, D. (2002). When Being Different Is Detrimental: Solo Status and the Performance of Women and Racial Minorities. *Analyses of Social Issues and Public Policy*, *2*(1), 183–203. https://doi.org/10.1111/j.1530-2415.2002.00037.x

Turban, D. B., Dougherty, T. W., & Lee, F. K. (2002). Gender, Race, and Perceived Similarity Effects in Developmental Relationships: The Moderating Role of Relationship Duration. *Journal of Vocational Behavior*, *61*(2), 240–262. https://doi.org/10.1006/jvbe.2001.1855

Turban, D. B., & Greening, D. W. (1997). Corporate Social Performance And Organizational Attractiveness To Prospective Employees. *Academy of Management Journal*, *40*(3), 658–672. https://doi.org/10.5465/257057

Turban, D. B., & Keon, T. L. (1993). Organizational attractiveness: An interactionist perspective. *Journal of Applied Psychology*, *78*(2), 184. https://doi.org/10.1037/0021-9010.78.2.184

Umphress, E. E., Smith-Crowe, K., Brief, A. P., Dietz, J., & Watkins, M. B. (2007). When birds

of a feather flock together and when they do not: Status composition, social dominance orientation, and organizational attractiveness. *Journal of Applied Psychology*, *92*(2), 396–409. https://doi.org/10.1037/0021-9010.92.2.396

Watkins, M. B., Simmons, A., & Umphress, E. (2019). It's Not Black and White: Toward a Contingency Perspective on the Consequences of Being a Token. *Academy of Management Perspectives*, *33*(3), 334–365. https://doi.org/10.5465/amp.2015.0154

Yoder, J. D. (1991). Rethinking Tokenism: Looking beyond Numbers. *Gender and Society*, *5*(2), 178–192.

Yoder, J. D., & Sinnett, L. M. (1985). Is it all in the Numbers? A Case Study of Tokenism. *Psychology of Women Quarterly*, *9*(3), 413–418. https://doi.org/10.1111/j.1471-6402.1985.tb00890.x

Tables

Table 2, Chapter 2. Full Correlation Table for Study 3

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
(1): "I think my work or performance will be distinct from that of other interns in my department"	1.0						
(2): "I think I bring a unique perspective to my department"	0.48***	1.0					
(3): "I think managers will want to ensure that at least one woman receives a full-time job from each department"	0.09	0.18***	1.0				
(4): "I think managers will be reluctant to give a full-time job only to men in each department"	0.14**	0.07	0.43***	1.0			
(5): "I feel tense competing against women"	0.00	-0.02	0.21***	0.27***	1.0		
(6): "I don't feel as comfortable competing against women as I do competing against men"	-0.01	-0.08	0.17***	0.18***	0.62***	1.0	
(7): Choice of the all-male group	0.21***	0.20***	0.15***	0.19***	0.31***	0.28***	1.0

† $p < 0.10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 3, Chapter 2. Summary Table of Results Across All Studies

	Total N	**Proportion choosing to be tokens in the *competitive* condition**	**Proportion choosing to be tokens in the *control* condition**	***z*-statistic for difference in proportions**	***p*-value for difference in proportions**
Study 1a	491	.461	.175	6.72	<.001
Study 1b	278	.366	.199	2.97	.003
Study 2	592	.228	.091	3.59	<.001
Study 3	396	.374	.192	3.91	<.001
Study 4	145	.233	.097	1.97	.048
Study 5	583	.416	.321	2.27	.023

Figures

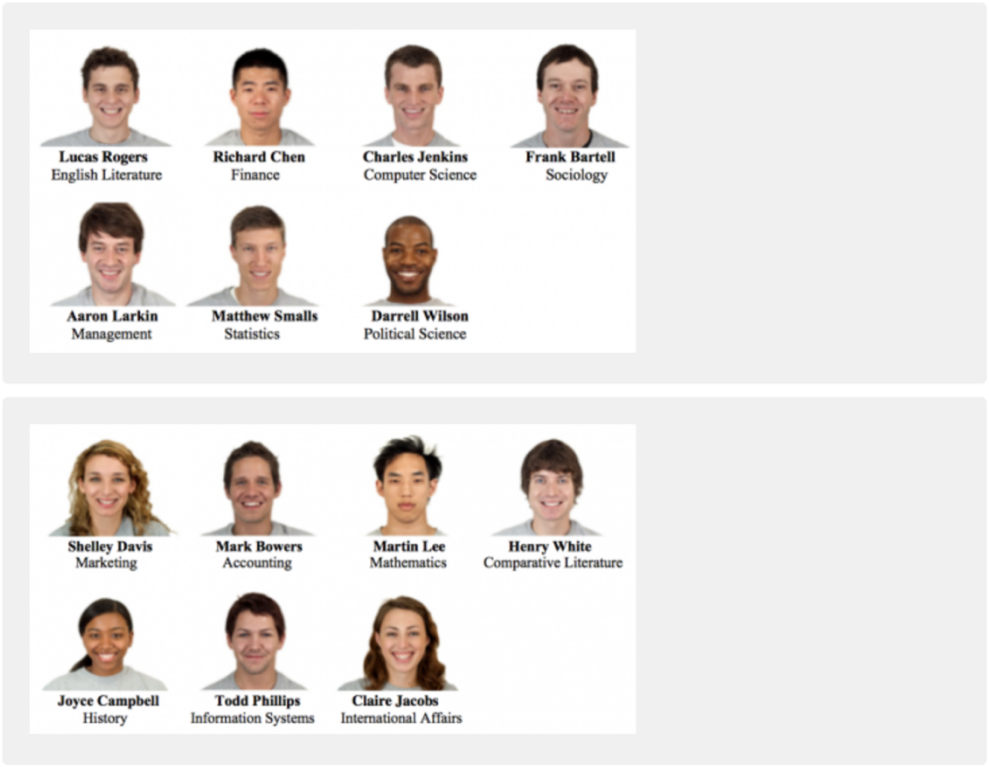
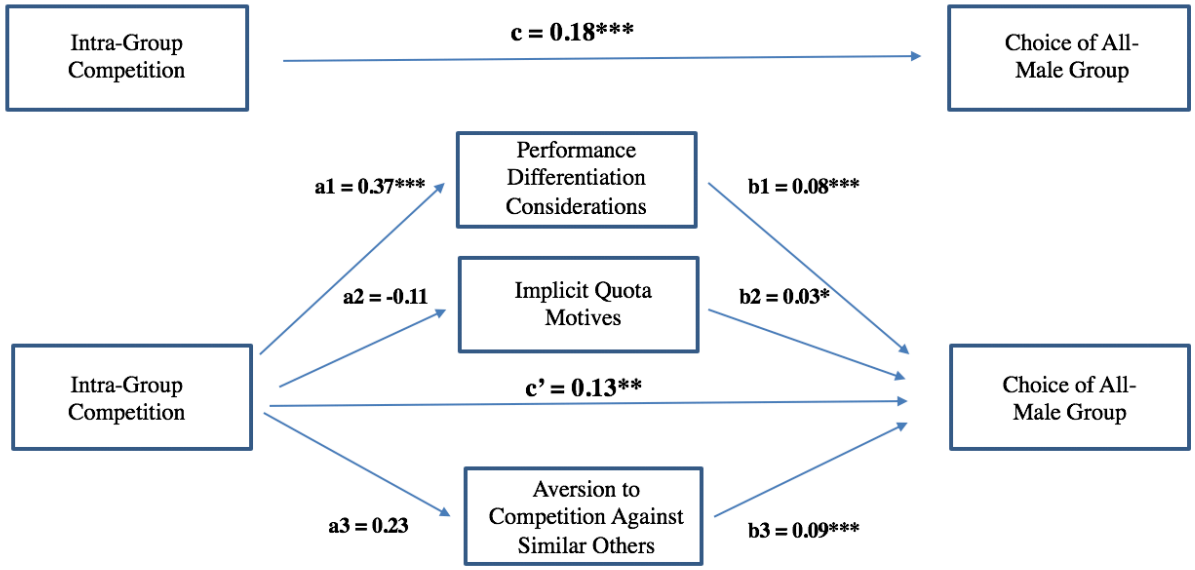


Figure 4, Chapter 2 | Example Stimuli from Study 1A. This is an example of the stimuli displayed to participants in Study 1A. The order of presentation of the two groups was randomized across participants. Racial diversity was held constant across the two groups, and college majors were matched across groups such that the majors in each group were similar but not identical (e.g., Computer Science vs. Information Systems), as presenting groups with identical majors could have appeared suspicious to participants.



* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 5, Chapter 2 | Mediation results from Study 3. Results of our Study 3 multiple mediator analysis Study 3 showed that performance differentiation mediated the relationship between intra-group competition and choice of the all-male group. Meanwhile implicit quota considerations and aversion to ingroup competition did not mediate choice of the all-male group.

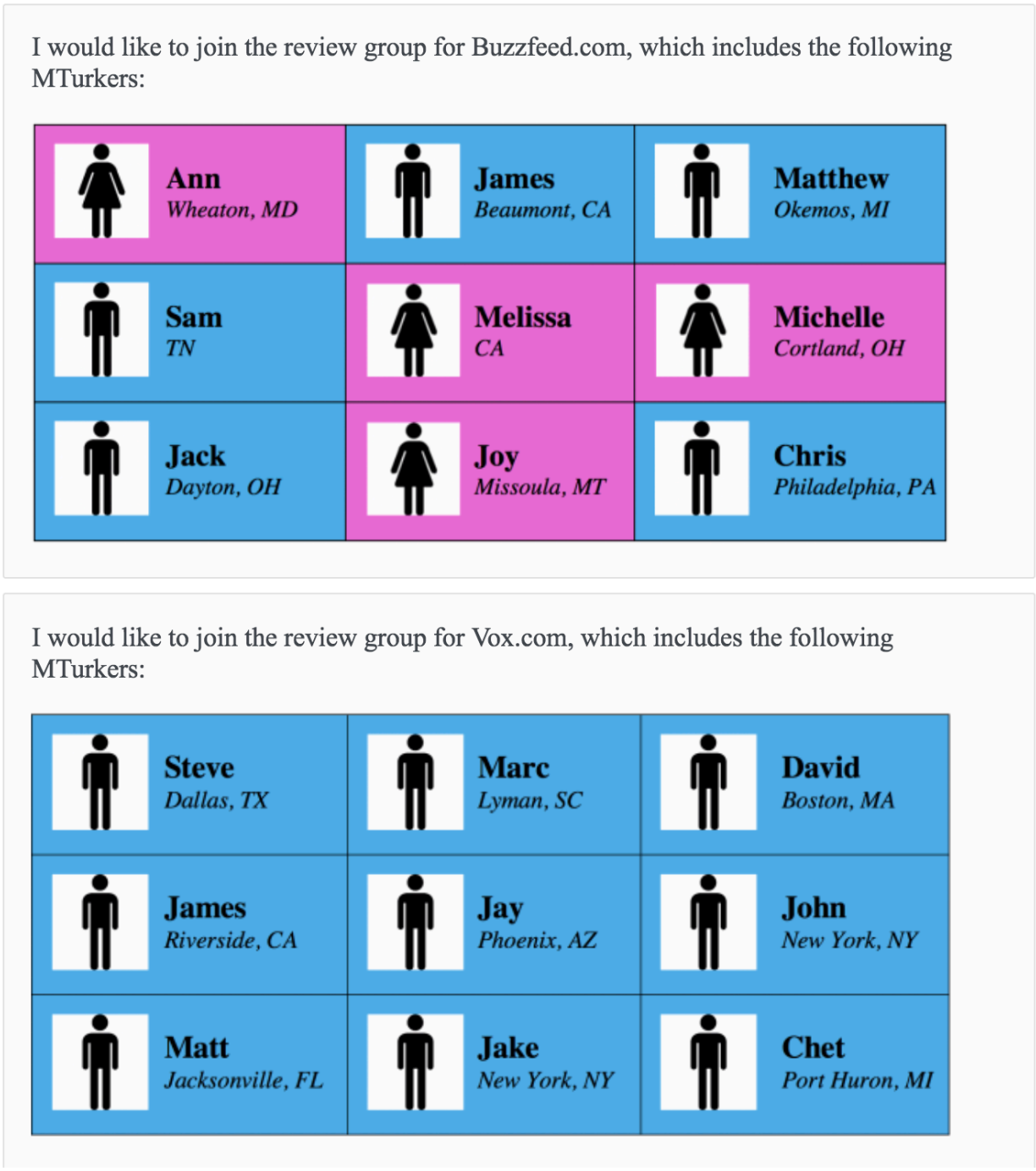


Figure 6, Chapter 2 | Example Stimuli from Study 5. This is an example of the stimuli displayed to participants in Study 5. Here we show two of the groups out of three pairs of groups from which we randomly sampled stimuli. Each group was associated with a randomly selected website from a set of four websites – Buzzfeed, HuffingtonPost, Vice, and Vox. Participants were asked to choose which of the two groups they wanted to join.

