Gender and race effects on job candidate evaluation over time

By

Emily Watkins
B.A., Rhodes College, 2018

Submitted to the graduate degree program in Social Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Arts.

_____

Chair: Monica Biernat, Ph.D.

_____

Nyla Branscombe, Ph.D.

_____

Ludwin Molina, Ph.D.

Date Defended: 26 July 2022

The thesis committee for Emily Watkins certifies that this is the approved version of the following thesis:

Gender and race effects on job candidate evaluation over time

Monica Biernat, Ph.D.

Date Approved: 29 July 2022

Abstract

Most experimental studies of the hiring process present all information about an applicant at a singular time point at which participants make a hiring decision. However, hiring contexts in the real world often present multiple time points for reviewers to form and change their impressions of a candidate. The current studies (total $N = 661$) investigate how changing performance information over time is differentially interpreted based on gender (Study 1) and race (Study 2). Participants were randomly assigned to a male or female (Study 1) or Black or White female (Study 2) applicant, whose job interview performance varied (Very Good, Average, or Poor). Trait ratings of the applicant were made at three time points, each after reviewing a resume, interview performance (where gender/race was revealed and performance was manipulated), and a writing sample. Participants also made a hiring decision and provided other objective ratings. Across both studies, there was little evidence of gender or race bias in hiring decisions and objective ratings among the full sample. Among a sample of participants who identified as White, the average-performing male applicant was judged more favorably than the average-performing female applicant, but a comparable pattern did not emerge in Study 2. No attitudinal moderators emerged in Study 1, but racism levels moderated nearly all judgments in Study 2. Potential explanations for the results, limitations of the study, and limitations of the typical social psychological paradigm used to study discrimination are discussed.

Keywords: gender bias, race bias, hiring decisions, workplace discrimination, updating impressions

Acknowledgments

I first want to thank my family—Julie, Bryan, and Sophie Watkins—for their support, not only through the process of writing this thesis, but throughout my whole life. You've listened to and encouraged me when I've been excited, when I've been sad, and everything in between. I honestly couldn't do any of this without you. I'm just trying to be another Dr. Watkins like you!

Second, I want to thank my advisor, Dr. Monica Biernat. Obviously this would not have happened without you – for so many reasons! I'm thankful to have an advisor who is so caring, experienced, and generous with her time. Thank you for teaching me so much, and I look forward to the next few years working closely with you. Also, thank you to my other committee members, Drs. Nyla Branscombe and Ludwin Molina, for their feedback and guidance on this project.

Third, I want to thank all of my friends. When I moved to Lawrence in the middle of the pandemic, I left all of my best college friends in Washington, DC. Despite the distance, they have supported, encouraged, and loved me well. Thank you, Laurie Williams, Elizabeth Simpson, Maggie Bradley, and Anna Conroy for always being there for me! And to Caroline Fortson, thanks for texting me daily, listening to my Covid paranoia, and always being one of the first people I text about anything!

Moving to Lawrence in the middle of the pandemic was probably the hardest thing I've ever done, but I'm thankful to my cohort—Jordan and Omar—for helping to make it an exciting transition, and for Charlotte Moser and Navanté Peacock for taking me under their wings and providing great mentorship and friendship. I have to give a special shout out to Charlotte for not only answering all my questions and listening to me talk non-stop about (fun) books, but also for

introducing me to the amazing Shannon Toalson, Kate Kennedy, and Carly Humes, who I feel very lucky to call my close friends.

Lastly, although it may seem silly, perhaps the biggest thank you of all goes to my sweet Xena girl. Despite the significant amount of my limited income and time that goes to caring for Xena (training, food, and rent she doesn't even contribute to!), I don't think I'd survive the stressful days and late nights studying and working without our twice daily walks and her goofy personality, expressive eyes, and wiggly ears.

Table of Contents

Introduction 1

The Problem: Gender and Race Bias in Employment 2

Models of Impression Formation 4

Race and Gender Stereotypes Contributing to Employment Decisions 7

Stereotypes, Ambiguity, and Performance 10

The Present Research 11

Hypotheses 12

Study 1 15

Method 15

Results 22

Discussion 30

Study 2 31

Method 32

Results 34

Discussion 43

General Discussion 44

References 70

Appendix A: Job Description 78

Appendix B: Resumes 79

Appendix C: Interviews 81

Appendix D: Writing Sample 85

Appendix E: Questionnaire Items 86

Appendix F: Criteria Ranking Analyses ... 93

List of Figures

Figure 1: Bayesian Racism moderation interaction, proportion hired, Study 2..... 66

Figure 2: Bayesian Racism moderation interaction, objective ratings, Study 2 67

Figure 3: Bayesian Racism moderation interaction, subjective ratings, Study 2..... 68

Figure 4: Structural Racism moderation interaction, subjective ratings, Study 2 69

List of Tables

Table 1: Sample sizes for each experimental condition, Studies 1 and 2 55

Table 2: Proportion hired, Study 1 56

Table 3: Descriptives of unstandardized objective ratings, Study 1 57

Table 4: Objective ratings, Study 1 58

Table 5: Engineering competence ratings, Study 1 59

Table 6: Assertiveness ratings, Study 1 60

Table 7: Proportion hired, Study 2 61

Table 8: Descriptives of unstandardized objective ratings, Study 2 62

Table 9: Objective ratings, Study 2 63

Table 10: Warmth ratings, Study 2 64

Table 11: Summary of moderation analyses, Study 2 65

Introduction

Women and people of color have been historically and empirically discriminated against in both the job selection process and the workplace thereafter. A typical job selection process involves multiple points in time for people to form (and potentially change) impressions of candidates, yet many experimental conceptualizations of the process provide only one point in time for an evaluation, rendering the process less realistic and valid. In contrast to this research, the job selection process in the real world is becoming increasingly complicated and generally involves several steps, including initial calls for resumes, multiple phone interviews, and conversations with high-ranking members of the team. In this scenario, there are several points in time for people to form and change their impressions of a candidate. Additionally, race information (and sometimes gender information, as well) is often not revealed until the interview stage, after the initial resume review. Prior research falls short of capturing the full picture of hiring and workplace discrimination because it typically provides only one point in time for evaluations after *all* information about the candidate has been reviewed.

Furthermore, the job market is one of the most important contexts in which the effects of historical, institutional, and interpersonal racism, sexism, and other forms of discrimination coalesce to disadvantage people, especially women of color (Castilla, 2008). In part because of laws that prohibit discrimination, modern-day discrimination is often no longer blatant (Dovidio & Gaertner, 2000). But, this subtlety may be dangerous because it is less noticed, less actionable, and more pervasive. The current research seeks to understand how impressions of job candidates and their performance differentially change over time based on gender and race information, especially when performance information is ambiguous, counter-stereotypic, or surprising. This

"updating" of impressions (Heilman et al., 2019) can happen at multiple times during an interview process, cumulatively and adversely affecting women and applicants of color.

The Problem: Gender and Race Bias in Employment

Employment discrimination along gender lines is well documented, as women face many disadvantages in the workplace. While women hold 52% of management and professional level jobs, they hold far fewer leadership positions (Warner et al., 2018). Only 6.6% of CEOs of Fortune 500 companies are women, and only 22.5% of board seats for Fortune 500 companies are held by women (AAUW, 2016). These numbers are even more dire for women of color. Only 0.8% of CEO positions in Fortune 500 companies and 3.2% of leadership roles at large U.S. companies are held by Black women (Brooks, 2019).

Race bias in employment can be traced not only to discrimination and prejudice in the present but also to many historical factors, including slavery, the Jim crow era, school segregation, and redlining, to name a few (Hanks et al., 2018). Research has demonstrated the relationship between the current racial gaps we see in education, wealth, and employment and slavery in the United States (Bertocchi & Dimico, 2010). In fact, the unemployment rate among African Americans has been, more often than not, double the rate of unemployment among White Americans, dating back to 1972 when the U.S. Bureau of Labor and Statistics first started collecting data on unemployment rates among African Americans (Ajilore, 2020). Not only do people of color face difficulties getting hired in the first place, but these inequities are magnified at higher levels of leadership. In 2018, only 3.3% of all executive or senior leadership roles were held by Black employees, and less than 1% of CEOs of Fortune 500 companies were Black (Sahadi, 2020). This inquity in hiring also extends to the wage gap. Black men are paid significantly less than their White colleagues; according to a 2020 report by the Economic Policy

Institute, in 2000, the wage gap between Black and White men was 21.8%, and this increased to 26.5% in 2019 (Gould, 2020).

A recent meta-analysis of 24 field experimental studies of hiring discrimination from 1989 to 2015 indicated that there has been no change in rates of discrimination against Black applicants in the last 25 years (Quillian et al., 2017). White applicants received 36% more callbacks on average compared to African American applicants who were just as qualified. The same meta-analysis indicated a marginally significant downward trend in the levels of discrimination over time against Latino applicants; White applicants received 30% more callbacks than equally qualified Latino applicants in 1990 but only 15% more callbacks than Latino applicants in 2010 (Quillian et al., 2017). However, small sample sizes of field experiments using Latino applicants (n = 9) warrant caution when interpreting these results.

In response to these dire statistics, applicants of color have even taken to "whitening" their resumes in an attempt to get past the initial resume review stage and get more interviews (Kang et al., 2016). Applicants who whiten their resumes engage in a variety of strategies, including changing the presentation of their name (using a more American-sounding name or a middle name), omitting experiences (e.g., not including being President of the Black Student Association on their resume), changing their experiences (e.g., saying they were President of the *Student Association* versus the *Black Student Association*), and adding stereotypically White experiences (e.g., including activities like snowboarding or hiking). Kang and colleagues (2016) suggest that this whitening of resumes can serve two purposes: getting one's foot in the door (i.e., getting past the resume reviewing portion of the process to an interview, where race/ethnicity information would later be revealed) or signaling assimilation and conformity to avoid the influence of stereotypes. Because race information is often easier to conceal on paper

than gender information (e.g., due to the gendered nature of names, etc.), female applicants often cannot undertake the same strategies.

In the studies reported here, I examine how evaluators form impressions of job candidates in a masculine work domain based on gender (Study 1) and race (Study 2). In both studies, clear gender/race information is delayed until after an initial impression is formed. I next review models of impression formation that form the basis of predictions outlined later.