CHAPTER 3: THE INFLUENCE OF POSITIVE AND NEGATIVE FEEDBACK ABOUT BIAS ON SUBSEQUENT DISCRIMINATION

Erika L. Kirgios

Working Paper

ABSTRACT:

Does making people aware of their biases reduce prejudice? Most people aim to feel or appear unprejudiced, so learning they have failed to do so should motivate behavior change. However, people may discount such feedback in an effort to protect their moral self-image. Instead, it might be better to make people aware of their *egalitarian* actions to encourage future conformity with past good behavior. In a two-stage audit experiment with U.S. city councilors (3,981 current and 885 former), I test whether people are more likely to offer support to racial minorities after receiving positive, negative, or no feedback about racial bias in their professional ingroup. Compared to a holdout control group that received no feedback, negative feedback emphasizing recent evidence of racial discrimination in city councils did not influence city councilors' willingness to provide career advice to Black men. Positive feedback emphasizing recent evidence of pro-diversity behavior in city councils, however, increased current city councilors' willingness to provide career advice to Black men by 36.3%. Additionally, city councilors who received positive feedback used significantly more polite language in responses to Black men. Heterogeneity analyses suggest these results may be driven by perceptions of descriptive norms: people who learn about a lack of bias in their professional ingroup exert more effort to support racial minorities to keep up with their colleagues' pro-diversity behavior. Prejudice reduction efforts may benefit from urging people to maintain ongoing pro-diversity efforts rather than to rectify past discrimination.

Link to *Online Supplement*, data, and code*:*

https://osf.io/7qnvf/?view_only=a0d1b144ca8e45ec93555552b404baa5.

Introduction

Prejudice is a sticky, pervasive problem (Paluck, Porat, Clark, & Green, 2021; Bertrand & Duflo, 2017). Among many other situations, marginalized group members face discrimination when applying to jobs, completing work tasks, seeking career advice, interacting with police, renting AirBnBs, and purchasing a car (Bertrand & Mullainathan, 2004; Milkman, Akinola, & Chugh, 2015; Edelman, Luca, & Svirsky, 2017; Ayres, 1991; Butler & Crabtree, 2017; Glover, Pallais, & Pariente, 2017; Donohue & Levitt, 2001; Voigt et al., 2017). Decision-makers often propose education as a prejudice-reduction strategy. Acting on the intuition that people can't correct their biases without being aware of them, organizations hold diversity trainings to teach employees about stereotyping and discrimination (Dobbin & Kalev, 2016), and news stories and individual activists call out instances of prejudice to highlight bad behavior (e.g., Wolfers, 2015; Solano & Robson, 2020; Holpuch, 2022). However, people often respond defensively when accused of exhibiting racial bias (Vitriol & Moskowitz, 2021), drawing attention to their Black friends and mixed-race grandchildren to "prove" they are not racist (Hains, 2019). Despite this defensiveness, can making people aware of their biases lead to behavior change?

On the one hand, sharing information about bias may be an effective bias-reduction strategy given that many people are motivated to avoid prejudiced behavior, whether it's to uphold their values or their reputation (Plant & Devine, 1998; Apfelbaum, Sommers, & Norton, 2008; Plant & Devine, 2009). In particular, informing people that they harbor biased attitudes or exhibit discriminatory behaviors should highlight a failure to avoid prejudice; this discrepancy between desired and actual behavior should, in turn, motivate people to increase their efforts towards egalitarian goals (Mullin & Monin, 2016; Locke & Latham, 1990; Monteith, 1993; Monteith, Ashburn-Nardo, Voils, & Czopp, 2002). Lab studies have demonstrated that after people are confronted with their own biased behavior they tend to report more attempts to suppress their biases, stronger egalitarian attitudes, and greater intentions to monitor their future

behaviors to ensure they are egalitarian (Parker, Monteith, Moss-Racusin, & Van Camp, 2018; Czopp, Monteith, & Mark, 2006; Monteith et al., 2002; Chaney & Sanchez, 2018). This theorizing and evidence points to a potential benefit of informing people about bias and discrimination. However, the story may not be so simple: it is likely to be deeply distressing to hear you may be prejudiced, destabilizing your sense of self as a moral, egalitarian person (Sherman & Cohen, 2006). This self-image threat could lead people to deny or minimize the information which, in turn, is likely to mitigate its positive impact on their behavior—even as they profess to hold stronger egalitarian attitudes in an attempt to deny that they are biased (Eskreis-Winkler & Fishbach, 2019; Levy & Maaravi, 2018).

Notably, extant research almost exclusively focuses on sharing information in the form of negative feedback, highlighting the prevalence of bias in one's professional ingroup. Instead, it might be beneficial to provide people with *positive* feedback, or information that spotlights the presence of pro-diversity behavior in one's professional ingroup. Positive feedback—which can reaffirm one's moral, unprejudiced self-image—may motivate efforts towards egalitarianism by encouraging consistency with prior good behavior (Mullin & Monin, 2016). Unlike negative feedback, this positive information is less likely to lead to a self-image threat and, therefore, to be ignored. On the other hand, positive feedback may create a moral licensing effect, reassuring people that they sufficiently affirmed their egalitarian values through past good behavior and can afford to reduce future efforts to support racial minorities (Mullin & Monin, 2016; Monin & Miller, 2001).

In this work, I conducted a large-scale field experiment with 3,981 current city council members serving in 765 of the largest cities in the U.S. (885 former city council members were also in the experiment and their data is included in robustness checks). I explored whether city councilors responded more often and more positively to Black students seeking career advice

after receiving positive, negative, or no feedback about racial bias in their professional ingroup. I first contacted city council members via email to share a public service announcement (PSA) containing a positive or negative message about evidence that city councilors tend to exhibit racial bias (city councilors in a control condition received no email). Specifically, PSA emails containing positive feedback highlighted recent research which suggests city councilors support Black constituents when their minority status is highlighted (Kirgios et al., 2022); PSA emails containing negative feedback highlighted recent research which suggests city councilors discriminate against Black constituents (Butler & Crabtree, 2017). The next day, city councilors received emails from (fictitious) Black or White male students seeking career advice. As an exploratory analysis, I examined the politeness of city councilors' replies to students. People in positions of power are often significantly less polite to Black people relative to White people, so I explored whether providing positive and negative feedback might mitigate this bias, reducing prejudice at both the intensive and extensive margins (Voigt et al., 2017).

Both positive and negative feedback may increase the salience of racial equity, which can in turn increase people's willingness to engage in pro-diversity behavior (Kirgios et al., 2022; Chang, Kirgios, Rai, & Milkman, 2020). However, positive and negative feedback have otherwise divergent psychological consequences that might influence their effectiveness, which I describe below.

The Psychological and Behavioral Consequences of Negative Feedback

Negative feedback, or feedback that reports evidence of discriminatory behaviors in one's professional community, may improve behavior towards racial minorities. People often underestimate their vulnerability to bias (Pronin, Lin, & Ross, 2002). So, it's likely that at baseline, people who are internally and externally motivated to control prejudice believe they are

successfully avoiding discriminatory behavior (Pronin et al., 2002; Plant & Devine, 1998; Ehrlinger, Gilovich, & Ross, 2005). Learning that they've fallen short of these goals may lead them to engage in compensatory behavior, seeking opportunities to support racial minorities to make up for past bias (Mullen & Monin, 2016; Locke & Latham, 1990; Eskreis-Winkler & Fishbach, 2020). In other words, negative feedback about bias may highlight the discrepancy between the equitable person you mean to be and your actual behavior, motivating you to seek pro-diversity actions. In fact, models of prejudice reduction have posited that being confronted with one's biases is critical for inspiring future equitable behavior (Monteith, 1993).

Some research supports this prediction (e.g., Son Hing, Bobocel, & Zanna, 2002, Pope, Price, & Wolfers, 2018; Chaney, Sanchez, Alt, & Shih, 2021; Parker et al., 2018; Chaney & Sanchez, 2018). For example, lab participants who were taught about the prevalence of workplace bias increased their support for affirmative action policies (Son Hing et al., 2002). And after a paper finding evidence of racial in-group bias amongst National Basketball Association (NBA) referees gained widespread media attention, the referees no longer exhibited the bias (Pope et al., 2018). While changes to NBA policy may have spurred this bias reduction rather than changes to individual behavior, these results suggest that awareness can reduce prejudice (Pope et al., 2018).

But there is also reason to doubt that negative feedback will actually motivate equitable behavior. In general, people learn less from negative feedback than positive feedback because receiving negative feedback undermines their self-confidence and leads to disengagement with the learning experience (Eskreis-Winkler & Fishbach, 2019). For example, people learning a new (invented) language learned less when their incorrect guesses were flagged than when their correct guesses were flagged, even though both forms of feedback conveyed the same information (Eskreis-Winkler & Fishbach, 2019). People struggle to learn from failure because their ego gets in the way.

Negative feedback about one's tendency to discriminate is likely even harder for people to accept. People who are informed about past discriminatory behavior may experience a profound threat to their self-image as a moral, good person (Sherman & Cohen, 2006). Importantly, people tend to experience threat whether the negative feedback is individualized or applies to their entire community, particularly if it's a community with which they highly identify (Sherman & Cohen, 2006; Terry & Hogg, 1996). This ego threat may lead them to discount or reject the information in order to protect their sense of self (Epley & Gilovich, 2016; Sherman & Cohen, 2006; Festinger, 1957; Schumann & Dweck, 2014). For example, White participants who were informed they would likely get a low score on the IAT (Implicit Association Test)—indicating high levels of prejudice—were more likely to decline to receive their IAT score than participants who were given no warning, preferring to not receive further evidence confirming their biases (Howell et al., 2013). Ultimately, research on ego threat suggests people are likely to respond defensively to negative feedback, either ignoring the information—and failing to change their behavior—or even justifying their actions and entrenching existing biases (Eskreis-Winkler & Fishbach, 2019; Howell, Redford, Pogge, & Ratliff, 2017; Levy & Maaravi, 2018).

Field evidence on the effectiveness of bias awareness interventions provides support for this pessimistic view. Diversity trainings rarely inspire behavior change, even when they lead to positive attitude change (Chang et al., 2019; Dobbin & Kalev, 2016; Kalev, Dobbin, & Kelly, 2006; Bezrukova, Jehn, & Spell, 2012; Bezrukova, Spell, Perry, & Jehn, 2016). People also don't seem to respond to information about prejudice in their professional circles. Butler and Crabtree (2017) emailed thousands of politicians asking them to participate in a survey about racial bias. In that email, they varied whether they provided negative feedback—sharing recent research suggesting that politicians display favoritism towards racial in-group members when handling requests from constituents—or not. Two weeks later, they conducted an audit experiment,

sending emails with practical requests (e.g., “How can I find out how well schools in our district are performing?”, “How should my nephew pay for his speeding ticket?”) from either White or Black constituents (Butler & Crabtree, 2017). Politicians were significantly less likely to respond to Black constituents relative to White constituents, and the negative feedback failed to produce even a directional reduction in this racial bias. However, if salience of racial equity drives some of the benefits of delivering feedback, the time lag between feedback delivery and measurement of behavior may have muted its effectiveness, making this a conservative test (Kirgios et al., 2022; Chang et al., 2020).

Ultimately, the evidence for the effectiveness of negative feedback remains mixed. On the one hand, it may motivate people to provide support to racial minorities to make up for past bias that they or their ingroup members have exhibited. On the other hand, the ego threatening nature of negative feedback may undermine its effectiveness as people ignore or dismiss the feedback in favor of maintaining a positive self-image.

The Psychological and Behavioral Consequences of Positive Feedback

Positive feedback about bias may prove more motivating. This information is unlikely to be rejected; instead, it may be interpreted as evidence of commitment to egalitarianism. At the individual level, perceptions of commitment can motivate consistency as people strive to continue upholding their newly reaffirmed egalitarian identity (Fishbach & Dhar, 2005; Mullen & Monin, 2016). At the group level, this evidence of commitment may create a stronger descriptive norm of non- or anti-racism. Sharing descriptive norms can motivate behavior change as people use peer behavior as a signal of how it is acceptable or appropriate to behave (Goldstein, Cialdini, & Griskevicius, 2008; Goldstein & Cialdini, 2007; Prentice & Paluck, 2020).

Descriptive norms may be particularly likely to encourage behavior change in the domain of prejudice where people might find deviations from socially acceptable behavior particularly risky. Being labeled as prejudiced is highly aversive, and people often adjust their behavior in social settings in an attempt to appear unbiased (e.g., by avoiding naming someone's race, even when it's relevant; Apfelbaum et al., 2008; Norton et al., 2006). In fact, local norms often shape people's attitudes and actions towards racial minorities (Stephan & Stephan, 1985; Duguid & Thomas-Hunt, 2015; Paluck & Green, 2009; Crandall & Stangor, 2005; Crandall, Eshleman, & O'Brien, 2002; Munger, 2017). For example, people who learn that stereotyping is rare express significantly less agreement with stereotypical statements (e.g., women are less career-oriented than men, older people are weaker than younger people) than those who learn that stereotyping is common (Duguid & Thomas-Hunt, 2015). Similarly, people exposed to a confederate with non-racist views expressed significantly more willingness to punish the writers of anonymous hateful letters sent to Black students on campus (Blanchard, Lilly, & Vaughn, 1991). Blanchard and colleagues (1991) asked students to participate in a survey about how their university should handle anonymous racist notes to students. Participants who heard confederates strongly advocate for punishing the perpetrators expressed significantly less agreement with racist statements than those exposed to neutral confederates (Blanchard et al., 1991). In essence, people's racial attitudes were swayed by the norms they observed their peers setting. Norms are often considered a necessary precondition of prejudice: people express prejudice to the extent that it's viewed as acceptable by their peers (Pettigrew, 1991; Crandall et al., 2002; Munger, 2017). Notably, most research on the relationship between norms and prejudice has focused on the explicit expression of prejudice in people's attitudes or statements. It's not clear whether information about norms will successfully shift behaviors, which often diverge from explicitly stated attitudes or intentions in the domain of prejudice (Chang et al., 2019; Kawakami, Dunn, Karmali, & Dovidio, 2009).

Because descriptive norms set a standard of socially acceptable behavior, they likely have the strongest influence on people who worry their behavior may receive negative scrutiny. Following norms ensures that people's behavior blends in with that of the majority, reducing the likelihood that they will stick out as bad actors and face social sanctions (Schultz et al., 2007). People who are worried about being labeled racists might therefore be motivated to fall in line with normative behavior to ensure they avoid punishment. Thus, to the extent the positive feedback about lack of bias in one's community shapes behavior by shifting perceptions of descriptive norms, it should have a stronger impact on people who are more concerned about appearing unprejudiced. So, for example, current public figures—who may expect more scrutiny relative to former public figures—and majority group members—who may be more worried about being labeled racists than minority group members—should be particularly likely to increase support for racial minorities after receiving positive group-level feedback (Apfelbaum et al., 2008; Chang et al., 2019; Norton et. al., 2006).

On the other hand, positive feedback may ultimately undermine pro-diversity behavior. Learning about prior support for racial minorities may lead to moral licensing: if people feel that their past actions—or those of ingroup members—effectively establish that they are not prejudiced, they may be comfortable subsequently exerting less effort towards egalitarian goals or even openly expressing bias (Mullen & Monin, 2016; Effron, Cameron & Monin, 2009; Monin & Miller, 2001; Kouchaki, 2011; Nguyen, 2021). Moral licensing is common in the domain of prejudice. For example, people given the opportunity to establish their credentials as a "non-racist" by disagreeing with blatantly racist statements or endorsing Barack Obama were subsequently more willing to favor White people during a simulated hiring process (Monin & Miller, 2001; Effron et al., 2009). Moral licensing can even happen vicariously: people who

witnessed the non-prejudiced behavior of professional ingroup members subsequently behaved in a more prejudiced manner themselves (Kouchaki, 2011).

Even if people don't feel licensed to behave badly after hearing positive feedback about lack of bias in their pofessional ingroup, they may take it as a signal that they and their colleagues have made sufficient progress on their racial equity goals (Fishbach & Dhar, 2005). Believing they've made progress on racial equity, they may reduce further efforts towards supporting racial minorities and turn their energy towards another goal, whether it's an adjacent goal—e.g., supporting women—or a distal goal—e.g., improving local parks (Fishbach & Dhar, 2005). Ultimately, moral licensing and progress perceptions would predict that positive feedback should either prove ineffective or backfire.

Thus, existing theory and empirical evidence provides competing predictions for the psychological effects of positive feedback about lack of bias in one's ingroup. Positive feedback may motivate people to provide support for racial minorities because they want to conform to descriptive norms or behave consistently with past good behavior. On the flip side, positive feedback may prove ineffective or backfire because people feel they (and their colleagues) have made sufficient progress towards promoting racial equity and can relax their efforts to support racial minorities.

Overview of the Current Work

I conducted a field experiment with U.S. city council members examining the quantity and quality of help they provided to Black vs. White men after receiving positive, negative, or no feedback about racial bias on city councils. City council members are local elected politicians responsible for handling city budgets and managing city programs, such as the fire department, parks and recreation department, and police department. I chose city council members as the

population for my experiment because prior audit experiments have found evidence that local and state politicians discriminate against Black constituents (Butler & Broockman, 2011; Butler & Crabtree, 2017; Kirgios et al., 2022). Specifically, relative to White male constituents, Black male constituents receive fewer responses from politicians to emails requesting information about how to vote, asking questions about basic city functions, or seeking career advice (Butler & Broockman, 2011; Butler & Crabtree, 2017; Kirgios et al., 2022). I focused on response rates to advice-seeking emails because prior research has documented discriminatory gaps in willingness to offer career advice to racial minorities relative to White men (Milkman et al., 2015; Kirgios et al., 2022), and such mentorship can be critical for long-term career success (Eby et al., 2008; Seibert, Kraimer, & Liden, 2001).

City councilors in the feedback conditions were either told about: (1) new evidence suggesting that city councilors discriminate against Black constituents, replying to them less often than White constituents (negative feedback); or (2) new evidence suggesting that city councilors support Black constituents, replying more often when their minority status is highlighted (positive feedback). Relative to delivering no feedback, negative feedback did not significantly influence subsequent response rates to Black men seeking career advice. Meanwhile, positive feedback increased response rates to Black men seeking career advice by a regression-estimated 36.3% relative to the no feedback control. This effect was muted to marginal significance when including former city councilors (people who had stepped down, retired, or been replaced shortly before the experiment was conducted). When combined, current and former city councilors were a regression-estimated 29.4% more likely to respond to Black students after receiving positive feedback relative to when they received no feedback at all. Moreover, city councilors who received positive feedback used significantly more polite language in their responses to Black men than city councilors who did not receive any feedback.

This work responds to recent calls to examine the effectiveness of prejudice reduction strategies in field experiments with consequential, behavioral dependent variables (Paluck et al., 2021). The vast majority of work on prejudice reduction takes place in the lab and measures attitudes and behavioral intentions (Paluck et al., 2021). Unfortunately, people's behavior in race-related contexts often diverges from their intended and predicted behavior, suggesting that it's particularly important to collect behavioral dependent variables when testing interventions to reduce prejudice (Kawakami et al., 2009).

I also begin to elucidate how we should best frame information about bias in order to motivate future positive behavior towards racial minorities. I resolve competing predictions about the effectiveness of bias feedback on downstream prejudice, finding (a) null effects of negative feedback, consistent with Butler and Crabtree (2017) and (b) benefits of positive feedback. These findings suggest that the laboratory literature on the benefits of confronting prejudice may be finding evidence of demand effects rather than true behavioral change (e.g., Chaney et al., 2021; Czopp et al., 2006; Chaney & Sanchez, 2018; Parker et al., 2018). In the field, ego threat seems to dominate people's psychological reaction to negative feedback about bias. As a result, instead of adjusting their behavior in the face of negative feedback, people seem to ignore or dismiss information that implies they may exhibit racial bias. Finally, I add to the literature on social norms and prejudice by demonstrating that information about positive peer behavior doesn't just influence explicitly stated attitudes, but also subtly measured, consequential real-world behaviors (Gomez et al., 2018; Patel, 2013; Paluck et al., 2021). This work also offers an actionable insight for practitioners: highlighting instances of good, egalitarian behavior may be the best way to improve support for racial minorities.

Field Experiment with Politicians

In a preregistered audit experiment,[13] I set out to test whether city councilors respond more often and more positively to Black men requesting career advice after receiving positive, negative, or no feedback about racial bias in their professional ingroup.

Methods

Participants included 3,981 current city councilors (36.6% female, 18.0% Black, 10.5% Latinx, 70.6% White, 0.9% Other)[14] who were serving in 765 of the most populous cities across the United States when this experiment was conducted in March 2022. Information about city council members was originally collected in 2019-2020 and updated in early 2022. As a result, 885 former city councilors (people who had retired, stepped down, or been replaced since 2019) were also included in the experiment. Current city councilors are the primary focus of my analysis because my goal was to test the behavioral effects of receiving feedback about bias in one's professional ingroup. Since the feedback is centered on the behavior of city council members, it is likely to resonate more with those who still occupy and identify with the role (Terry & Hogg, 1996; Goldstein & Cialdini, 2007). However, I also present results including the full, preregistered sample as robustness checks.

The experiment included two stages, the feedback delivery stage and then the help-seeking audit stage. The Stage 1 and Stage 2 emails were sent one day apart.

---

[13] Preregistration link: https://aspredicted.org/R7X_M4M. All data and code from this experiment is in this OSF link: https://osf.io/7qnvf/?view_only=a0d1b144ca8e45ec93555552b404baa5.

[14] A team of research assistants categorized each city councilor's identity based on publicly available information (i.e., photos and bios on city council websites, interviews in local news, membership in affiliation groups and social media sites). When two research assistants disagreed about a city councilor's identity, the author resolved the disagreement.

Stage 1: Feedback Delivery. In the first stage, city councilors received email PSAs with feedback about bias in their professional ingroup. Given that both positive and negative feedback may effectively improve behavior—or backfire—and that their effects are likely to be driven by independent mechanisms, I included a no feedback control condition as a benchmark.

City councilors were randomly assigned to receive either no email (*no feedback control* condition), an email with negative feedback about racial bias in their ingroup (*negative feedback* condition), or an email with positive feedback about a lack of racial bias in their ingroup (*positive feedback* condition). These emails were sent on March 2nd, 2022. Each email came from the "Equity and Inclusion Research Lab", a group I created for the purposes of this experiment. The emails included a link to the lab's website in the email signature.[15] On the website, the lab was described as "committed to conducting and communicating research related to diversity, equity, and inclusion."

City councilors were assigned to a bias feedback condition through clustered random assignment by city, such that each city councilor in the same city received the same feedback (see Figure 7 to visualize the random assignment strategy). Randomization was conducted at the city level to account for potential spillover effects if city councilors decided to share the information with their colleagues given its importance and potentially upsetting nature.

City councilors assigned to the *no feedback control* condition were not contacted on March 2nd, 2022. Meanwhile, city councilors assigned to the *negative feedback* condition received an email that informed them about new research suggesting that city councilors were

[15] Due to rate limits on email accounts, I sent these emails from two accounts with slightly different names: communications@equityresearchlab.org and comms@equityresearchlab.org. However, all emails were signed "The Equity and Inclusion Research Group." I included the lab's website URL in all emails. This led to some emails bouncing because they were flagged as suspicious by the city's email servers, so I manually re-sent about 1200 emails without the lab's website in the signature.

less likely to provide support to Black constituents relative to White constituents. They received emails with subject lines reading: "City councilors discriminate against Black constituents." Those assigned to the *positive feedback* condition received an email that informed them about new research suggesting that city councilors were more likely to provide support to Black male constituents who highlighted their minority status. They received emails with subject lines reading: "City councilors support Black constituents." Importantly, both pieces of information were true (see: Butler & Crabtree, 2017 and Kirgios et al., 2022). The negative and positive feedback emails went on to say: "This research indicates that **city councilors (like you) can and do [contribute to unequal] / [actively improve] outcomes for racial minorities.**" The full email text is included in Figure 8.

Stage 2: Help-Seeking Audit. In the second stage, each city councilor received a help-seeking email from either a Black male student (*Black help seeker* condition) or White male student (*White help seeker* condition). These help-seeking emails were delivered on the morning of March 3rd, 2022. Each email came from a (fictitious) college student who expressed an interest in pursuing a role in local politics. In the email, the student briefly described their interest in politics and requested some advice about how to kick-start their involvement.