Models of Impression Formation

Minimal research has been conducted in the domain of workplace discrimination that examines how multiple pieces of information (i.e., performance over time) are integrated into an impression on which a final hiring decision is made. However, much research *has* been dedicated to understanding, more generally, how people blend together what they know about a person into a cohesive impression (Asch, 1946; Anderson, 1965; Brewer, 1988; Fiske & Neuberg, 1990; Kunda & Thargard, 1996). Each impression formation model considers the intake and use of information to create an impression, but each differs in the ways in which information is integrated into a cohesive whole.

Both Asch (1946) and Anderson (1965) argued that when forming an impression, people use all of the information they know about a person (Kunda & Thagard, 1996); however, Asch and Anderson differed in their accounts of how this integration of information occurs. Asch (1946) took a gestalt perspective and assumed that perceivers create a coherent, singular impression, taking into account individual features and how those features are related to each other. Anderson (1965), on the other hand, took a more piecemeal approach that assumed perceivers consider the implications of each different piece of information separately, and then algebraically combine them into a summary impression.

Asch (1964) and Anderson (1965) both consider the individual person as the basis for person perception and impression formation (Brewer, 1988). However, later models of impression formation broadly distinguished between individuating and category-based information. Category-based information is information about a person's perceived social categorical membership (e.g., woman, African American, etc.), while individuating information is information about a person's particular attributes or any other information known about the person outside of their social category membership (e.g., is a dentist, has two kids, etc.) (Brewer, 1988; Kunda & Thagard, 1996).

Brewer's (1988) dual process model of impression formation processes distinguishes between both automatic and controlled processing and category-based versus individuating information. When a person is initially perceived, automatic processes lead to categorization, a stage called "identification." If the target is irrelevant to the self, no further processing occurs, but if the target is self-relevant, controlled processing begins. When self-involvement is low (the person is deemed unimportant to know), category-based processing occurs, involving reliance on schemas, heuristics, and prior expectancies about people from the categories engaged (i.e., stereotypes). Stereotype-based perceptions continue unless the target is perceived as a poor fit to the category, when individuation occurs (the target is viewed as an exception to the category). When self-involvement is high, the perceiver engages in personalization based on individuating information, which is more detailed and requires more effort (Moskowitz, 2005).

Similar to Brewer's dual-process model, Fiske and Neuberg's (1990) continuum model of impression formation argues that impression formation occurs along a continuum. Category-based information lies on one end of the continuum while individuating or attribute-based information lies on the other. Fiske and Neuberg's model posits that social categories form the

organizational basis for social information, and that category-based impression formation processes are attempted before any individuating information is incorporated into the impression. If this process is successful, which depends on how easy it is to match attribute information with a category, then attribute-oriented impression formation processes do not occur (Fiske & Neuberg, 1990). Processing individuating information requires more attention than category-based information. In this model, compared to the dual process model (Brewer, 1998), individual, piecemeal information is only used if and when categorization fails. Anderson (1965), on the other hand, conceptualized the entire impression formation process as this type of piecemeal information integration process, and ignored the role of category information.

Finally, Kunda and Thagard's (1996) parallel-constraint satisfaction model of impression formation specifically concerns the use of stereotypes versus individuating information. In the case of their model, stereotypes are types of category-based information and concern group membership and are about traits or behaviors believed to be associated with membership in that group (Kunda & Thagard, 1996). Kunda and Thagard (1996) take an approach most similar to Asch. In their parallel-constraint satisfaction model of impression formation, stereotypes and individuating information are processed simultaneously, rather than in a piecemeal, ordered fashion. They work together to influence each other—through both excitatory and inhibitory links—and create an overall impression. Importantly, stereotypes/category-based information and individuating information are treated similarly—as pieces of information to add to the puzzle—with neither being given "special status," as in Brewer and Fiske and Neuberg's models (Kunda & Thagard, 1996, p. 287). An example best illustrates the model: if one observes a Black person pushing someone, that activates thoughts of "violent push" and "friendly shove," but the image of a Black person also (stereotypically) activates thoughts of "aggression." This then

further activates “violent push” and deactivates “friendly shove.” However, if the individual pushing someone was White, the “aggression” stereotype would not be activated, and “violent push” does not receive the higher activation that it did with the Black man. In this case, stereotypes constrain the meaning of a particular behavior, and the level of activation of a certain characteristic is indicative of how much of a characteristic the person is assumed to have (Kunda & Thagard, 1996).

The hiring process provides multiple opportunities for impression formation and impression maintenance or updating. The hiring discrimination literature has primarily focused on providing participants with all the information about a potential employee at one point in time and asking them to make a decision. In the real world, impressions are formed over time and may change depending on new information provided later in the process. There are multiple opportunities for categorical information, stereotypes, and individuating information to be integrated, discarded, categorized, or recategorized to form an impression. I now turn to a review of the stereotypes facing women and applicants of color in the hiring process to understand how these stereotypes could affect the impression formation process.

Race and Gender Stereotypes Contributing to Employment Decisions

During the impression formation process, at least according to some models (Brewer, 1988; Fiske & Neuberg, 1990), categorization (e.g., based on gender and race) is a first, automatic step (e.g., see Brewer, 1988; Devine, 1989). Once these social categories are activated, the stereotypes accompanying those categories are also activated, and the perceiver must engage in effortful processing in an attempt to incorporate other information (Brewer, 1988; Fiske & Neuberg, 1990) or disengage the stereotype (Devine, 1989). Much social psychological research

has been dedicated to understanding how stereotypes impact perceptions of job candidates, especially female candidates and candidates of color.

Both the Lack of Fit Model (Heilman, 1983, 2001, 2012; Heilman & Caleo, 2018; Heilman et al., 2019) and Role Congruity Theory (Eagly & Karau, 2002; Eagly, 2004) were developed as ways of explaining gender discrepancies in the workplace. The Lack of Fit Model, based on research about gender stereotypes, suggests that gender bias is a product of a perceived lack of fit between the assumed requirements of a job (or task or role) and what men and women are perceived to be like. This is particularly a problem when women are applying for or working in jobs that are historically male-dominated (e.g., computer science), but the same effect has also been demonstrated for men working in more stereotypically female jobs (e.g., education) (Heilman et al., 2019). This expectation of negative performance for the incongruent applicant then leads to an assumption and/or perception of incompetence. Consistent with traditional role expectations, in a study concerning what employers value in applicants, Quadlin (2018) demonstrated that employers value competence and commitment in male applicants, but likeability more than competence or commitment in female applicants.

Similarly, Role Congruity Theory (Eagly & Karau, 2002; Eagly, 2004), developed in the context of studying leadership specifically, posits that an incongruency between the stereotypes about a social group and a certain social role creates a context in which prejudicial evaluations could occur. Prejudice resulting from this incongruency can take two forms: regarding women as less qualified than men for leadership roles and evaluating women's behavior in leadership roles less positively than the same behavior from men (Eagly & Karau, 2002).

Additionally, in higher status, STEM contexts especially, the field-specific ability beliefs hypothesis (FAB; Meyer et al., 2015) may be able to explain some persistent gender gaps. The

FAB hypothesis posits that women are more likely to be underrepresented in certain fields that are thought to require "raw intellectual talent" (Meyer et al., 2015, p. 1) or brilliance (i.e., primarily STEM fields), an attribute more stereotypical of men than women. Leslie and colleagues (2015) demonstrated that women were underrepresented in fields thought, by academics in the same fields, to require "stable, raw talent" (p. 3). This "brilliance" factor explained gender discrepancies better than other potential explanations, such as willingness/ability to work long hours, aptitude/selectivity, or differences in systematizing versus empathizing (Leslie et al., 2015). This same pattern was found with non-academics as well. In addition to perceivers under-valuing women, beliefs that certain fields require "brilliance" could result in young women doubting their ability to succeed in those fields; this, in turn, could hinder their pursuit of careers in the same disciplines, further perpetuating the gender gap.

Research examining the Lack of Fit Model, Role Congruity Theory, and the FAB hypothesis has primarily focused on gender, but there is no reason to believe these same processes could not be extended to a perceived lack of fit or incongruency in expectations between a certain job and perceptions of a person based on racial and other stereotypes. Many of the same stereotypical characteristics implicated for women in both models (e.g., competence assumptions, etc.) have been established as stereotypes for Black candidates as well (Devine & Elliot, 1995; Biernat & Kobrynowicz, 1997; Biernat et al., 2009). Additionally, research suggests that although individual belief in or endorsement of these stereotypes, as well as the specific content of the stereotypes, may be changing over time, the overall valence of the stereotypes against Blacks is consistently negative (Devine & Elliot, 1995).

The Shifting Standards Model (SSM; Biernat et al., 1991; Biernat & Manis, 1994) provides another account of race and gender differences in employment. The SSM posits that

when making judgments about a stereotyped group, people will use *within-category standards* as references for their judgments (Biernat et al., 1991; Biernat & Manis, 1994). On an objective scale (i.e., a concrete or universal scale, like GPA), ratings tend to follow a stereotypic pattern (i.e., assimilate to the stereotype), in that the person thought to be "deficient" in an attribute is rated lower than the person considered to have more of the attribute. However, on subjective rating scales, this pattern disappears or is reversed; the negatively stereotyped individual is rated as having just as much or more of the attribute than the non-stereotyped individual, presumably because judgment is made with reference to (lower) within-group standards. Examining stereotypes about academic competence, Biernat and colleagues (2009) found that people rated White students as having higher ACT scores than Black students, but when the same ratings were made on subjective scales, the pattern was non-existent. Research applying the SSM to race-based stereotypes has demonstrated that Black individuals often have to work harder to be perceived as just as good as their White counterparts (Biernat & Kobrynowicz, 1997; DeSante, 2013). Because of stereotype-based expectations, women and Black employees (or other low-status individuals) may meet minimum standards fairly easily, but they must work harder to meet confirmatory standards of competence (Biernat & Kobrynowicz, 1997). Many of these studies have focused on White males and females or Black and White males, but fewer have taken an intersectional approach to examine stereotypes and evaluations of individuals with multiple marginalized identities (e.g., Black women). The SSM suggests that the effects of category membership (race and gender) in employee evaluations may vary depending on the type of judgment being made (i.e., objective judgments and zero-sum decisions versus subjective trait ratings).