Random assignment to the *Black* or *White help seeker* conditions was conducted at the individual level through blocked random assignment by city. As in prior audit experiments, I manipulated the identity of the help seeker through a name that strongly signaled racial identity (e.g., Bertrand & Mullainathan, 2004; Bertrand & Duflo, 2017; Edelman et al., 2017; Milkman et al., 2015; Butler & Broockman, 2011; Butler & Crabtree, 2017; Kirgios et al., 2022). Specifically, I identified common surnames with a high likelihood ratio of belonging to Black families and White families, respectively, and I used online baby name lists to generate names that were likely to belong to Black men and White men. I then pre-tested these names to ensure high gender and

race recognizability (see Supplement page 8 for more details on the pre-testing procedure and results). Ultimately, 24 names (12 for each racial group) were included in the experiment. The student's name was repeated in their email address, the first line of their help-seeking email, and in the email signature in order to increase the likelihood that the recipient would see the name (see stimuli from both Stage 1 and Stage 2 in Supplement pages 3 and 6).

I discovered during pre-testing that including software to track open rates increased the likelihood that emails would be flagged as spam, so I did not include the tracking software in the Stage 1 or Stage 2 emails. As a result, all analyses were conducted intent-to-treat. Any city councilor who received a Stage 2 email is included in my analyses; only city councilors whose emails failed to send (e.g., because their email address was no longer in use) are excluded.

Further details about the methods and results are included in the Supplement.

Primary Dependent Variable. The preregistered primary dependent variable of interest was whether a city councilor replied to the Stage 2 help-seeking emails within 24 hours. Automatic replies, emails from aides and assistants, and replies that arrived more than 24 hours after the help-seeking email were not considered replies, as preregistered. I preregistered a 24-hour response window due to concerns about contagion: given the potentially inflammatory nature of the negative feedback, I worried that city councilors might circulate the messages outside of their city (e.g., by posting on social media) over time. Moreover, I expected any benefits of feedback to wane over time as the information became less salient. Exploratory response windows of 48 hours and 1 week are discussed in the Results.

Exploratory Dependent Variable. To investigate the impact of feedback on the politeness of city councilors' replies, I used the Linguistic Inquiry Word Count platform (LIWC-22) (Pennebaker, Boyd, Booth, Ashokkumar, & Francis, 2022). The LIWC-22 platform uses

validated dictionaries to identify the frequency of, for example, terms that convey anger in a text. When reporting the frequency of angry language in a text, LIWC-22 reports the proportion of total words in the text that conveyed anger (i.e., the proportion of words in the text that are in their anger dictionary). So, for example, an anger score of 12.5 suggests that 12.5% of a text's words or phrases connote anger. Thus, LIWC-22 scores can take values between 0 and 100. In my analyses, I focus on the proportion of polite language included in the city councilors' replies given recent research demonstrating that people in positions of power are significantly less polite when talking to Black citizens relative to White citizens (Voigt et al., 2017). The LIWC-22 politeness dictionary includes words or phrases that have been validated as signaling courteous speech (e.g., "Thank you", "Good morning").

Results

Summary statistics of participant characteristics for current city councilors and the overall sample (including former city councilors) are included in Table 3. Balance checks confirming randomization was successful for both current city councilors and the overall sample (including former city councilors) are included in Supplement Tables 2-5.

Responses to Stage 1 Emails. Stage 1 emails were phrased as public service announcements, so they did not invite a response from city councilors. Nevertheless, a very small minority of city council members chose to respond to Stage 1 emails. Of the 3.308 current and former city councilors who received a Stage 1 email, 20 responded. Fifteen of these were responses to the negative feedback about bias in city councils. The responses were a mix of gratitude (e.g., "Thank you for sharing these important facts."), agreement and support (e.g., "I wholeheartedly agree."), requests for more information, and, in four cases, disagreement and

anger (e.g., “This is baloney.”). The five responses to the positive feedback about lack of bias in city councils were all positive and expressed gratitude.

Responses to Stage 2 Emails. Response rates by condition are visualized in Figure 9a. Current city councilors replied to emails from White men and Black men 15.1% and 12.6% of the time, respectively, in the *no feedback control* condition ($z = 1.213$, $p = .225$). Meanwhile, city councilors replied to White and Black men 14.0% and 16.9% of the time, respectively, in the *positive feedback* condition ($z = 1.385$, $p = .166$) and 11.0% and 11.8% of the time, respectively, in the *negative feedback* condition ($z = 0.412$, $p = .680$). These response rates give a rough lower bound estimate of how many Stage 1 emails were opened and read (13.6%); a slightly less conservative lower bound is 23.3%, the proportion of current city councilors who responded within one week of receiving the help-seeking email.

The discriminatory gap in the *no feedback control* condition is a non-significant 2.5 percentage-points (see Figure 9b), diverging from the 5 to 5.5 percentage-point gaps in response rates to Black vs. White men identified in prior audit experiments with local and state politicians (Butler & Broockman, 2011; Butler & Crabtree, 2017; Kirgios et al., 2022).

Primary Analysis. My primary preregistered analysis was an ordinary least squares regression with robust standard errors clustered by city to account for cluster randomization and my binary dependent variable measuring whether each politician responded to the help-seeking email (Arceneaux, 2005). Predictor variables in the regression included indicators for assignment to each bias feedback condition (with the *no feedback control* condition omitted), an indicator for assignment to the *Black help seeker* condition, and the interaction between the feedback condition indicators and the *Black help seeker* condition indicator. The regression also included the following preregistered control variables: an indicator for the city councilor’s gender (male

omitted), indicators for the city councilor's race (Black, Latinx, White, or Other, with White omitted), indicators for the email variant used (four email variants were used in order to stimulus sample), indicators for the city councilor's region, indicators for the city councilor's political party (Democrat, Republican, or Other, with Other omitted), and a continuous control for the city's logged population size.[16]

There was no significant main effect of assignment to the *Black help seeker* condition (beta = -0.024, SE = 0.020, p = .230), assignment to the *positive feedback* condition (beta = -0.007, SE = 0.021, p = .717), or assignment to the *negative feedback* condition (beta = -0.030, SE = 0.020, p = .126). The interaction between assignment to the *Black help seeker* condition and assignment to the *negative feedback* condition was positive but non-significant (beta = 0.033, SE = 0.026, p = .205). Meanwhile, the interaction between assignment to the *Black help seeker* condition and assignment to the *positive feedback* condition was positive and significant (beta = 0.053, SE = 0.027, p = .049).[17] Delivering positive feedback about pro-diversity behavior in city councils led to a 4.6 percentage-point (36.3%) regression-estimated boost in response rates to Black help seekers relative to delivering no feedback. Wald tests reveal no significant difference in the effects of the positive and negative feedback on response rates to Black help seekers (beta = 0.020, SE = 0.025, p = .427). See Table 5, Column 2 for full regression results.

---

[16] The primary analysis specified in my preregistration is for the full sample, including both the 3,981 city councilors actively serving at the time of my experiment and the 885 former city councilors no longer serving at the time of my experiment. However, I am focusing on current city councilors for my primary analysis because feedback about one's professional group is likely to feel more relevant and applicable to people still serving in the profession. For those people, the professional group is more likely to be an ingroup with which they identify, so they are more likely to feel that the feedback applies to them (Terry & Hogg, 1996; Goldstein & Cialdini, 2007). As a result, I have modified my preregistered primary analysis to focus on this subgroup, and have excluded a preregistered control variable: an indicator for whether the politician was still serving on the council. I present my preregistered primary analysis as a robustness check. I also present all exploratory analyses for both the sample of current city councilors alone and for the full sample of city councilors in the Supplement.

[17] Results remain consistent when I use logistic regression rather than OLS regression (see Supplement Table 6) and when I do not include control variables (see Table 5, Column 1).

As a robustness check, I conducted my primary preregistered analysis with the full sample (4,866 city councilors, 885 of whom were no longer serving on city councils at the time of the experiment). The pattern of results for this full sample can be visualized in Supplement Figure 1. Although the results are consistent with the effects for current city councilors, the benefits of the positive feedback are muted in this sample: positive feedback increases response rates to Black help seekers by a regression-estimated 29.4% (beta = 0.037, SE = 0.023, p = .096; see Supplement Table 7 for full regression results and Supplement Table 8 to see these results are robust to using logistic regression rather than OLS regression). That the benefits of positive feedback are dimmed for the full sample is not surprising; feedback about bias in city councils is likely to feel less applicable to former city councilors.

Preregistered Exploratory Analyses: Former vs. Current City Councilors. To explore whether current and former city councilors' behavior towards racial minorities following positive, negative, or no feedback about bias differed, I conducted a preregistered heterogeneity analysis. This analysis also allows me to test the theory that descriptive norms may be driving the benefits of positive feedback. Descriptive norms are more likely to sway behavior when the norm is set within a relevant referent group (Goldstein et al., 2008). For people no longer serving on the city council, city councilors are less likely to feel like a relevant referent group. As a result, I would expect that, to the extent that positive feedback sets a descriptive norm for behavior, it should be less likely to improve response rates to Black men amongst former city councilors relative to current city councilors.

I examined whether the current and former city councilors differed significantly by modifying my primary regression to include the interactions between the indicator for whether the recipient was a former (not current) city councilor and each of the bias feedback condition indicators, the interaction between the indicator for whether the recipient was a former city

councilor and the *Black help seeker* condition indicator, and three-way interactions between the indicator for whether the recipient was a former city councilor, the *Black help seeker* condition indicator, and each of the bias feedback condition indicators.

There was a directional, negative three-way interaction between assignment to the *Black help seeker* condition, assignment to the *positive feedback* condition, and the city councilor's status (beta = -0.083, SE = 0.046, p = .075) such that former city councilors were a regression-estimated 1.9 percentage-points less likely to reply to Black men after receiving positive feedback about bias in city councils. Wald tests suggest that current city councilors, meanwhile, were 4.5 percentage-points more likely to reply to Black men after receiving positive feedback. There was a significant, negative three-way interaction between assignment to the *Black help seeker* condition, assignment to the *negative feedback* condition, and the city councilor's status (beta = -0.105, SE = 0.046, p = .023). Former city councilors were a regression-estimated 2.6 percentage-points less likely to reply to Black men after receiving negative feedback, whereas Wald tests suggest that current city councilors were unaffected by the negative feedback: the regression-estimated effect of receiving negative feedback on current city councilors' responses to Black men was 0.0. Full regression results are included in Table 6.

Preregistered Exploratory Heterogeneity Analyses. The effects of both positive and negative feedback about city councilors' bias on their response rates to Black men were not moderated by the political affiliation of the city councilor, the demographic identity of the city councilor, the political leaning of the city councilor's city, or the size of the city councilor's city (see Supplement pages 26-32).

One city-level participant characteristic did directionally moderate the benefits of positive feedback: the demographic diversity of the city in which the city councilor was serving.

Specifically, the positive feedback seemed to improve response rates to Black men by a directionally greater margin in Whiter cities (beta = 0.002, SE = 0.001, p = .095; see full regression results in Supplement Table 8).

To further explore these results, I analyzed the effects of positive and negative feedback on subsequent discrimination among (1) city councilors serving in the least White cities in my sample (cities whose population was 61.6% White or less, the median value in my sample) and (2) city councilors serving in the Whitest cities in my sample (cities whose population was more than 61.6% White). Cities with fewer White residents than the median included places like Miami, Florida (14.5% White); Jackson, Mississippi (25.7% White); and Inglewood, California (26.7% White). Cities with more White residents than the median included places like Duluth, Minnesota (91.6% White); Muncie, Indiana (87.4% White); and Eugene, Oregon (83.1% White). Positive feedback did not influence response rates to Black men among city councilors serving in the least White cities (beta = -0.013, SE = 0.037, p = .731; see full regression results in Supplement Table 9, column 1). Meanwhile, city councilors serving in the Whitest cities were a regression-estimated 10.8 percentage-points, or 95.8%, more likely to respond to Black men after receiving positive feedback (beta = 0.112, SE = 0.039, p = .004; see full regression results in Supplement Table 9, column 2; results are consistent when considering the full sample including former city councilors, see Supplement pages 28-29). Response rates in cities above and below the median White population are depicted in Figure 10a and Figure 10b and Supplement Figure 2.