Stereotypes, Ambiguity, and Performance

Stereotypes are more likely to influence judgments of individuals when individuating information about them (e.g., performance) is ambiguous (Dovidio & Gaertner, 2000) and/or when the person being judged does not fit the stereotype associated with the judgment context (Heilman et al., 2019). In a college application review paradigm, discrimination against Black students occurred among high-prejudice participants when the student's qualifications were mixed or ambiguous, but not when the student's credentials were consistent (strong or weak) (Hodson et al., 2002). This same pattern was found when participants reviewed candidates for a peer counseling program, even when self-reported prejudice was low (Dovidio & Gaertner, 2000). In another study, when a student's performance declined across a semester in a male-stereotyped academic course, female students were disproportionately penalized on competence ratings and performance evaluations compared to male students; conversely, when performance improved, this improved performance was less beneficial for female than for male students (Heilman et al., 2019). The opposite pattern was found in a female-stereotyped academic course. Similarly, Steinpreis and colleagues (1999) found that academic psychologists were more likely to hire a male applicant for an assistant professor position than an equally qualified female applicant, but this bias did not occur for applicants up for tenure. Steinpreis et al. (1999) suggest that qualifications were very high for the applicant up for tenure, and therefore there was no room (i.e., no ambiguity) for stereotypes to influence the hiring decision (see also Moss-Racusin et al., 2010). Expectancies in general are more likely to influence judgment when information is ambiguous rather than clear-cut (Olson et al., 1996).

The Present Research

The current research addresses gaps in the literature by employing a methodology that mimics the temporally extended hiring process in a more ecologically valid way. These studies:

(a) will model how impressions of ambiguous performance information are *differentially* interpreted based on gender and race; (b) are among the first to examine the potential role of bias across multiple time points in the job selection process; and (c) examine how the revelation of gender and/or race information *after* initial impressions are made changes subsequent performance evaluations. Study 1 will examine gender (female/male), and Study 2 will examine race (Black female/White female). Participants, recruited for a study on perceptions of job application materials, will view a job description for a Mechanical Engineering Internship (a White, masculine domain), followed by the application package of a potential candidate. The application package includes a resume, an excerpt from their interview (where gender [Study 1] or race [Study 2] information is revealed), and a writing sample. While both the resume and the writing sample will be perceived as average, performance will be manipulated in the interview excerpt, creating three conditions: poor, average, and very good interviews. After each application component, participants will rate the applicant on a variety of characteristics, including competence and likelihood of success in the position, as well as other attributes implicated in gender and race stereotypes, such as assertiveness and warmth. After all application components have been viewed, participants will make a final hiring decision. Finally, participants will rank the importance of each criteria (e.g., resume, interview, writing sample) for their hiring decision.

Hypotheses

Prior research on shifting evaluation standards shows that perceivers view female/Black targets as subjectively more competent than comparable male/White targets, yet they continue to favor male/White targets in selection decisions, following stereotyped beliefs that male/White candidates are better fits (Biernat & Manis, 1994; Biernat & Kobrynowicz, 1997; Heilman et al,

2019). Based on these findings, I expect that Studies 1 and 2 will show that, after candidate gender/race is revealed, participants will *subjectively* rate the Black/female candidate higher on competence and intellect (within each interview performance condition), yet their hiring decision will nevertheless favor White/male candidates. Results in this direction would be in line with previous work on the Shifting Standards Model, where subjective ratings are equal or higher for the female (Study 1) or Black female applicant (Study 2), but objective judgments (i.e., the hiring decision) follow the stereotypic pattern where the female or Black female candidate are not hired at the same rates.

The role of ambiguity. Ambiguous situations create the perfect context for stereotyping to occur. Previous work has conceptualized ambiguity differently in terms of performance. In some instances, ambiguity is conceptualized as *average performance* (i.e., an evaluator cannot really determine how good or how bad performance is; Dovidio & Gaertner, 2000). In other instances, ambiguous performance is characterized by *inconsistency in performance* (i.e., poor performance in one instance and high performance in another; Hodson, Dovidio, & Gaertner, 2002). For the purposes of these studies, ambiguity is conceptualized as average performance across all three time points[1], but it should be noted that some ambiguity is created in the other two conditions (poor vs. very good interview performance) because of the change from average

[1] I argue that the average performance condition (i.e., average across all three time points) is the context in which the most ambiguity is created over time. The study design always includes average performances at Times 1 and 3; performance is only manipulated at Time 2 (the interview). Therefore, I believe that the general perception of performance will match that of performance during the interview. Thus, when interview performance is poor, general perceptions of overall performance will be low, and when interview performance is very good, general perceptions of overall performance will be higher. When performance in the interview is average, there is no additional information that gives any more indication of how performance should be perceived, and evaluators will thus use their own expectations of performance based on stereotypical information.

performance on the resume, poor/very good interview performance, and an average writing sample. I hypothesize that there will be greater discrimination in the average interview condition. Average performance across all three time points will result in evaluators having to make inferences about performance, and those inferences will be influenced by expectations of how people should perform in certain contexts. Therefore, in Study 1, competence stereotypes about female applicants should be activated, and they will be hired less (even if rated as good as or better than the male candidate on other characteristics). The same should happen for Black female candidates in Study 2.

In both studies, I expect to see a main effect of performance, such that when interview performance is very good, the applicant is hired more often, and when interview performance is poor, the applicant is hired less often, regardless of gender or race. We would not expect to see high levels of discrimination in the poor interview performance condition, as poor performance at the interview stage should indicate the candidate is not suitable for the job; however, we may see slightly more discrimination against the female/Black female applicants, as a poor interview might confirm stereotypes of incompetence. The same pattern could be seen in the very good interview performance condition, such that the very good interview is enough to overcome any stereotypes of incompetence, and the female and Black female applicants may be hired at the same rate as the male and White female applicants.

Additionally, I hypothesize that the relative importance of specific criteria should shift depending on target gender/race. When interview performance is poor, participants should overrate the importance of the interview for the female and Black applicants compared to the male/White applicants. When performance in the interview is high, the interview should be seen as less important for the female and Black applicants. When performance is average, the female

and Black applicant's interview performance should be rated as less important than resume strength. Hypotheses for Study 1 were registered at AsPredicted.org under AsPredicted#92897.[2]

Study 1

Study 1 examines how people's initial impressions of a candidate change when presented with new information about both performance and gender. The study uses a 2 (Gender: Female, Male) x 3 (Interview Performance: Poor, Average, Very Good) between-subjects design, with Time of Evaluation (3) as a within-subjects factor. Subjective trait ratings (e.g., judgments of competence) were made at each time point, but key objective judgments (e.g., hiring, salary, etc.) were collected only at Time 3.

Method

Participants

Participants were 350 U.S. adults, 18+ years-old, recruited through the Prolific crowdsourcing site and balanced based on gender. An a priori statistical power analysis was conducted using G*Power 3.1 (Faul et al., 2009). With an alpha = .05 and power = 0.90, the projected sample size needed for an effect size of 0.20 (F-test) was approximately $N = 320$. All participants successfully passed all three attention check questions, but a total of 30 participants were removed for failing manipulation checks: 13 incorrectly remembered the target's race, and 17 incorrectly remembered the target's gender. One additional participant was removed for not consenting to participate in the research. The final sample consisted of 319 participants *(M_{Age}* = 36.5 years, SD_{Age} = 13.6 years). One hundred and sixty participants identified as male, 148

[2] I did not specifically preregister Study 2, but predictions are identical for those in Study 1, though the focus is on race in Study 2 rather than gender in Study 1.

participants identified as female, 9 participants identified as gender non-binary, and 2 participants identified as trans. Additionally, 227 participants identified as White, 15 as Black/African American, 19 as Latinx, 28 as Asian/Pacific Islander, 1 as Native American, 2 as Middle Eastern, and 27 as mixed race or "other."

Procedure

The study aimed to mimic a job selection process in which participants formed an initial impression of a candidate based on a name and resume, and later learned information about gender. Participants were recruited for a study examining impressions of job application materials. They were told that they would first view a job description followed by the candidate's application package, which included the applicant's resume, excerpts from an interview with the applicant, and a writing sample; after each piece of information, they answered some questions about the application materials they saw and rated the applicant on a variety of characteristics (including competence, effectiveness, likelihood of success, assertiveness, and warmth, before making a final hiring decision, among other key objective outcome variables including salary, projects completed, and projected GRE scores. Participants were also asked to rank the relative importance of criteria (e.g., GPA, previous work experience, quality of interview) for their hiring recommendation. Previous research has shown that participants shift the criteria they indicate as being most important to their hiring decisions. For example, Phelan and colleagues (2008) demonstrated that when participants evaluated an agentic male job applicant, they rated competence and social skills as being equal determining factors in their decision, but for agentic female applicants, social skills were rated as more important.

Materials

Job description. Participants first saw a job description for a Mechanical Engineering Intern at Xenatech Engineering, Inc., a fictional engineering firm on the West coast. A Mechanical Engineering Intern position was chosen because it is a high-profile, male-dominated, primarily White occupation; of mechanical engineers, 91.2% are male, and 74.4% are non-Hispanic Whites (Deloitte & Datawheel, n.d.). The company was described as looking for a "brilliant" (Leslie et al., 2015) soon-to-graduate student or recent graduate for their 10-week summer internship program, which has the potential to turn into a full-time job in the future. Requirements for the job included a degree in Engineering, comfort working in a laboratory environment, attention to detail, and good scholastic standing. The full job description can be found in Appendix A. Pretesting, using an independent sample of 65 U.S. adults, 18+ years-old, from Prolific, indicated that a majority of participants viewed the Mechanical Engineering Internship as one that a White male would be mostly likely to apply for. Ninety-four percent of participants indicated that they would expect people who apply for this job to be White, and 95% of participants expected that applicants would be male. Additionally, participants said they would expect 72% of employees in a Mechanical Engineering Internship to be White (M = 72.38, SD = 12.90) and 77% of employees in the Internship to be male (M = 77.32, SD = 12.44). Finally, 88% of participants thought there was a stereotype that a woman would be *less* competent at this job than a man, and 77% of participants thought there was a stereotype that a Black person would be *less* competent at this job than a White person.

Resume. Next, participants viewed the applicant's resume, which was kept the same across all participants. The single page resume was for a candidate named Charlie W*****, who is looking for a job after graduation and has average qualifications. The name "Charlie" was

chosen because while it is often viewed primarily as a male name (Van Fleet & Atwater, 1997), it is a name seen for some women (e.g., often short for Charlotte); pretesting indicated that 97% of participants thought "Charlie W." was a male name, and 84% said it was a White name. One aim of this research is to understand how gender information impacts the interpretation of performance information, especially when the gender information is *surprising* or *incongruent* with the job context. By starting with a stereotypical assumption of "maleness" for both the job type and the resume, the revelation of gender in the female candidate condition should be "surprising," and this should interact with performance condition to negatively impact the female candidate, particularly when performance is ambiguous/average. Furthermore, in typical job hiring processes, gender (and more often—race) information may not be revealed until the interview portion of the process (i.e., gender and race information is not always clear from the resume). The resume was pretested to ensure that the information conveyed reflects an average student whose GPA (3.03/4.00), extracurricular activities (e.g., soccer team, volunteer at soccer tournament), and previous work experience (e.g., summer intern at another engineering firm, crewmember at fast-food restaurant) are modest. Pretesting indicated that the resume was perceived as average (M = 4.185 on a 7-point scale, SD = 1.102). Upon viewing the resume, 89% of pretesting participants expected the candidate to be male, and 86% expected the candidate to be White. The full resume can be found in Appendix B.

Interview. Participants were randomized into one of six conditions where they saw a photo (to manipulate gender) of the applicant (either White female or male) and read an excerpt of their interview (manipulated to be either poor, average, or very good; see Appendix C). The photos were selected from the Chicago Face Database and were matched on age, gender prototypicality, and attractiveness (Ma et al., 2015). Pretesting confirmed that each interview was

perceived as intended (on 7-point scales, M_{Poor} = 2.59, SD_{Poor} = 1.18; M_{Avg} = 4.15, SD_{Avg} = 1.46; M_{VGood} = 5.30, SD_{VGood} = 1.18). *T*-tests indicated that all interviews were perceived as significantly different from each other (*p*s < .01).

Writing sample. To further simulate the accrual of additional information about job candidates, participants were then asked review a brief writing sample (the same across all conditions) that is part of the candidate's application package (see Appendix D). The writing sample was intended to be perceived as average, and pretesting indicated that this was the case (M = 3.45, SD = 1.26).

Measures

Manipulation check on quality of the materials. After viewing the resume, interview transcript, and writing sample, participants rated each on a scale from 1 (Very Poor) to 7 (Very Good). The question read: "How would you rate the quality of the [resume/interview/writing sample]?" All measures and answer scales can be seen in Appendix E.

Subjective ratings. At each of three time points (after viewing the resume, interview, and writing sample), participants assessed the applicant's competence, likelihood of success, and effectiveness in the Engineering internship, as well as the applicant's perceived competence, assertiveness, and warmth more generally. All attributes were assessed on a scale from 1 (e.g., Not at all [competent, successful, warm, etc.]) to 7 (Very [competent, successful, warm, etc.]). The Engineering-specific questions read: "How [competent/successful/effective] do you think Charlie would be in the Engineering Internship?" The general attribute questions were assessed by asking: "How competent, in general, do you think Charlie is?", "How assertive do you think Charlie is?", and "How warm (i.e., friendly) do you think Charlie is?"

Hiring decision. After viewing all the job application materials, participants decided whether or not they would hire the applicant for the Mechanical Engineering Internship by answering "Yes, I would hire Charlie for the internship" or "No, I would not hire Charlie for the internship" to the question "Would you hire Charlie for the Mechanical Engineering Internship?" Additionally, participants answered an open-ended question about why they made their decision.

Criteria ranking. Participants were also asked, "How important were the following criteria in your decision to hire or not hire Charlie for the Engineering Internships?" Participants were instructed to rank the following criteria from 1 (most important) to 6 (least important) in their hiring decision: undergraduate GPA, previous work experience, extracurricular/service activities, interview quality, and writing sample quality.

Objective judgments. An objective rating variable was created by computing a composite score of three standardized dependent variables: salary, projects completed, and projected GRE score ($\alpha = 0.72$). For the salary variable, participants indicated how much salary/stipend they felt the applicant should be paid during the 10 weeks. The question read: "Internships like this, given the limited time frame, are usually salaried, rather than paid by the hour. The salary/stipend can range from about $3,000 to $10,000 for the 10 weeks. If Charlie was hired, how much do you think Charlie should be paid for the 10-week Internship?" Participants marked, on a sliding scale, a dollar amount between $3,000 and $10,000.

Participants also estimated the number of projects they thought the applicant would complete during the 10 weeks. On a sliding scale, marked from 1 to 6, participants answered, "If [the applicant] was hired, how many projects do you think [the applicant] would complete during the Internship?" if "An Intern in this type of program generally completes between 1 and 6 projects during the course of their 10-week Internship."

Finally, participants estimated the applicant's GRE score (if they were to take it), on a sliding scale from 260 to 340, after reading that "the GRE (Graduate Record Examination) is a standardized test often used for admission to graduate school. Total GRE scores can range from 260 to 340."

Hired full-time. Finally, participants who hired the applicant for the internship decided ("Yes" or "No") if they should be hired at full-time after the Internship. The question read: "Do you think Charlie should be hired at Xenatech full-time after the Internship is completed?"

Manipulation checks. Participants also confirmed the applicant's race and gender as manipulation checks (see Participant section for information on deletions based on missed manipulation checks).

Potential Moderators

Structural sexism. Participants completed a measure of structural sexism. The question read: "Please indicate to what extent you believe sexism is an individual problem (i.e., sexism is individual bias towards another person or group of people) or a structural problem (i.e., sexism is part of the structure and laws of our institutions." Responses were on a scale from 1 (Individual problem) to 8 (Structural problem).

Attitudes toward Affirmative Action. Participants also completed the six-item Attitude Toward Affirmative Action Scale ($\alpha = 0.86$; Kravitz & Platania, 1993). Affirmative action plans might be seen as threatening to men or Whites, and therefore may impact their own conceptions of fair hiring processes and their simulated hiring decisions. Responses on the Attitude toward Affirmative Action Scale were given on a 5-point Likert-type scale, from 1 (strongly disagree) to 5 (strongly agree). Items include "*Affirmative action is a good policy*," "*I would* not *like to work at an organization with an affirmative action plan*" (reverse scored), and "*Employees should be*

actively involved in attempts to improve the affirmative action conditions at their place of employment."

Modern Sexism Scale. Participants completed the 8-item Modern Sexism Scale (MS; Swim et al., 1995). Modern sexism is characterized by beliefs that sexism is "a thing of the past" feelings of resentment towards women's demands for equal economic and political power and policies that strive to create (p. 200). Items on the Modern Sexism Scale ($\alpha = 0.83$) include "*Women often miss out on good jobs due to sexual discrimination*" (reverse-scored), "*Society has reached the point where women and men have equal opportunities for achievement*," and "*Discrimination against women is no longer a problem in the United States*."

Demographics. Participants were also asked a series of demographic questions that could potentially moderate evaluations and hiring decisions, including participant race, gender, and political ideology/identification. Political ideology was assessed by asking: "How would you describe your political views? [Very Conservative, Conservative, Moderate, Liberal, Very Liberal]" and political identification was assessed by asking: "In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent? [Republican, Democrat, Independent, or Other]."

Results

In line with previous research, hypotheses, and the pre-registration, results were analyzed using the full sample ($N = 319$). However, initial analyses (reported below) indicated few findings involving target gender. Exploratory analyses were then conducted among White participants, and participants initially excluded for failing the manipulation checks were maintained in an effort to increase the resulting sample size ($N = 249$). Results will be reported using the full sample, and any different patterns in the White sample will be noted.

Manipulation check on quality of application materials

Similar to the pre-testing, the materials were perceived as intended. The resume was generally perceived as average (M = 4.56, SD = 1.20). The interviews were also perceived as intended (M_{Poor} = 2.75, SD_{Poor} = 1.16; M_{Avg} = 4.28, SD_{Avg} = 1.44; M_{VGood} = 5.56, SD_{VGood} = 1.02), and the writing sample was perceived as average (M = 3.85, SD = 1.54). The 3 interview conditions were statistically significantly different from each other, $F(2, 316) = 143$, $p < .001$. Pairwise comparisons using the LSD procedure indicate that each interview performance condition was perceived as significantly different from the others (all ps < .001).

Hiring decision

The primary research question concerned whether hiring rates differ among male and female targets based on performance, represented by a 2 (Target Gender: Female, Male) x 3 (Performance: Poor, Average, Very Good) interaction. Table 1 shows the sample sizes for each condition in both the full and White samples, for both Study 1 and Study 2. Table 2 indicates the proportion of male and female targets hired in each performance condition. A chi-squared test of independence showed no interaction between target gender and performance, $\chi^2(2, N = 319) = 0.467$, $p = 0.792$. The gender main effect was also non-significant, $\chi^2(1, N = 319) = 2.065$, $p = 0.151$, but there was main effect of performance, $\chi^2(2, N = 319) = 30.065$, $p < .001$, with targets in the Very Good performance condition hired at a greater rate (M = 0.639, SD = 0.483) than targets in the Average performance condition (M = 0.514, SD = 0.502) and targets in the Poor performance condition (M = 0.274, SD = 0.448). Pairwise comparisons using the LSD method indicated that the mean differences between the Poor and Average performance conditions and the Poor and Very Good conditions were statistically significant, ps < .001, and the difference between the Average and Very Good conditions was marginally significant, $p = 0.061$. This main

effect of performance was in line with hypotheses, but my hypothesis concerning the male candidate being hired at greater rates than the female candidate was not supported.