These results provide further corroboration that descriptive norms may be driving the benefits of positive feedback. In Whiter cities, there may be more ambiguity about how to handle requests from non-White constituents, so city councilors may be more inclined to use their colleagues' behavior as a cue to inform their own (Chang et al., 2019).

Preregistered Exploratory Subgroup Analyses. Neither positive nor negative feedback significantly influenced female, Black, or Latinx city councilors' willingness to reply to Black male help seekers (see Supplement pages 33-37 for subgroup analyses). However, both male city councilors and White city councilors were significantly more likely to reply to help requests from Black men after receiving positive feedback relative to when they received no feedback (see Supplement Table 10 for full regression results and Supplement Figure 3 to visualize these results; results are consistent when I consider the full sample of both current and former city councilors). After receiving positive feedback, men currently serving on city councils were a regression-estimated 42.6% more likely to respond to Black help seekers (beta = 0.075, SE = 0.035, p = .031) and White city councilors were a regression-estimated 33.6% more likely to respond to Black help seekers (beta = 0.071, SE = 0.035, p = .039). Again, these results offer suggestive evidence that descriptive norms may drive the benefits of positive feedback. Prior work suggests that dominant group members might exert more effort towards ensuring they do not appear to be racist, perhaps because they expect to face more negative scrutiny if they exhibit bias (Apfelbaum et al., 2008). As a result, dominant group members may be more likely to attempt to follow anti-racist social norms in order to avoid being labeled racist.

Preregistered Alternative Reply Timelines. As a robustness check, I planned to examine whether the effects of sharing positive and negative feedback about bias in city councils on response rates to Black men would hold across two longer response windows, considering all replies that arrived within 48 hours or within 1 week of the help-seeking email. Overall, 13.6% of city councilors replied within 24 hours, 17.2% within 48 hours, and 23.3% within one week. The effect of positive feedback was not robust across response windows; instead, it seemed to fade over time, weakening to marginal significance at 48 hours and to a null effect at 1 week. Specifically, the interaction between assignment to the *Black help seeker* condition and

assignment to the *positive feedback* condition was directionally positive when considering a 48-hour response window (beta = 0.049, SE = 0.029, p = .092) and positive but non-significant when considering a one-week response window (beta = 0.030, SE = 0.031, p = .328). The weakening of the positive feedback message over time may suggest that the salience of the feedback is important for determining its effectiveness (Tiefenbeck et al., 2018). See Supplement Tables 11-14 for full regression results.

Exploratory Analyses: Polite Language in Replies. To explore whether the qualitative dimensions of city councilors' responses improved after receiving feedback, I analyzed the influence of experimental condition on the politeness of current city councilors' replies. This analysis was not formally preregistered.

The difference in city councilors' usage of polite language (e.g., "please" and "thank you") in responses to Black vs. White men across conditions is depicted in Figure 11. When city councilors didn't reply, I replaced the missing politeness value with the mean. Specifically, city councilors who did not reply to Black men were counted as having used 2.65% polite terms in their responses, which was the mean level of politeness used in responses to Black men in the *no feedback control* condition. City councilors who did not reply to White men were counted as having used 4.56% polite terms in their responses, which was the mean level of politeness used in responses to White men in the *no feedback control* condition.

On average, city councilors used significantly less polite language when responding to Black men in the *no feedback control* condition: replies from city councilors who received no feedback about bias included 4.56% polite terms when responding to White men and 2.65% polite terms when responding to Black men (t= -18.138, p < .001). City councilors who received positive feedback about bias showed a similar though directionally smaller gap: their responses

included 4.35% polite terms when responding to White men and 2.84% polite terms when responding to Black men (t = -13.618, $p < .001$). City councilors who received negative feedback about bias included 4.52% polite terms when responding to White men and 2.67% polite terms when responding to Black men (t = -21.137, $p < .001$).

Analyzing the politeness of city councilors' replies with my preregistered primary regression confirmed that positive feedback reduced bias. City councilors' responses to Black men contained significantly less polite language than responses to White men (beta = -1.904, SE = 0.107, $p < .001$). This tendency did not change significantly amongst city councilors who received negative feedback about bias (beta = 0.043, SE = 0.136, $p = .751$). However, city councilors who received positive feedback about bias used significantly more polite language in their responses to Black men relative to city councilors who did not receive any feedback (beta = 0.393, SE = 0.162, $p = .016$; regression-estimated boost in polite language = 0.197, 24.0% increase; see Table 7 for full regression results). The benefits of positive feedback are consistent when I consider the full sample of city councilors rather than current city councilors alone (beta = 0.377, SE = 0.140, $p = .008$; regression-estimated boost in polite language = 0.185, 25.8% increase) and when I conduct conditional analyses considering only the sample of city councilors who replied (beta = 2.665, SE = 1.017, $p = .009$; regression-estimated boost in polite language = 1.133, 32.1% increase).

Discussion

Study 1 offers some suggestive evidence that giving acting city councilors positive feedback about a lack of racial bias in their professional referent group can increase both the rate at which they reply to help requests from Black men and the politeness of those replies relative to delivering no feedback. The benefits of positive feedback seem to be primarily driven by male

and White current city councilors and city councilors serving in predominantly White cities, for whom the boost in response rates to Black men after receiving positive feedback was most pronounced. Meanwhile, delivering negative feedback about bias (i.e., informing city councilors that there is evidence that city councilors discriminate against Black constituents and contribute to unequal outcomes for racial minorities) does not seem to affect response rates or response quality relative to delivering no feedback.

The benefit of positive bias messages on response rates to Black men is larger for current than former city councilors, providing suggestive evidence that the positive bias message may operate in part by setting a descriptive group norm or standard of behavior (Goldstein et al., 2008; Duguid & Thomas-Hunt, 2015; Crandall et al., 2002). Those norms are less likely to provide a meaningful benchmark and, ultimately, less likely to lead to behavior change for those who are no longer part of the professional referent group (Goldstein et al., 2008; Crandall et al., 2002; Terry & Hogg, 1996; Prentice & Paluck, 2020). Current group members, meanwhile, are public figures, so they are in a more visible position than former politicians. As a result, they may be more concerned about negative scrutiny, driving them to conform to group norms when deciding how to behave towards racial minorities (Chang et al., 2019).

The directional benefits of the positive feedback message on response rates faded when examining longer response windows. In other words, city councilors who waited longer to reply to students seemed less influenced by the bias feedback. This may be because the feedback had faded from their memory by then, suggesting that salience or recency of the feedback may matter for its effectiveness (Tiefenbeck et. al., 2018; Butler & Crabtree, 2017; Chang et al., 2020; Kirgios et al., 2022).

General Discussion

In a field experiment with 3,981 current U.S. city council members, I found that sharing positive feedback about racial bias in one's professional referent group induced a regression-estimated 36.3% increase in response rates to Black men seeking career advice relative to sharing no feedback about racial bias. Incorporating 885 former city councilors into the sample reduced the size of the effect of positive feedback to a marginal but substantial regression-estimated 29.4% increase in response rates to advice-seeking Black men. In exploratory, non-preregistered analyses, I found that sharing positive feedback about lack of bias also increased the use of polite language in replies to Black men by a regression-estimated 24.0%, reducing the discriminatory politeness gap identified in prior research and replicated in this experiment (Voigt et al., 2017). Negative feedback about bias, on the other hand, did not increase either the likelihood that city councilors reply to Black men or the politeness of those replies.

The benefits of positive feedback seemed to be primarily driven by current city councilors—those still serving as elected officials when the experiment was conducted. Moreover, White and male current council members were a regression-estimated 36.6% and 49.9% more likely to reply to Black men when they received positive feedback, respectively. These subgroups—visible public figures and dominant group members—may be particularly motivated to avoid the appearance of prejudice by conforming to descriptive norms around behavior towards racial minorities (Apfelbaum et al., 2008; Chang et al., 2019; Norton et. al., 2006). City councilors serving in cities with a larger White population than the median were also more responsive to positive feedback, perhaps because there is more ambiguity about how to handle requests from racial minority constituents in those cities. This ambiguity may have increased the extent to which people used descriptive norms to inform their own behavior (Chang et al., 2020). Thus, these subgroup effects provide some evidence that normative pressures are driving the benefits of positive feedback.

In this work, I demonstrate that sharing *positive* information about support for racial minorities in one's professional referent group can reduce racial bias. The idea of leveraging social norms to reduce prejudice is not new, but extant research has tended to focus on attitudes rather than behaviors, sharing information about peers' tolerant beliefs in order to push people towards a more egalitarian consensus (Gomez et al., 2018; Patel, 2013; Paluck et al., 2021; cf. Munger, 2017). However, behavior change and attitude change can be independent, particularly in the domain of prejudice reduction (Paluck et al., 2021; Chang et al., 2019). The results of this work suggest that highlighting positive behavior towards racial minorities can also lead to behavior change, at least in the short term. Moreover, this work suggests that sharing positive feedback about past egalitarian behavior at the group level rather than the individual level may mitigate motivation laundering effects, reducing the backlash that sometimes stems from highlighting past moral actions (Monin & Miller, 2001). Organizations can leverage this insight by publicly highlighting patterns of pro-diversity behavior in groups of employees (e.g., in departments, offices, or the organization as a whole) in order to encourage employees to engage in future behavior that supports racial minorities. When people receive this positive feedback, it seems to change their perceptions of how their peers are behaving. Knowing their colleagues have been exerting extra effort to help racial minorities, they seek to follow suit. It would be valuable for future work to explore whether other forms of positive feedback about bias (e.g., individual-level feedback, feedback about an outstanding ally, etc.) would prove equally or more effective.

Theorizing on the self-regulation of prejudice suggests people need to be confronted with their own biases in order to be motivated to control them (Monteith, 1993; Monteith, Ashburn-Nardo, Voils, & Czopp, 2002). This work posits that people improve their behavior towards racial minorities only after experiencing the guilt that stems from the discrepancy between their desired

(non-prejudiced) reactions and their actual (prejudiced) reactions (Monteith, 1993; Monteith et al., 2002; Monteith & Mark, 2005). In other words, prejudice reduction is discrepancy-motivated. Here, however, I show that the opposite can also be true: people confronted with evidence that members of their professional ingroup have behaved *consistently* with their egalitarian values respond more positively to racial minorities in the future, suggesting that when feedback is provided at the group level, correspondence between desired and actual behavior can be just as (if not more) motivating as divergence.

This work identifies an effective strategy for reducing the effects of racial bias on people's willingness to provide career advice. In doing so, I focus on reducing bias in a pathway process (i.e., an informal process that affects one's downstream success). Most work on bias reduction examines gateway processes (e.g., hiring or promotions), but pathways are more frequent and precede gateways, so it is critically important to understand how to reduce inequalities at pathways (see Milkman, Akinola, & Chugh, 2015). It's worth exploring whether positive feedback can reduce bias in other pathway processes (e.g., provision of mentorship, help on tasks, or referrals).

These findings also add to recent work suggesting that people learn more from positive feedback than negative feedback even though negative stimuli are more attention-grabbing (Eskreis-Winkler & Fishbach, 2019; Eskreis-Winkler & Fishbach, 2020). Because the negative feedback is more actionable and informative than the positive feedback, this work may be a particularly conservative test of that hypothesis. In the negative feedback condition, politicians learned that city councilors were less likely to respond to Black constituents than White constituents. People motivated to avoid prejudice can infer that they should change their behavior and respond more often to emails from Black constituents. Meanwhile, in the positive feedback condition, politicians learned that city councilors responded to Black constituents more when they

emphasized their minority status. If city councilors want to further support Black constituents, the positive feedback doesn't give a direct suggestion for how to do so. Despite this asymmetry—which should favor the negative message—positive feedback is the only feedback that effectively changed behavior in my audit experiment.