Objective ratings

Participants were asked to assign the target a salary amount, indicate the number of projects the target would complete during the internship, and estimated the target's projected GRE score. Table 3 contains the means and standard deviations for each variable, as well as intercorrelations. Scores on each of these variables were then standardized and combined into a mean composite score ($\alpha = 0.72$). Table 4 provides the means and standard deviations on the standardized objective rating index by target gender and performance, for both the full and White samples. In the full sample, the Target Gender X Performance ANOVA indicated a significant main effect of performance, $F(2, 313) = 6.642$, $p = 0.001$, $\eta_p^2 = 0.040$, no main effect of gender, $F(1, 313) = 1.323$, $p > .05$, $\eta_p^2 = 0.004$, and a significant Target Gender x Performance interaction, $F(2, 313) = 3.191$, $p = 0.042$, $\eta_p^2 = 0.020$. Pairwise comparisons indicated that in the Poor performance condition, men ($M = -0.415$, $SD = 0.627$) were rated more negatively than women ($M = -0.010$, $SD = 0.845$), $F(2, 313) = 7.074$, $p = 0.008$, $\eta_p^2 = 0.002$. Objective judgments did not differ by gender in the Average and Very Good conditions ($ps > .05$).

In the *White Sample,* the main effect of performance was again significant, $F(2, 243) = 9.393$, $p < .001$, and the gender effect was nonsignificant, $F(1, 243) = 0.899$, $p = 0.344$. The interaction was also statistically significant, $F(2, 243) = 4.558$, $p = 0.011$, but the pattern of findings was different. Consistent with predictions, it was in the Average performance condition that male targets ($M = .287$, $SD = 0.754$) were rated significantly higher than female targets ($M = -0.142$, $SD = 0.799$), $F(1, 243) = 6.304$, $p = .013$, $\eta_p^2 = 0.025$. Gender effects were nonsignificant ($ps > 0.05$) in the Poor and Very Good conditions. Thus, my predictions regarding discrimination

occurring in the average condition on objective measures was supported among White participants only.

Subjective judgments over time

At each of the three time points, participants rated the applicant's competence, likelihood of success, and effectiveness in the Engineering Internship specifically; these three items were combined at each time point into a general engineering competence score (αs = .94-.97). Participants also rated the applicant's (general) competence, assertiveness, and warmth (one item each at each time point).

Engineering Competence. A 2 (Target Gender: Female, Male) x 3 (Performance: Poor, Average, Very Good) x 3 (Time) repeated measures ANOVA was conducted, with Time as a within-subjects factor, for the engineering competence composite scores at all three time points. There was a main effect of Time, $F(2, 626) = 75.466$, $p < 0.001$, $\eta_p^2 = 0.194$, Performance, $F(2, 313) = 19.329$, $p < .001$, $\eta_p^2 = .110$, and a significant Time x Performance interaction, $F(4, 626) = 46.326$, $p < 0.001$, $\eta_p^2 = 0.228$. In the Poor performance condition, engineering competence scores were significantly higher at Time 1 ($M = 5.11$, $SD = 1.16$), before performance information was known, than at Time 2 ($M = 3.69$, $SD = 1.31$), $p < .001$. Scores then increased, but non-significantly, p = .280, at Time 3 ($M = 3.78$, $SD = 1.27$). In the Average performance condition, scores were once again significantly higher at Time 1 ($M = 5.01$, $SD = 1.05$) than at Time 2 ($M = 4.62$, $SD = 1.36$), $p < .001$; however, scores then significantly decreased at Time 3 ($M = 4.34$, $SD = 1.56$) compared to Time 2, $p = .002$. Finally, in the Very Good performance condition, scores at Time 2 ($M = 5.45$, $SD = 1.09$) were significantly higher than at Time 1 ($M = 5.00$, $SD = 1.12$), $p < .001$, and scores then significantly decreased at Time 3 ($M = 4.92$, $SD = 1.30$), $p < .001$, to similar levels as at Time 1, $p = .481$. In other words, perceived engineering

competence varied across time in accordance with the manipulated performance information. The ANOVA also indicated a Target Gender X Performance interaction, $F(2, 313) = 3.486$, $p = .032$, $\eta_p^2 = .022$, which was not moderated by time, $p = .778$. Engineering competence means (collapsed across time) by Target Gender and Performance appear in Table 5. In the Poor condition, the female was rated significantly more competent than the male target, $p = .006$, but there were no significant target gender differences in the Average, $p = .641$, or Very Good, $p = .618$, conditions.

For the *White Sample*, the Target Gender X Performance interaction (collapsed across time) was also significant, $F(2, 243) = 3.554$, $p = .030$, $\eta_p^2 = .028$ (not moderated by time, $p = .904$), although the pattern was slightly different. The male target was judged more competent than the female target in the Average condition, $p = .045$, but there were no significant target gender differences in the Poor, $p = .081$, or Very Good, $p = .907$, conditions (see Table 5).

Competence. A 2 (Target Gender: Female, Male) x 3 (Performance: Poor, Average, Very Good) x 3 (Time) repeated measures ANOVA was conducted, with Time as a within-subjects factor, for general competence ratings at all three time points. There was a main effect of Time, $F(2, 626) = 74.478$, $p < .001$, $\eta_p^2 = 0.192$, a main effect of Performance, $F(2, 313) = 14.325$, $p < .001$, $\eta_p^2 = .084$, and a significant interaction between Time and Performance, $F(4, 626) = 31.163$, $p < .001$, $\eta_p^2 = 0.166$. Within the Poor performance condition, competence ratings were significantly higher, $p < .001$, at Time 1 ($M = 5.20$, $SD = 1.15$), before performance information was known, compared to Time 2 ($M = 3.91$, $SD = 1.32$), and then increased non-significantly at Time 3 ($M = 3.96$, $SD = 1.19$), $p = .553$. In the Average performance condition, competence ratings were significantly higher, $p < .001$, at Time 1 ($M = 5.32$, $SD = 1.14$) compared to Time 2 ($M = 4.81$, $SD = 1.37$). Competence ratings then significantly dropped again at Time 3 ($M =$

4.51, *SD* = 1.54), $p < .001$. In the Very Good performance condition, the pattern was slightly different: competence ratings at Time 2 (*M* = 5.46, *SD* = 1.09) were significantly higher than at Time 1 (*M* = 5.07, *SD* = 1.10), $p < .001$, and then they significantly dropped at Time 3 (*M* = 4.86, *SD* = 1.29), $p < .001$. The difference between Time 1 and Time 3 was not significant, $p = .096$. No effects involving target gender were significant, *p*s > .874.

Assertiveness. Once again, a 2 (Target Gender: Female, Male) x 3 (Performance: Poor, Average, Very Good) x 3 (Time) repeated measures ANOVA was conducted on perceived assertiveness, with Time as a within-subjects factor. There was a main effect of Time, $F(2, 626) = 38.533$, $p < 0.001$, $\eta_p^2 = 0.110$, as well as a significant Time X Performance interaction, $F(4, 626) = 66.844$, $p < 0.001$, $\eta_p^2 = 0.299$. Within the Poor performance condition, assertiveness ratings were significantly higher at Time 1 (*M* = 4.43, *SD* = 1.14) than at Time 2 (*M* = 2.69, *SD* = 1.22), and then significantly increased again at Time 3 (*M* = 3.05, *SD* = 1.28), all *p*s < .001. In the Average condition, assertiveness scores were highest at Time 1 (*M* = 4.54, *SD* = 1.20), significantly decreased at Time 2 (*M* = 4.25, *SD* = 1.43), $p = .009$, and non-significantly decreased again at Time 3 (*M* = 4.16, *SD* = 1.48), $p = .293$. In the Very Good performance condition, assertiveness scores were lowest at Time 1 (*M* = 4.14, *SD* = 1.07), significantly increased at Time 2 (*M* = 4.73, *SD* = 1.24), $p < .001$, and then significantly decreased at Time 3 (*M* = 4.56, *SD* = 1.33), $p = .036$. Additionally, the interaction between Target Gender and Performance was significant, $F(2, 313) = 6.10$, $p < .003$, $\eta_p^2 = .038$, and was not moderated by time, $p = .725$. Assertiveness means by Target Gender and Performance (collapsed across time) appear in Table 6. In the Poor condition, the female target was judged more assertive than the male target, $p = .003$, no gender difference emerged in the Average condition, $p = .527$, and in

the Very Good condition, the male target was judged non-significantly more assertive than the female target, $p = .089$.

In the *White Sample*, the Target Gender X Performance interaction (not moderated by time, $p = .759$) was again significant, $F(2, 243) = 5.354$, $p = .005$, $\eta_p^2 = .042$; see Table 6 for means. The male target was judged significantly more assertive than the female target in the Very Good condition, $p = .044$ (compared to a non-significant comparison in the full sample), but the female target was rated significantly more assertive than the male target in the Poor condition, $p = .015$. There was no significant gender differences in the Average condition, $p = .419$.

Warmth. In a comparable analysis of perceived warmth, the main effect of performance, $F(2, 313) = 7.265$, $p < .001$, $\eta_p^2 = .044$, and the Time X Performance interaction, $F(4, 626) = 20.078$, $p < .001$, $\eta_p^2 = .114$, were again significant. Within the Poor performance condition, warmth ratings were significantly higher at Time 1 ($M = 4.92$, $SD = 0.88$) than at Time 2 ($M = 4.27$, $SD = 1.40$), $p < .001$, and then increased non-significantly again at Time 3 ($M = 4.38$, $SD = 1.25$), $p = .167$. Within the Average condition, no differences between time points were significant, all ps $> .05$. Within the Very Good performance condition, the opposite pattern of the Poor condition appeared: warmth ratings were significantly higher at Time 2 ($M = 5.21$, $SD = 1.14$) than at Time 1 ($M = 4.73$, $SD = 1.06$), $p < .001$, but then were non-significantly lowered again at Time 3, $p = .118$, although not to the original level of Time 1. The Time X Target Gender interaction was also significant, $F(2, 626) = 3.503$, $p = .031$, $\eta_p^2 = 0.011$.[3] However, for

[3] This interaction was not significant for the *White Sample, p* = .434.

this interaction, the Target Gender effect was never significant at any time point. At Time 1, the male target (M = 4.95, SD = 0.95) was judged somewhat warmer than the female target (M = 4.76, SD = 1.11), p = .116. At Time 2, the male target (M = 4.84, SD = 1.33) and female target (M = 4.83, SD = 1.40) were judged equally, and at Time 3, there was a non-significant, p = .428, reversal where the female target (M = 4.90, SD = 1.18) was judged warmer than the male target (M = 4.78, SD = 1.35). The three-way Time X Target Gender X Performance interaction was not significant, $F(4, 626) = .185$, $p = .946$, $\eta_p^2 = .001$.

Criteria Ranking

A 2 (Target Gender: Female, Male) X 3 (Performance: Poor, Average, Very Good) ANOVA was performed for ratings of the importance of each criterion used for hiring decisions. Other than main effects of performance for the importance of GPA and Interview Quality (in the full sample), and Writing Sample Quality (in the White sample only), no other significant effects emerged. A full report of these results can be found in Appendix E.

Hire Full-time

Among the 152 participants who indicated they would hire the applicant for the internship, I examined the effects of target gender and performance on intention to hire full-time. A chi-squared test of independence showed no interaction between target gender and performance for the hiring decision, $\chi^2(2, N = 152) = 1.085$, $p = 0.581$. The gender main effect was also non-significant, $\chi^2(1, N = 152) = 1.898$, $p = 0.168$, as was the main effect of performance, $\chi^2(2, N = 152) = 4.015$, $p = 0.134$.

Moderation Analyses

Moderation analyses were conducted using PROCESS Macro v3.4 for SAS (Hayes, 2022). For each moderator (structural sexism, attitudes towards affirmative action, and modern

sexism), 18 regressions (based on 6 dependent variables: hiring decision, objective ratings, and engineering competence, competence, assertiveness, and warmth ratings at Time 3) were run with Target Gender, Performance (dummy-coded), the moderator, and all interactions. Across all analyses, there were no significant main effects of the moderator, and only one significant three-way interaction on warmth ratings with pro-Affirmative Action attitudes as the moderator. Rather interpreting this singular significant effect, I conclude that there is no evidence of moderation by the structural sexism, modern sexism, or affirmative action constructs.

Although I did not anticipate effects of participant gender, I also re-ran the ANOVAs described above (for the six dependent variables) with participant gender as an additional predictor. Other than a main effect of participant gender on hiring, $F(1, 296) = 6.31$, $p = .0126$, $\eta_p^2 = .02$, women recommended hiring at a higher rate than men, 55% vs 41% of the time), all main effects and interaction effects involving participant gender were nonsignificant, ps > .12.

Discussion

Results from Study 1 generally did not support my hypotheses. I predicted that hiring rates and objective ratings would favor the male candidate (particularly when performance was average), even while the female candidate was rated subjectively more competent than the male candidate. Results in this direction would have represented a shifting standards pattern. Instead, male and female applicants were hired at equal rates. Across outcome variables, however, participants were responsive to performance information that changed over time, as predicted. Applicants who performed better were generally rated as subjectively and objectively better—and hired more often—than applicants who performed worse, but applicant gender did not moderate these effects. Additionally, none of the other hypothesized moderators (structural sexism, modern sexism, or attitudes towards affirmative action) moderated any of the effects.

Compared to the full sample, when only White participants were included in analyses, a few findings were in line with predictions. I expected that gender bias would be strongest in the Average condition. Although this was not the case for hiring decisions, this pattern did emerge on the standardized objective index: Average-performing male targets were rated higher (i.e., given a higher salary, thought to complete more projects, and have a higher GRE score) than average-performing female targets. Similarly, for ratings of engineering competence, collapsed across time, males were judged significantly more competent than females in the Average condition only. For assertiveness, it was only in the Very Good condition that the male target was judged more assertive than the female target (significantly in the White sample), and in the Poor condition, the female target was judged more assertive than the male target.

Overall, while a few rating patterns emerged in the hypothesized direction, the results did not support the main hypotheses regarding hiring decisions and objective ratings. Judgments across time were responsive to the actual performance information, but not in a manner that differed for men and women (cf, Heilman et al., 2019). Study 1 examined target gender only, and perhaps my mostly null findings are a result of gender stereotypes being less salient in current times. I will return to this point later in the paper. In Study 2, I focused instead on racial stereotypes.

Study 2

Study 2 was conducted to examine the same effects as in Study 1, but with a focus on target race, rather than gender. Black and White *female* applicants were used for three reasons: 1) recent previous research I have conducted has shown a pattern of generalized negative attitudes towards White men that disrupt typical race effects, 2) the research literature in general is biased toward studying race effects using only male targets (see Sesko & Biernat, 2011), and 3) in a

move toward understanding intersectionality, I wanted to examine how Black women are perceived in the hiring process. By using Black and While female applicants, I could avoid any of the generalized negativity towards White men while examining the role of race, but also diving deeper into perceptions of Black women in the workplace.

Method

Participants

Participants were 351 U.S. adults, 18+ years-old, recruited from Prolific and balanced based on gender. All participants successfully passed the three attention checks. Nine participants were removed for failing manipulation checks: 5 for incorrectly remembering the target's race, and 4 for incorrectly remembering the target's gender. The final sample consisted of $N = 342$ participants ($M_{age} = 37.7$ years, $SD_{age} = 13.1$ years). One hundred and sixty-nine participants identified as male, 167 as female, 4 as gender non-binary, 2 as trans, and 1 indicated another gender identity. The sample was 67.5% White ($N = 231$), with 32 participants identifying as Black/African American, 12 as Latinx, 27 as Asian/Pacific Islander, 1 as Native American, 2 as Middle Eastern, and 37 as mixed race or "other."

Procedure

All procedures in Study 2 were the same as Study 1. Participants viewed a White female or a Black female applicant, creating a 2 (Race) x 3 (Interview performance) factorial design, again with Time of Evaluation (3) as a within-subjects factor.

Materials

The materials were the kept same as Study 1, except for the target's name and photos. Cassie W****** was chosen as the target's name because previous research indicates that slightly over sixty percent of people consider the name "Cassie" to be White (Gaddis, 2017).

Similar to Study 1, the goal of the name is to originally indicate Whiteness to examine how race information impacts the interpretation of performance information, especially when the race information is *surprising* or *incongruent* with the job context. However, the name does have to be a realistic name for a Black female candidate as well, and since Cassie is only perceived as White approximately sixty percent of the time, this seemed a reasonable option. In pretesting the materials, 86% of participants thought that the name was a White name. Photos were chosen to indicate race, rather than gender, and were matched on age, attractiveness, and racial prototypicality (Ma et al., 2015).

Measures

All dependent measures remained the same as in Study 1, while substituting the White male and female applicants for the Black and White female applicants.

Potential Moderators

Bayesian Racism Scale. Participants completed the Bayesian Racism Scale (BRS). Bayesian racism is the belief that using racial stereotypes to discriminate against individuals is rational (Uhlmann et al., 2010). Some items in the measure include "*If you want to make accurate predictions, you should use information about a person's ethnic group when deciding if they will perform well*" and "*When the only thing you know about someone is their race, it makes sense to use your knowledge of their racial group to form an impression of them*," $\alpha = .74$. The measure has been found to be positively correlated with Social Dominance Orientation (SDO; Sidanius & Pratto, 1999) and is significantly associated with negative attitudes about Black, Hispanic, and Arab (but not White) Americans (Uhlmann et al., 2010).

Measure of belief in and acknowledgement of structural racism. Participants completed a measure of their belief in structural racism. Early theories of racism conceptualize it

as an interpersonal process due to individual prejudiced values (Allport, 1954), and this is where much of social psychological research has focused its attention. The role of institutional and structural racism in the maintenance of discriminatory practices today has been less frequently acknowledged or studied in the psychological literature. Recently, Rucker and Richeson (2021) have argued that recognition of the role of structural racism in the American criminal justice system is inherently threatening to the existence of the system itself, and as a result, individuals are motivated to continue to conceptualize racism as interpersonal. In 2016, 70% of Whites said that "discrimination based on the prejudice of individuals is the bigger problem when it comes to discrimination against black people in the U.S. today" (as compared to discrimination "built into laws and institutions") compared to only 48% of Black people (Pew Research Center, 2016). This difference in the conceptualization of racism as interpersonal versus structural may also impact potentially prejudicial evaluations of Black versus White candidates in hiring decisions. On a Likert-type scale from 1 (Interpersonal) to 8 (Structural/Institutional), participants indicated the extent to which they believe that racism is more of an interpersonal versus structural/institutional issue.

Demographics. Participants were also asked a series of demographic questions that could potentially moderate evaluations and hiring decisions, including participant race, gender, and political identification/ideology.

Results

To allow for direct comparisons between Study 1 and Study 2, analyses for both the full sample and the White sample were conducted. The White sample consisted of $N = 238$ participants, and as in Study 1, participants who failed the manipulation check were retained in

order to maximize the size of the White sample. The White sample findings will only be noted when they differed from those using the full sample.

Manipulation check on quality of the materials

The materials were perceived as intended. The resume ($M = 4.608$, $SD = 1.181$) and writing sample ($M = 4.000$, $SD = 1.520$) were both perceived as average (on a 7-point scale). Each interview performance was also perceived as intended ($M_{Poor} = 3.051$, $SD_{Poor} = 1.345$; $M_{Avg} = 4.301$, $SD_{Avg} = 1.394$; $M_{VGood} = 5.830$, $SD_{VGood} = 1.064$). A one-way ANOVA indicated a significant main effect of performance, $F(2, 339) = 135.68$, $p < .001$, $\eta_p^2 = 0.445$, and pairwise comparisons using the LSD procedure indicated that each interview performance condition significantly differed from the others, all ps $< .001$.

Hiring decision

A 2 (Target Race: Black, White) x 3 (Performance: Poor, Average, Very Good) analysis was conducted to determine whether hiring rates differed for the Black or White candidates based on their performance in the interview. Table 7 illustrates the proportion of Black and White targets hired in each performance condition for both the full and White samples. A chi-squared test of independence indicated a significant main effect of performance, $\chi^2(2, N = 342) = 46.516$, $p < .001$, but no main effect of target race, $\chi^2(1, N = 342) = 0.829$, $p < 0.363$, and no interaction, $\chi^2(2, N = 342) = 0.408$, $p = 0.816$. Pairwise comparisons using the LSD method indicated that hiring rates in each condition were all significantly different from each other, all ps $< .001$. As in Study 1, my hypothesis that the White female applicant would be hired at significantly higher rates than the Black female applicant was not supported.