Although the negligible impact of the negative feedback is consistent with prior work by Butler and Crabtree (2017), it does conflict with important findings by Pope, Price, and Wolfers (2018) which suggest that raising awareness of racial bias in one's professional context can reduce or eliminate biased behavior. Specifically, Pope and colleagues (2018) found that after a study identifying racial bias in NBA referees' game-time calls gained widespread media attention, the referees' bias disappeared—thus, hearing negative feedback about bias in their ingroup eliminated their biased behavior. These divergent results may be explained in several ways. On the one hand, it's possible that the mechanism underlying the benefits of bias awareness documented by Pope and colleagues (2018) was structural change rather than individual change. While the authors were unable to identify a policy change in the NBA in response to the media attention the original study garnered, they acknowledge that some policy changes may have occurred without being advertised. In that case, the negative feedback may have operated by spurring organizational leaders to change the decision-making environment rather than by changing individual behavior. On the other hand, it's possible that the awareness interventions in this work and in the Butler and Crabtree (2017) work were simply too subtle. Because I did not track email open rates in my field experiment to avoid being flagged as spam, I cannot be sure how many people opened the negative feedback emails I sent in Stage 1. Given that the subject line of my Stage 1 email alluded to the (potentially ego threatening) contents, people may have simply deleted the message without reading it. Even if politicians read the message, a single, brief email is different than prolonged, widespread media attention. The former is easier to dismiss

without internalizing and learning from the information than the latter (Sherman & Cohen, 2006). Moreover, the feedback came from an unknown research lab rather than a known, trusted source (e.g., a boss or a major media outlet). Councilors may have been more likely to dismiss the negative feedback than they would have been had it come from a trusted source (Audia & Locke, 2003). It would be valuable for future work to explore the impact of sustained attempts to impart negative feedback about bias and to examine whether negative feedback more effectively reduces prejudice when it comes from a trusted source.

While I find suggestive evidence that descriptive norms drive the benefits of positive feedback, it would be valuable to further explore the psychological processes underlying the effects of positive feedback. Future work should, for example, probe whether people's perceptions of descriptive group norms actually shift after they receive positive feedback about pro-diversity behaviors in their ingroup. Alternative mechanisms are also worthy of future examination. For example, positive feedback may boost people's feelings of self-efficacy around egalitarian behaviors, creating a self-fulfilling prophecy (Eden, 1992). Specifically, positive feedback may increase people's sense that they (and their colleagues) know how to enact their pro-diversity intentions and, as a result, their willingness to do so in the future. Self-efficacy boosts may explain why people who may typically feel less confident about their pro-diversity habits (e.g., majority group members, councilors serving in Whiter cities) showed a greater reduction in bias after receiving positive feedback. This and other alternative mechanisms (e.g., positive affect, monitoring) should be tested in future work.

More generally, it would be valuable for future work to replicate and extend these findings in other contexts, for different identity groups, and with different types of feedback. For example, future work could examine more individualized feedback, repeated vs. one-time feedback, and feedback that contains multiple actionable tips for changing behavior. It would also

be worthwhile for future work to examine whether this strategy works for other marginalized groups (e.g., when sharing feedback about discrimination against Asian people, women, sexual minorities, low income students, etc.). It would be particularly useful to examine the long-term impact of feedback provision rather than only measuring next-day behavior. This work provides some suggestive evidence that timeliness matters: the benefits of sharing positive feedback waned if people waited longer to reply to the student after receiving the feedback. However, I did not explicitly manipulate the timing of the message, so all participants received the two emails within one day of each other. Future work should explicitly vary the time lag between feedback provision and behavior measurement in order to explore the role of salience and forgetfulness in determining the impact of feedback.

In an effort to reduce prejudiced behavior, our instinct is often to highlight bad behavior: we teach people about the prevalence of stereotyping in diversity training, share viral news stories about instances of discrimination, and readily call people out on Twitter when their most recent Tweet leaks evidence of their biases. These may all be functional actions, establishing the bounds of acceptable behavior and evidencing the size and scope of the impact of bias. However, they may not be the best behavior change strategies to encourage people to support racial minorities. Instead, it may be more effective to amplify group-level patterns of good behavior, shining a light on a new standard for egalitarian behavior that people may be compelled to follow.

References

Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. Journal of personality and social psychology, 95(4), 918.

Arceneaux, K. (2005). Using cluster randomized field experiments to study voting behavior. The Annals of the American Academy of Political and Social Science, 601(1), 169-179.

Audia, P. G., & Locke, E. A. (2003). Benefiting from negative feedback. Human Resource Management Review, 13(4), 631-646.

Ayres, I. (1991). Fair driving: Gender and race discrimination in retail car negotiations. Harvard Law Review, 817-872.

Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. Handbook of economic field experiments, 1, 309-393.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American economic review, 94(4), 991-1013.

Bezrukova, K., Jehn, K. A., & Spell, C. S. (2012). Reviewing diversity training: Where we have been and where we should go. Academy of Management Learning & Education, 11(2), 207-227.

Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016). A meta-analytical integration of over 40 years of research on diversity training evaluation. Psychological bulletin, 142(11), 1227.

Blanchard, F. A., Lilly, T., & Vaughn, L. A. (1991). Reducing the expression of racial prejudice. Psychological Science, 2(2), 101-105.

Butler, D. M., & Broockman, D. E. (2011). Do politicians racially discriminate against constituents? A field experiment on state legislators. American Journal of Political Science, 55(3), 463-477.

Butler, D. M., & Crabtree, C. (2017). Moving beyond measurement: Adapting audit studies to test bias-reducing interventions. Journal of Experimental Political Science, 4(1), 57-67.

Chaney, K. E., & Sanchez, D. T. (2018). The endurance of interpersonal confrontations as a prejudice reduction strategy. Personality and Social Psychology Bulletin, 44(3), 418-429.

Chaney, K. E., Sanchez, D. T., Alt, N. P., & Shih, M. J. (2021). The breadth of confrontations as a prejudice reduction strategy. Social Psychological and Personality Science, 12(3), 314-322.

Chang, E. H., Kirgios, E. L., Rai, A., & Milkman, K. L. (2020). The isolated choice effect and its implications for gender diversity in organizations. Management Science, 66(6), 2752-2761.

Chang, E. H., Milkman, K. L., Chugh, D., & Akinola, M. (2019). Diversity thresholds: How social norms, visibility, and scrutiny relate to group composition. Academy of Management Journal, 62(1), 144-171.

Chang, E. H., Milkman, K. L., Gromet, D. M., Rebele, R. W., Massey, C., Duckworth, A. L., & Grant, A. M. (2019). The mixed effects of online diversity training. Proceedings of the National Academy of Sciences, 116(16), 7778-7783.

Crandall, C. S., Eshleman, A., & O'brien, L. (2002). Social norms and the expression and suppression of prejudice: the struggle for internalization. Journal of personality and social psychology, 82(3), 359.

Crandall, C. S., & Stangor, C. (2005). Conformity and prejudice. On the nature of prejudice: Fifty years after Allport, 295-309.

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. Journal of personality and social psychology, 90(5), 784.

Dobbin, F., & Kalev, A. (2016). Why diversity programs fail. Harvard Business Review, 94(7), 14.

Donohue III, J. J., & Levitt, S. D. (2001). The impact of race on policing and arrests. The Journal of Law and Economics, 44(2), 367-394.

Duguid, M. M., & Thomas-Hunt, M. C. (2015). Condoning stereotyping? How awareness of stereotyping prevalence impacts expression of stereotypes. Journal of Applied Psychology, 100(2), 343.

Eby, L. T., Allen, T. D., Evans, S. C., Ng, T., & DuBois, D. L. (2008). Does mentoring matter? A multidisciplinary meta-analysis comparing mentored and non-mentored individuals. Journal of vocational behavior, 72(2), 254-267.

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. American economic journal: applied economics, 9(2), 1-22.

Eden, D. (1992). Leadership and expectations: Pygmalion effects and other self-fulfilling prophecies in organizations. *The Leadership Quarterly*, *3*(4), 271-305.

Effron, D. A., Cameron, J. S., & Monin, B. (2009). Endorsing Obama licenses favoring whites. Journal of experimental social psychology, 45(3), 590-593.

Ehrlinger, J., Gilovich, T., & Ross, L. (2005). Peering into the bias blind spot: People's assessments of bias in themselves and others. Personality and Social Psychology Bulletin, 31(5), 680-692.

Epley, N., & Gilovich, T. (2016). The mechanics of motivated reasoning. Journal of Economic perspectives, 30(3), 133-40.

Eskreis-Winkler, L., & Fishbach, A. (2019). Not learning from failure—The greatest failure of all. Psychological science, 30(12), 1733-1744.

Eskreis-Winkler, L., & Fishbach, A. (2020). When praise—versus criticism—motivates goal pursuit. Psychological perspectives on praise, 47-54.

Festinger, L. (1957). A theory of cognitive dissonance (Vol. 2). Stanford university press.

Fishbach, A., & Dhar, R. (2005). Goals as excuses or guides: The liberating effect of perceived goal progress on choice. Journal of Consumer Research, 32(3), 370-377.

Glover, D., Pallais, A., & Pariente, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. The Quarterly Journal of Economics, 132(3), 1219-1260.

Goldstein, N. J., & Cialdini, R. B. (2007). The spyglass self: a model of vicarious self-perception. Journal of personality and social psychology, 92(3), 402.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. Journal of consumer Research, 35(3), 472-482.

Gómez, Á., Tropp, L. R., Vázquez, A., Voci, A., & Hewstone, M. (2018). Depersonalized extended contact and injunctive norms about cross-group friendship impact intergroup orientations. Journal of Experimental Social Psychology, 76, 356-370.

Hains, R. (2019, August 23). Perspective | dear fellow white people: Here's what to do when you're called racist. The Washington Post. Retrieved March 24, 2022, from https://www.washingtonpost.com/outlook/dear-fellow-white-people-heres-what-to-do-when-youre-called-racist/2019/08/20/6e31941a-beda-11e9-b873-63ace636af08_story.html

Holpuch, A. (2022, March 9). Twitter bot highlights gender pay gap one company at a Time. The New York Times. Retrieved March 18, 2022, from https://www.nytimes.com/2022/03/09/business/pay-gap-international-womens-day-twitter.html

Howell, J. L., Collisson, B., Crysel, L., Garrido, C. O., Newell, S. M., Cottrell, C. A., ... & Shepperd, J. A. (2013). Managing the threat of impending implicit attitude feedback. Social Psychological and Personality Science, 4(6), 714-720.

Howell, J. L., Redford, L., Pogge, G., & Ratliff, K. A. (2017). Defensive responding to IAT feedback. Social Cognition, 35(5), 520-562.

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. American sociological review, 71(4), 589-617.

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J. F. (2009). Mispredicting affective and behavioral responses to racism. Science, 323(5911), 276-278.

Kirgios, E. L., Rai, A., Chang, E. H., & Milkman, K. L. (2022). When seeking help, women and racial/ethnic minorities benefit from explicitly stating their identity. Nature Human Behaviour, 1-9.

Kouchaki, M. (2011). Vicarious moral licensing: the influence of others' past moral actions on moral behavior. Journal of personality and social psychology, 101(4), 702.

Levy, A., & Maaravi, Y. (2018). The boomerang effect of psychological interventions. Social Influence, 13(1), 39-51.

Locke, E. A., & Latham, G. P. (1990). A theory of goal setting & task performance. Prentice-Hall, Inc.

Milkman, K. L., Akinola, M., & Chugh, D. (2015). What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. Journal of Applied Psychology, 100(6), 1678.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of personality and social psychology*, *81*(1), 33.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. Journal of personality and social psychology, 65(3), 469.

Monteith, M. J., Ashburn-Nardo, L., Voils, C. I., & Czopp, A. M. (2002). Putting the brakes on prejudice: on the development and operation of cues for control. Journal of personality

and social psychology, 83(5), 1029.

Monteith, M. J., & Mark, A. Y. (2005). Changing one's prejudiced ways: Awareness, affect, and self-regulation. European review of social psychology, 16(1), 113-154.

Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. Annual review of psychology, 67, 363-385.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. Political Behavior, 39(3), 629-649.

Nguyen, C. M. (2021). The effect of other in-group members' organizational citizenship behavior on employees' organizational deviance: a moral licensing perspective. Journal of Asian Business and Economic Studies.

Norton, M. I., Sommers, S. R., Apfelbaum, E. P., Pura, N., & Ariely, D. (2006). Color blindness and interracial interaction: Playing the political correctness game. Psychological Science, 17(11), 949-953.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. Annual review of psychology, 60, 339-367.

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. Annual review of psychology, 72, 533-560.

Parker, L. R., Monteith, M. J., Moss-Racusin, C. A., & Van Camp, A. R. (2018). Promoting

concern about gender bias with evidence-based confrontation. Journal of Experimental Social Psychology, 74, 8-23.

Patel, S. L. (2013). Examining the influence of perceived social consensus information on weight prejudice across development. The University of Texas at Dallas.

Pennebaker, J. W., Boyd, R. L., Booth, R. J., Ashokkumar, A., & Francis, M. E. (2022). Linguistic Inquiry and Word Count: LIWC-22.

Pettigrew, T. F. (1991). Normative theory in intergroup relations: Explaining bo th harmony an d conflict. Psychology and developing societies, 3(1), 3-16.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. Journal of personality and social psychology, 75(3), 811.