Objective ratings

Participants again assigned the applicant a salary amount, indicated the number of projects they thought the applicant would complete during the internship, and estimated the applicant's projected GRE score. Table 8 shows the means, standard deviations, and intercorrelations for each of these variables. Each of these variables was then standardized and combined into a mean objective rating score ($\alpha = 0.75$); the means and standard deviations for the objective rating scores for each group can be seen in Table 9. The main effect of performance was significant, $F(2, 336) = 7.083$, $p = 0.001$, $\eta_p^2 = 0.040$, but there was no main effect of target race, $F(1, 336) = 1.021$, $p > .05$, $\eta_p^2 = 0.003$. Pairwise comparisons using the LSD method indicated that objective ratings were significantly different between the Average ($M = -0.077$, $SD = 0.781$) and Very Good ($M = 0.233$, $SD = 0.772$) conditions, $p = .004$, as well as the Poor ($M = -0.143$, $SD = 0.848$) and Very Good conditions, $p < .001$. The difference between the Poor and Average conditions was not significant, $p = .537$. Although the Target Race X Performance interaction was significant in Study 1, the two-way interaction in Study 2 was not significant, $F(2, 336) = 0.927$, $p = .397$, $\eta_p^2 = 0.005$.

Subjective judgments over time

Ratings of the applicant's competence, likelihood of success, and effectiveness in the Engineering Internship were combined into a general engineering competence score (αs = 0.95–0.97). General competence, assertiveness, and warmth were analyzed individually. In each case, I conducted a 2 (Target Race: Black, White) X 3 (Performance: Poor, Average, Very Good) X 3 (Time) repeated measures ANOVA, with Time as a within-subjects factor.

Engineering Competence. There was a significant main effect of Time, $F(2, 672) = 63.815$, $p < .001$, $\eta_p^2 = 0.160$, Performance, $F(2, 336) = 22.957$, $p < .001$, $\eta_p^2 = 0.120$, and a

Time X Performance interaction, $F(4, 672) = 54.213$, $p < .001$, $\eta_p^2 = 0.244$. In the Poor condition, engineering competence ratings were significantly higher at Time 1 ($M = 5.140$, $SD = 1.081$) than at either Time 2 ($M = 3.840$, $SD = 1.354$) or Time 3 ($M = 3.798$, $SD = 1.503$), ps < .001. Engineering competence ratings decreased again at Time 3, but non-significantly, $p = .642$. In the Average condition, ratings were the highest at Time 1 ($M = 5.157$, $SD = 1.180$), decreased significantly at Time 2 ($M = 4.761$, $SD = 1.260$), $p < .001$, and decreased significantly again at Time 3 ($M = 4.435$, $SD = 1.428$), $p < .001$. In the Very Good condition, scores were highest at Time 2 ($M = 5.690$, $SD = 1.059$) and were significantly different from ratings at Time 1 ($M = 4.988$, $SD = 1.128$), $p < .001$, or Time 3 ($M = 5.083$, $SD = 1.286$), $p < .001$. Although ratings at Time 3 were higher than at Time 1, this difference was not significant, p = .356. No effects involving target race were significant, ps > .51.

Competence. Main effects of Time, $F(2, 672) = 72.560$, $p < .001$, $\eta_p^2 = 0.178$, Performance, F(2, 336) = 16.687, $p < .001$, $\eta_p^2 = 0.090$, and the Time X Performance interaction, $F(4, 672) = 34.208$, $p < .001$, $\eta_p^2 = 0.169$, were significant. In the Poor performance condition, competence scores were highest at Time 1 ($M = 5.462$, $SD = 1.095$), significantly decreased at Time 2 ($M = 4.170$, $SD = 1.373$), $p < .001$, and non-significantly increased at Time 3 ($M = 4.188$, $SD = 1.520$), $p = .853$. In the Average condition, scores were once again highest at Time 1 ($M = 5.395$, $SD = 1.048$), then significantly decreased at Time 2 ($M = 5.008$, $SD = 1.271$), $p < .001$, but then significantly decreased again at Time 3 ($M = 4.573$, $SD = 1.400$), $p < .001$. Finally, in the Very Good condition, competence ratings were highest at Time 2 ($M = 5.762$, $SD = 1.082$), a significant increase from Time 1 ($M = 5.330$, $SD = 1.165$), $p < .001$. Scores then significant decreased from Time 2 to Time 3 ($M = 5.188$, $SD = 1.241$), $p < .001$, back to levels similar to Time 1, $p = .279$. No effects involving target race were significant, ps > .29.

Assertiveness. Once again, there was a main effect of Time, $F(2, 672) = 17.689$, $p < .001$, $\eta_p^2 = 0.050$, Performance, $F(2, 336) = 33.047$, $p < .001$, $\eta_p^2 = 0.164$, and their two-way interaction, $F(4, 672) = 53.971$, $p < .001$, $\eta_p^2 = 0.243$. In the Poor performance condition, assertiveness ratings were highest at Time 1 ($M = 4.376$, $SD = 1.291$), decreased significantly at Time 2 ($M = 2.889$, $SD = 1.363$), $p < .001$, and increased significantly at Time 3 ($M = 3.265$, $SD = 1.499$), but not nearly to the original level of Time 1, $p < .001$. In the Average condition, the only comparisons that were significant were between Time 1 and Time 3, $p = .032$. Scores were highest at Time 1 ($M = 4.266$, $SD = 1.296$), slightly decreased at Time 2 ($M = 4.044$, $SD = 1.442$), $p = .061$, and slightly decreased again at Time 3 ($M = 3.999$, $SD = 1.464$), $p = .577$. In the Very Good condition, scores were lowest at Time 1 ($M = 4.379$, $SD = 1.172$), increased significantly at Time 2 ($M = 5.049$, $SD = 1.269$), $p < .001$, and then decreased significantly at Time 3 ($M = 4.874$, $SD = 1.238$), $p = .031$. No effects involving target race were significant, $ps > .57$.

Warmth. The main effect of Performance, $F(2, 336) = 12.627$, $p < .001$, $\eta_p^2 = 0.070$, and the Time X Performance interaction were again significant, $F(4, 672) = 25.659$, $p < .001$, $\eta_p^2 = 0.132$. In the Poor performance condition, warmth ratings were significantly higher at Time 1 ($M = 5.077$, $SD = 1.068$) than at Time 2 ($M = 4.485$, $SD = 1.324$) or Time 3 ($M = 4.580$, $SD = 1.341$), $ps < .001$. The difference in ratings between Times 2 and 3 was not significant, $p = .140$. In the Average condition, there were no statistically significant differences between time points, all $ps > .05$. In the Very Good condition, warmth ratings were lowest at Time 1 ($M = 4.937$, $SD = 1.034$), significantly increased at Time 2 ($M = 5.634$, $SD = 1.108$), $p < .001$, and then non-significantly decreased at Time 3 ($M = 5.535$, $SD = 1.115$), $p = .127$.

In a pattern different from the other ratings in Study 2, there was both and a main effect of Target Race, $F(1, 336) = 4.565$, $p = .033$, $\eta_p^2 = 0.013$, and a significant Time X Target Race interaction, $F(2, 672) = 6.899$, $p = .001$, $\eta_p^2 = 0.020$. The means and standard deviations by Target Race and Time can be seen in Table 10. At Time 1, both targets were rated equally warm, $p = .951$. At Times 2 and 3, the Black target was rated significantly warmer than the White target, $p = .003$ and $p = .017$, respectively. In the *White Sample*, neither the main effect of race, $F(2, 232) = 2.130$, $p = .146$, $\eta_p^2 = 0.009$, nor the interaction, $F(2, 464) = 2.867$, $p = .058$, $\eta_p^2 = 0.012$, were significant, though the pattern was similar.

Criteria Ranking

A 2 (Target Race: Black, White) X 3 (Performance: Poor, Average, Very Good) ANOVA was performed for each criteria to determine whether applicant race or performance shifted the top criteria used for their hiring decision. Analyses revealed no meaningful significant findings, but a full report can be found in Appendix E.

Hire Full-time

Among the 181 participants who indicated they would hire the applicant for the internship, I examined the effects of Target Race and Performance on intention to hire full-time. A chi-squared test of independence showed no interaction between target gender and performance for the hiring decision, $\chi^2(2, N = 181) = 0.479$, $p = 0.787$. The target race main effect was also non-significant, $\chi^2(1, N = 181) = 0.536$, $p = 0.464$, as was the main effect of performance, $\chi^2(2, N = 181) = 2.053$, $p = 0.358$.

Moderation Analyses

As in Study 1, for each dependent variable (hiring, objective judgment, and the four subjective trait judgments at time 3), I conducted additional regression analyses to test for

potential moderating effects[4]. In this study, Bayesian racism and the Structural racism item were examined as potential moderators. Unlike Study 1, there was considerable evidence that Bayesian racism in particular (and Structural racism to a less consistent extent), moderated the effects of target race and performance on judgments. I used PROCESS macros (Hayes, 2022) to regress each dependent variable on target race (coded -1=White, 1=Black), two dummy variables to reflect the three performance conditions, with the average condition as the referent (D1 therefore compares the poor and average conditions, D2 compares the very good and average conditions), the moderator (Bayesian racism or Structural racism), and all interactions. Table 11 summarizes the three-way interactions that emerged for each analysis. Because the patterns of results were similar across outcomes, I also combined the four subjective ratings at Time 3 (the engineering competence index, competence, assertiveness, and warmth) into a single subjective index. Regression results for this index appear in the last row of the table and are discussed below.

Bayesian Racism. As can be seen in Table 11, 3-way interactions among Target Race, Performance, and Bayesian Racism emerged for every dependent variable. Figure 1 depicts the interaction for *proportion of hiring* decisions, with levels of Bayesian racism depicted at low (16^{th} percentile), moderate (50^{th} percentile) and high (84^{th} percentile) levels. The Target Race X Bayesian Racism interaction was significant only in the Average performance condition, χ^2 (1, N = 342) = 4.786, p = .0289 (ps in other conditions > .17). The simple effect of race was

[4] Although this study focused on race rather than gender, I also examined whether gender of participant moderated any findings. The only findings of any interest were that for objective judgments and judgments of warmth, Target Race X Participant Gender interactions were significant, ps < .05 with women rating Black targets more favorably than White targets, and men showing no race effect.