Plant, E. A., & Devine, P. G. (2009). The active control of prejudice: unpacking the intentions guiding control efforts. Journal of personality and social psychology, 96(3), 640.

Pope, D. G., Price, J., & Wolfers, J. (2018). Awareness reduces racial bias. Management Science, 64(11), 4988-4995.

Prentice, D., & Paluck, E. L. (2020). Engineering social change using social norms: Lessons from the study of collective action. Current Opinion in Psychology, 35, 138-142.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. Personality and Social Psychology Bulletin, 28(3), 369-381.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. Psychological science, 18(5), 429-434.

Schumann, K., & Dweck, C. S. (2014). Who accepts responsibility for their transgressions?. Personality and social psychology bulletin, 40(12), 1598-1610.

Seibert, S. E., Kraimer, M. L., & Liden, R. C. (2001). A social capital theory of career success. Academy of management journal, 44(2), 219-237.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. Advances in experimental social psychology, 38, 183-242.

Solano, E., & Robson, A. (2020). How viral videos of racist incidents are changing society. Time. Retrieved March 18, 2022, from https://time.com/5875479/viral-videos-racism-impact-protests/

Son Hing, L. S., Bobocel, D. R., & Zanna, M. P. (2002). Meritocracy and opposition to affirmative action: making concessions in the face of discrimination. Journal of personality and social psychology, 83(3), 493.

Stephan, W. G., & Stephan, C. W. (1985). Intergroup anxiety. Journal of social issues.

Terry, D. J., & Hogg, M. A. (1996). Group norms and the attitude-behavior relationship: A role for group identification. Personality and social psychology bulletin, 22(8), 776-793.

Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., & Staake, T. (2018). Overcoming salience bias: How real-time feedback fosters resource conservation. Management science, 64(3), 1458-1476.

Vitriol, J. A., & Moskowitz, G. B. (2021). Reducing defensive responding to implicit bias feedback: On the role of perceived moral threat and efficacy to change. *Journal of Experimental Social Psychology*, *96*, 104165.

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., ... & Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. Proceedings of the National Academy of Sciences, 114(25), 6521-6526.

Wolfers, J. (2015, March 2). Fewer women run big companies than men named John. The New York Times. Retrieved March 18, 2022, from https://www.nytimes.com/2015/03/03/upshot/fewer-women-run-big-companies-than-men-named-john.html

Tables

Table 4, Chapter 3. Summary Statistics of Participant Characteristics.

	Current City Councilors (N = 3,981)	All City Councilors, Current and Former (N = 4,866)
City Councilor Gender		
Female	1,457 (36.60%)	1,740 (35.76%)
Male	2,524 (63.40%)	3,126 (64.24%)
City Councilor Race		
Black	715 (17.96%)	881 (18.10%)
Latinx	420 (10.55%)	502 (10.32%)
White	2,812 (70.64%)	3,449 (70.88%)
Other	34 (0.85%)	34 (0.70%)
City Councilor Political Party		
Democrat	1,043 (26.20%)	1,306 (26.84%)
Republican	380 (9.54%)	470 (9.66%)
Other	2,558 (64.26%)	3,090 (63.50%)
City Councilor Region		
Northeast	618 (15.52%)	807 (16.58%)
Midwest	1,007 (25.30%)	1,207 (24.80%)
South	1,293 (32.48%)	1,540 (31.65%)
West	1,063 (26.70%)	1,312 (26.96%)
City Councilor Currently Serving		
Former City Councilor	0 (0%)	885 (18.19%)
Current City Councilor	3,981 (100%)	3,981 (81.81%)
Average City Population Size	323,660.1 (S.D. = 1,008,026)	345,715 (S.D. = 1,118,328)
Percent of Voters Who Voted Republican in 2016 Presidential Election in City Councilor's County	40.68% (S.D. = 14.29)	40.40% (S.D. = 14.20)
Percent White Population in City Councilor's County	58.92% (S.D. = 19.45)	59.10% (S.D. = 19.35)

Table 5, Chapter 3. Regression-Estimated Effects of Bias Feedback and Help-Seeker Race on Current City Councilors' Response Rates.

	Did the Participant Reply? (1=Yes, 0=No)					
	Model 1			**Model 2**		
	b	95% CI	p	b	95% CI	p
Black Help Seeker	-0.025	[-0.064, 0.014]	.212	-0.024	[-0.064, 0.015]	.230
Negative Feedback	-0.041	[-0.080, -0.001]	.043	-0.030	[-0.068, 0.008]	.126
Positive Feedback	-0.010	[-0.051, 0.030]	.621	-0.007	[-0.048, 0.033]	.717
Black Help Seeker*Negative Feedback	0.034	[-0.018, 0.085]	.197	0.033	[-0.018, 0.085]	.205
Black Help Seeker*Positive Feedback	0.054*	[0.001, 0.106]	.047	0.053*	[0.0001, 0.106]	.049
Population Size in Recipient's County				-0.019***	[-0.029, -0.009]	<.001
Recipient is a Woman				-0.010	[-0.033, 0.012]	.367
Recipient is a Democrat				-0.012	[-0.042, 0.018]	.437
Recipient is a Republican				-0.038*	[-0.073, -0.002]	.036
Fixed Effects for Recipient Race	No			Yes		
Fixed Effects for Email Variant	No			Yes		
Fixed Effects for Region	No			Yes		
Observations	3981			3981		
Adjusted R^2	0.002			0.018		

Note. This table reports the results of two ordinary least squares (OLS) regression models predicting whether a given city councilor responded to an email from a student requesting career advice. Only the 3,981 current city councilors are included in these models—the 885 former city councilors who had been replaced before the experiment was conducted were not included. Model 1 shows the main effect of assignment to the *Black help seeker* condition, assignment to the *negative feedback* condition, and assignment to the *positive feedback* condition, as well as the interaction between the *Black help seeker* condition indicator and each of the two *feedback* condition indicators. Model 2 includes the same predictors with the addition of the following preregistered covariates: the log-transformed population size of the city councilor's city, a binary indicator for whether the city councilor is a woman, a binary indicators for whether the city councilor is a Democrat or a Republican, fixed effects for the city councilor's race (Black, Latinx, White, or Other), fixed effects for the email variant a city councilor received (I stimulus sampled by including four similar emails, all requesting career advice), and fixed effects for the city councilor's region (as determined by the U.S. Census). Robust standard errors clustered by the city councilor's city are reported in parentheses.

†, *, **, and *** denote significance at the 10%, 5%, 1%, and 0.1% levels, respectively.

Table 6, Chapter 3. Regression-Estimated Effects of Interaction Between City Councilor Position, Bias Feedback, and Help-Seeker Race on Current City Councilors' Response Rates.

	Did the Participant Reply? (1=Yes, 0=No)		
	Model 1		
	b	95% CI	p
Black Help Seeker	-0.024	[-0.064, 0.015]	.230
Negative Feedback	-0.032	[-0.070, 0.006]	.102
Positive Feedback	-0.008	[-0.048, 0.032]	.700
Recipient is a Former City Councilor	-0.133***	[-0.173, -0.092]	<.001
Black Help Seeker * Recipient is a Former City Councilor	0.062†	[-0.009, 0.133]	.087
Black Help Seeker * Negative Feedback	0.033	[-0.018, 0.085]	.206
Black Help Seeker * Positive Feedback	0.053*	[0.0001, 0.106]	0.049
Recipient is a Former City Councilor * Negative Feedback	0.077*	[0.017, 0.138]	.012
Recipient is a Former City Councilor * Positive Feedback	0.019	[-0.041, 0.079]	.528
Recipient is a Former City Councilor * Black Help Seeker * Negative Feedback	-0.105*	[-0.194, -0.015]	.023
Recipient is a Former City Councilor * Black Help Seeker * Positive Feedback	-0.083†	[-0.174, 0.008]	.075
Population Size in Recipient's County	-0.015***	[-0.023, -0.007]	<.001
Recipient is a Woman	-0.011	[-0.030, 0.008]	.243
Recipient is a Democrat	-0.012	[-0.037, 0.014]	.365
Recipient is a Republican	-0.032*	[-0.062, -0.002]	.038
Fixed Effects for Recipient Race		Yes	
Fixed Effects for Email Variant		Yes	
Fixed Effects for Region		Yes	
Observations		4866	
Adjusted R^2		0.028	

Note. This table reports the results of a preregistered ordinary least squares (OLS) regression model predicting whether a given city councilor responded to an email from a student requesting career advice in Study 1. The model shows the main effect of assignment to the *Black help seeker* condition, assignment to the *negative feedback* condition, assignment to the *positive feedback* condition, an indicator for whether the recipient is a former rather than current city councilor (because they stepped down or were replaced shortly before the experiment was conducted), the interaction between the *Black help seeker* condition indicator and each of the two *feedback* condition indicators, the interaction between the *Black help seeker* condition indicator and whether the recipient is a former city councilor, the interactions between each of the two *feedback* conditions and whether the recipient is a former city councilor, and the interactions between assignment to the *Black help seeker* condition, whether the recipient is a former city councilor, and assignment to each of the two *feedback* conditions. Moreover, the model includes the following preregistered covariates: the log-transformed population size of the city councilor's city, a binary indicator for whether the city councilor is a woman, a binary indicator for whether the city councilor is a Democrat, a binary indicator for whether the city councilor is a Republican, fixed effects for the city councilor's race (Black, Latinx, White, or Other), fixed effects for the email variant a city councilor received (I stimulus sampled by including four similar emails, all requesting career advice), and fixed effects for the city councilor's region (as determined by the U.S. Census; categories include Northeast, Midwest, South, and West). Robust standard errors clustered by the city councilor's city are reported in parentheses.

†, *, **, and *** denote significance at the 10%, 5%, 1%, and 0.1% levels, respectively.

Table 7, Chapter 3. Regression-Estimated Effects of Bias Feedback and Help-Seeker Race on Politeness of City Councilors' Replies.

	Model 1			**Model 2**		
	Outcome: Politeness			*Outcome: Politeness*		
	b	**95% CI**	**p**	**b**	**95% CI**	**p**
Black Help Seeker	-1.904**	[-0.604, 0.103]	<.001	-1.904**	[-2.113, -1.694]	<.001
Negative Feedback	-0.037	[-0.501, 0.055]	.747	-0.028	[-0.253, 0.196]	.804
Positive Feedback	-0.205†	[-0.507, 0.003]	.054	-0.196†	[-0.407, 0.015]	.069
Black Help Seeker*Negative Feedback	0.054	[-0.101, 0.543]	.694	0.043	[-0.224, 0.310]	.751
Black Help Seeker*Positive Feedback	0.388*	[0.183, 0.917]	.017	0.393*	[0.075, 0.711]	.016
Preregistered Control Variables Included	No			Yes		
Fixed Effects for Recipient Race	No			Yes		
Fixed Effects for Email Variant	No			Yes		
Fixed Effects for Region	No			Yes		
Observations	3981			3981		
Adjusted R^2	0.184			0.187		

Note. This table reports the results of two exploratory preregistered ordinary least squares (OLS) regression models predicting the use of polite language in city councilors' responses to emails from students requesting career advice. The dependent variable is expressed as the proportion of language in the city councilor's reply that connotes politeness based on the LIWC dictionary (Pennebaker et al., 2022). Politeness values for city councilors who do not respond are replaced with the mean level of politeness in the *no feedback control* condition, where the means are calculated separately for Black and White help-seekers. Models 1 and 2 both show the main effect of assignment to the *Black help seeker* condition, assignment to the *negative feedback* condition, and assignment to the *positive feedback* condition, as well as the interaction between the *Black help seeker* condition indicator and each of the two *feedback* condition indicators. Models 2 includes the following covariates preregistered in my primary analysis: the log-

transformed population size of the city councilor's city, a binary indicator for whether the city councilor is a woman, a binary indicator for whether the city councilor is a Democrat, a binary indicator for whether the city councilor is a Republican, fixed effects for the city councilor's race (Black, Latinx, White, or Other), fixed effects for the email variant a city councilor received (I stimulus sampled by including four similar emails, all requesting career advice), and fixed effects for the city councilor's region (as determined by the U.S. Census; categories include Northeast, Midwest, South, and West). Robust standard errors clustered by the city councilor's city are reported in parentheses.

†, *, **, and *** denote significance at the 10%, 5%, 1%, and 0.1% levels, respectively.

Figures

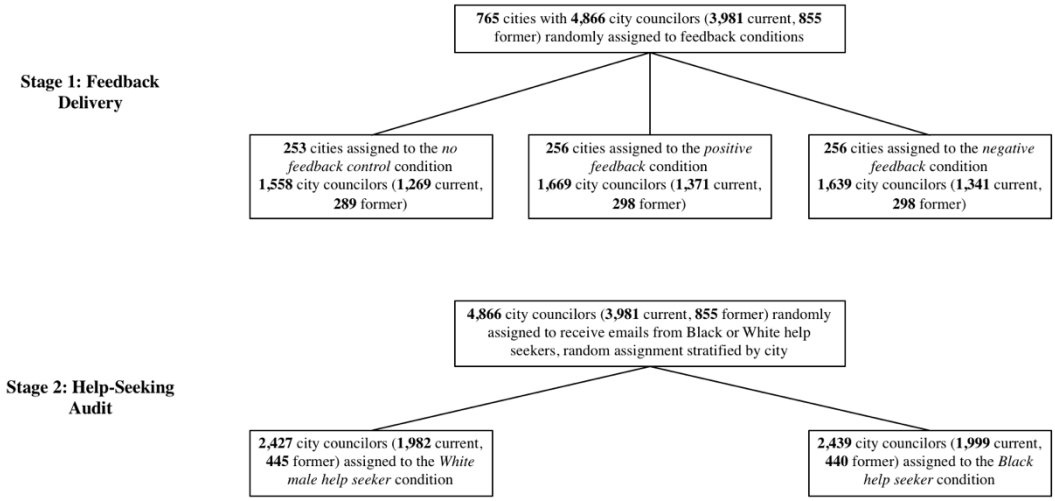


Figure 7, Chapter 3 | Randomization flow chart for the audit experiment. City councilors were assigned to a bias feedback condition for Stage 1 through clustered random assignment by city, so city councilors in the same city received the same message. Assignment to the help seeker conditions for Stage 2 was conducted at the individual level but was stratified by city. Not depicted in the flow chart: I originally emailed 5,537 city councilors, but 671 emails bounced. Those 671 individuals were excluded from analysis, as preregistered, because they could not reply to an email they never received.

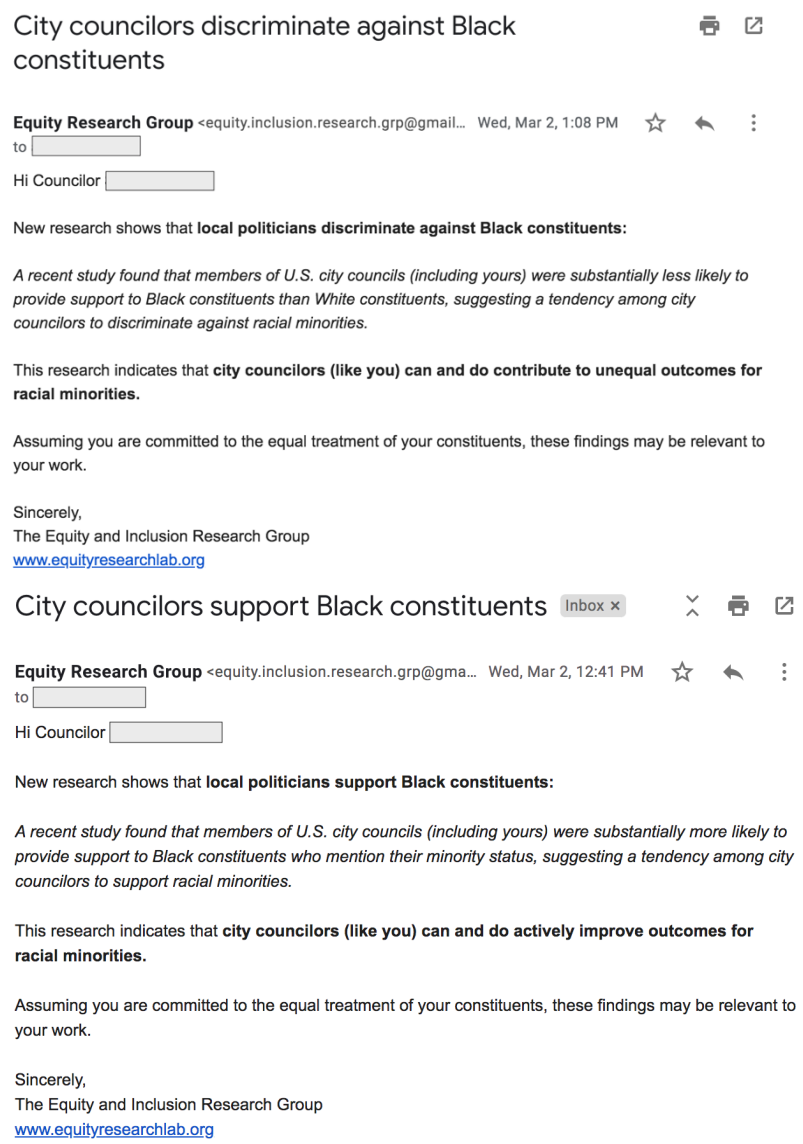


Figure 8, Chapter 3 | Emails from the feedback delivery stage of the field experiment with politicians. Emails are de-identified to maintain anonymity. The left panel displays the email sent in the negative feedback condition and the right panel displays the email sent in the positive feedback condition. No emails were sent to city councilors assigned to the no feedback control condition.

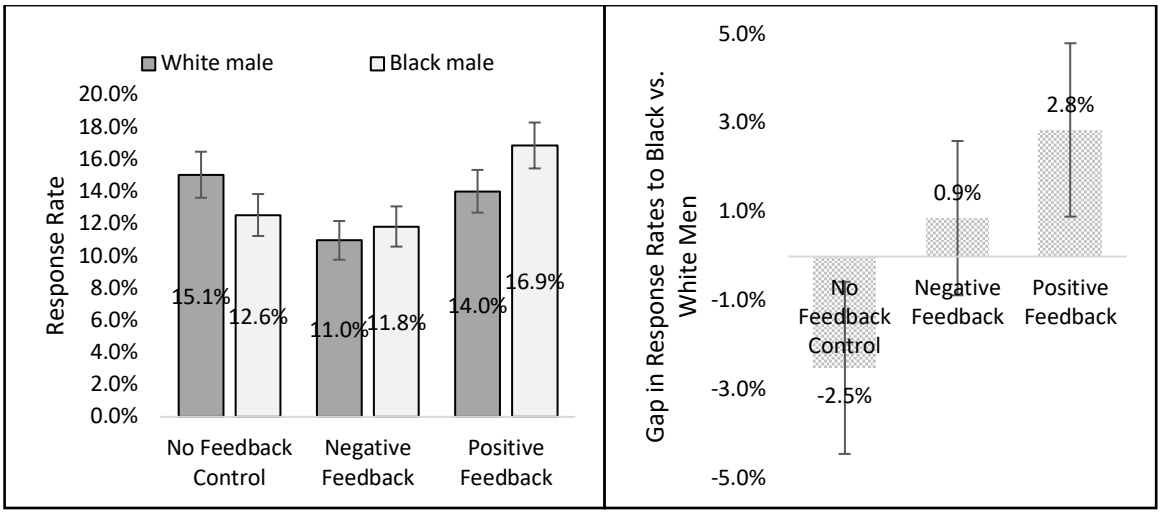


Figure 9a and 9b, Chapter 3 | Current city councilors' response rates to help-seeking emails across conditions. The left panel displays current city councilors' (N = 3,981) response rates to emails from fictitious students seeking career advice. The dark grey bars represent response rates to students whose names signaled that they were White men and the light grey bars represent response rates to students whose names signaled that they were Black men. The two bars on the left display response rates from city councilors who did not receive any feedback about city councilors' behavior towards racial minorities. The two bars in the middle represent response rates from city councilors who received negative feedback suggesting that city councilors discriminate against Black constituents. The two bars on the right represent response rates from city councilors who received positive feedback suggesting that city councilors support Black constituents. The right panel displays the gap in current city councilors' (N = 3,981) response rates to Black men vs. White men in each of the three feedback conditions. Negative values indicate that Black men received fewer responses than White men while positive values indicate that Black men received more responses than White men. In both panels, standard error bars are depicted around each proportion.

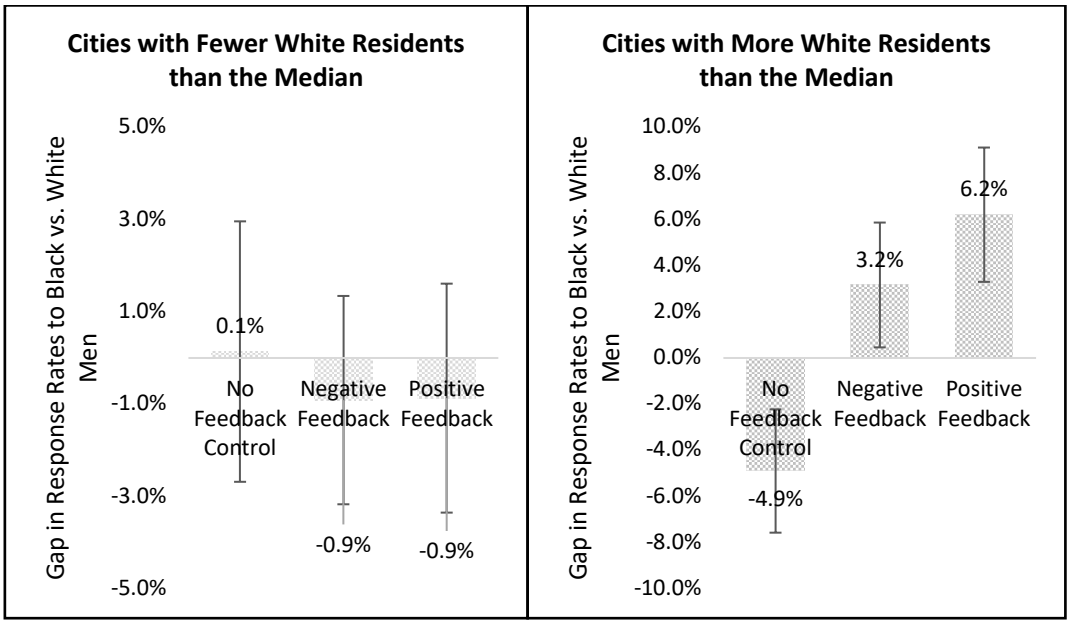


Figure 10a and 10b, Chapter 3 | Gap in response rates to Black vs. White men across conditions in cities with fewer and more White residents than the median. The left panel of the figure displays the gap in city councilors' response rates to Black vs. White help-seeking students for councilors serving in cities with fewer White residents than the median (i.e., 61.6% or less; N = 2,002). The right panel of the figure displays the gap in city councilors' response rates to Black vs. White help-seeking students for councilors serving in cities with more White residents than the median (i.e., more than 61.6%; N = 1,979). Standard error bars are depicted.

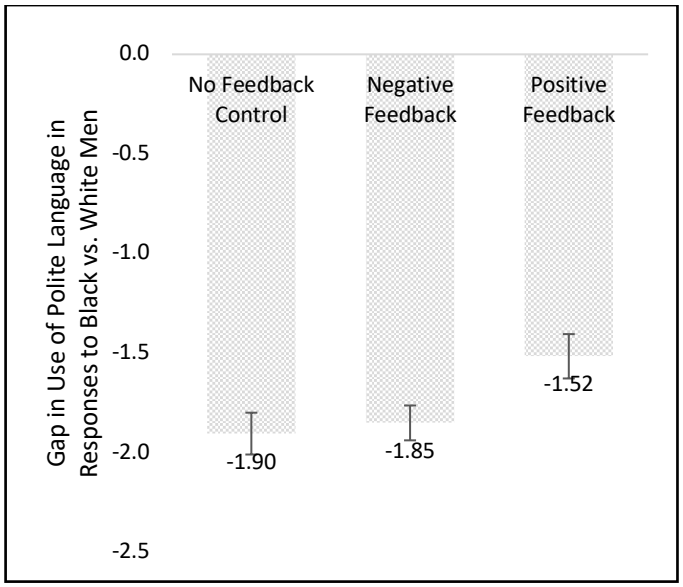


Figure 11, Chapter 3 | Use of polite language in city councilors' replies to help-seeking students across conditions. The figure displays differences in the use of polite language to Black vs. White help-seekers in current city councilors' (N = 3,981) responses to emails from fictitious students. Standard error bars are depicted.