(marginally) statistically significant only when performance was average, among those *low* in Bayesian racism, *effect* = -0.627, *SE* = 0.33, *z* = -1.9277, *p* = .0539 (all other race effects *p* > .10). The average-performing White woman was *more* likely to be hired than the average-performing Black woman among those who did *not* endorse Bayesian racism. Further, with increasing Bayesian racism, the average-performing Black target was *more likely* to be hired, *z* = 2.2628, *p* = .0237, but hiring of the White target did not vary, *p* > .48.

Figure 2 depicts the comparable interaction for *objective judgments*. The Target Race X Bayesian Racism interaction was significant in the Average performance condition, $F(1, 330) = 4.0624$, $p = .0447$, and marginally significant in the Poor performance condition, $F(1, 330) = 3.7304$, $p = .0543$ ($p = .1274$ in the Very Good condition). For this outcome, the simple effect of race was statistically significant when performance was Very Good, among those *low* in Bayesian racism, *effect* = .2655, *SE* = .12, $t(330) = 2.2276$, $p = .0266$, and marginally significant when performance was Very Good among those *moderate* in Bayesian racism, *effect* = .148, *SE* = .18, $t(330) = 1.9329$, $p = .0541$. The strong-performing Black woman was rated objectively *better* than the strong-performing White woman among those low and moderate in Bayesian racism. In the average-performing condition, however, the pattern was reversed (the average-performing Black woman was rated objectively *worse* than the average-performing White woman, among those both low and moderate in Bayesian racism, but the differences did not reach significance (*p*s > .0753)). Further analyses indicated that with increasing levels of Bayesian Racism, the average-performing Black target was judged objectively *more* favorably, $t(330) = 1.8998$, $p = .0583$ ($p = .3607$ for the average White target), and the poor-performing Black target was judged *less* objectively favorably, $t(330) = -2.7637$, $p = .0060$ ($p = .90$ for the poor-performing White target).

Figure 3 presents the comparable interaction for the overall subjective index (a composite of all the subjective ratings at Time 3). The Target Race X Bayesian Racism interaction was significant in both the Average performance condition, $F(1, 330) = 5.7150$, $p = .0174$, and the Poor performance condition, $F(1, 330) = 5.9697$, $p = .0151$ ($p = .5407$ in the Very Good condition). The simple effect of race for the subjective index was statistically significant when performance was Poor among those *low* in Bayesian racism, *effect* = .3152, *SE* = .16, $t(330) = 2.0297$, $p = .0432$, and marginally significant when performance was Average among those *high* in Bayesian Racism, *effect* = .3181, *SE* = .16, $t(330) = 1.9493$, $p = .0521$. The poor-performing Black woman was rated subjectively higher than the poor-performing White woman among those low in Bayesian racism, and the average-performing Black woman was also rated subjectively higher than the average-performing White woman among those high in Bayesian racism. In the Average condition, the Black woman was rated subjectively lower when Bayesian Racism was low, but the difference did not reach significance ($p = .0749$). Additional analyses indicated that with increasing levels of Bayesian racism, the average-performing Black woman was rated subjectively *higher,* $t(330) = 2.5142$, $p = .0124$, but judgments of the average-performing White woman did not vary across levels of Bayesian racism. When performance was poor, the Black woman was judged more negatively with increasing levels of Bayesian racism, $t(330) = -1.9300$, $p = .0545$, but judgments of the White woman did not vary ($p = .1311$).

Structural Racism. As can be seen in Table 10, comparable interactions were significant when Structural Racism was the moderator for three of the subjective judgments (engineering competence, competence, and assertiveness), and for the overall Time 3 subjective index, which is depicted in Figure 4. Lower levels of structural racism indicate a belief that racism an individual problem, while higher levels indicate a belief that racism is more of a

structural/institutional problem. The Target Race X Structural Racism interaction was marginally significant in the Average condition, $F(1, 330) = 2.9858$, $p = .0849$, and in the Poor performance condition, $F(1, 330) = 3.7763$, $p = .0528$ (but not in the Very Good condition, $p = .4588$), although simple effects of race were never significant, all ps > .084. However, further analysis indicated that as belief in racism shifted from a focus on the individual to a focus on the structural, the poor-performing Black woman was rated subjectively *higher*, $t(330) = 2.8423$, $p = .0048$ ($p = .7818$ for the poor-performing White woman).

Discussion

In Study 2, I predicted that hiring rates and objective ratings would favor the White candidate (particularly in the average performance condition), even while the Black candidate was rated as subjectively better (more competent, etc.). The results did not support my hypotheses. Again, participants were quite responsive to the performance information, updating their impressions of the candidate over time based on their performance, but these impressions generally did not differ by target race. I also examined all of the outcomes and ratings among participants who identified as White, and unlike Study 1, the results did not differ significantly from that of the full sample.

It was only on judgments of applicant warmth that target race effects emerged in an interaction between time and target gender. At Times 2 and 3 (after performance information was revealed), the Black applicant was rated as significantly warmer than the White applicant. This single finding aligns with a shifting standards pattern, but the lack of race effects on other outcomes reduces any confidence in the interpretation of this finding.

Finally, unlike in Study 1, there were significant effects of the moderators on the dependent variables. Bayesian Racism interacted with race and performance on hiring decisions,

objective ratings, and the subjective ratings, including the total subjective index created for ratings at Time 3. However, the results were in an unexpected pattern. I hypothesized that as endorsement of Bayesian racism (the belief that reliance on stereotypes is a rational way to form judgments about people) increased, so would reliance on stereotypes in making hiring decisions and ratings, especially in the average condition.

While Bayesian racism most consistently interacted with race in the average condition, it was among those who do *not* believe stereotypes should be used to form judgments that the average-performing Black woman was less likely to be hired, and judged less objectively and subjectively positively than the average-performing White woman. The low endorsers of Bayesian racism generally favored the Black woman in the Poor and Very Good performance conditions, as might be expected, but the pattern of anti-Black judgment in the average condition is surprising. Those who highly endorsed the use of stereotypes to make judgments showed the converse pattern: They favored the average Black woman over the average White woman, but favored the Very Good and Poor White woman over the Black woman (though these race effects were rarely statistically significant). Additionally puzzling, as endorsement of Bayesian Racism increased, the average-performing Black woman was hired *more often* and rated objectively and subjectively *more favorably,* while racism had no effect on judgments of the White woman. In a more straightforward pattern, those who believe that racism is more of a structural problem than an individual problem rated the Poor-performing Black woman more favorably than those low in structural racism endorsement.

General Discussion

Previous research suggested that incongruence with a job context and changes in performance over time would result in differential hiring rates for members of historically

marginalized and underrepresented groups (e.g., White women, Black women) compared to members of the dominant group (i.e., White men) (Heilman et al., 2019; Dovidio & Gaertner, 2000; Biernat & Kobrynowicz, 1997; Quillian et al., 2017). The two studies presented here were designed to understand how people update their impressions of a job candidate when performance changes over time, and how revealing gender (Study 1) and race (Study 2) information interacts with that performance information, especially if the gender and/or race of the applicant is incongruent with the job description and field. I approached this question by attempting to replicate a realistic job application process in which there are multiple times for impressions of the candidate to be formed, and potentially changed, based on performance information changing over time. As in a typical hiring scenario, participants made judgments about a candidate after reviewing a resume, then gained more information about performance after reading the transcript of the candidate's interview (where gender and race were revealed), and then reviewed a writing sample as the final piece of the application package. By making the same ratings over multiple time points, this allowed participants (role playing as hiring managers) the ability to take in new information (about performance, gender, and/or race) and incorporate that into the impression of the candidate as a whole, as they would in a real hiring context.

Across two studies, I did not find evidence of gender (Study 1) or race (Study 2) bias in hiring decisions or other objective outcomes, such as salary, perceptions of projects completed in the internship, or projected GRE score. Applicants were hired at equal rates regardless of their gender or race. Participants were, however, very receptive to changes in performance information, as hypothesized. Performance information seemed to outweigh any potential influence of gender or race bias. Although these studies were not designed to test among the

various models of impression formation described in the introduction, the pattern of findings presented here indicates that participants may be relying more on individuating information (i.e., performance information) than on category-based information (i.e., gender or race) to form an overall impression of the candidate. Additionally, patterns of change on the subjective ratings over time indicated that participants may be algebraically combining information at each time point to create an overall impression (Anderson, 1965); however, most of the patterns also conformed to what would be expected based on changes in performance.

One particularly puzzling finding involved the effect of endorsement of Bayesian Racism, or "the belief that it is rational to discriminate against individuals based on stereotypes about their racial group" (Uhlmann et al., 2010, p. 1), on hiring decisions, objective ratings, and subjective ratings in Study 2. I hypothesized that those who believe it is rational to discriminate based on stereotypes would indeed discriminate based on gender (anti-female) and race (anti-Black) stereotypes when making hiring and other decisions, particularly in the average performance condition where ambiguity might necessitate inference. Bayesian Racism has been found to be correlated with negative attitudes towards Black people, as well as with Social Dominance Orientation (Uhlmann et al., 2010; Sidanius & Pratto, 1999), both of which are correlated with discriminatory practices (e.g., Pratto et al., 1994; Rudman & Ashmore, 2007). In Study 2, while endorsement of Bayesian Racism did have the strongest effect in the Average performance condition, it was only among those *low* (or sometimes *moderate*) in Bayesian Racism where there was evidence of anti-Black bias. Conversely, those who were *high* in Bayesian Racism tended to favor the Black target (though significantly so only on the subjective rating index). This pattern is paradoxical, as it is people who *do not* believe in the use of stereotypes who show anti-Black bias when performance is ambiguous (while those who *do*

believe in the use of stereotypes tend to be pro-Black). This effect—opposite from what was expected—is difficult to explain, but it was a consistent pattern, emerging across all the dependent measures. The fact that the Race X Bayesian Racism interaction patterns generally reversed in the Poor and Very Good conditions (in a pattern more consistent with expectations) is worth noting, but does not explain the odd pattern in the Average condition. The distribution of composite scores on the Bayesian Racism Scale (BRS) was highly positively skewed, indicating that most participants do not feel that stereotypes should be used as a basis for discrimination. Thus, those labeled high were not extreme endorsers of the use of stereotypes, and those labeled low may include "real" lows as well as those merely offering the normative, socially desirable response. A next step will be a replication study, to examine the reliability of the Study 2 patterns.

I also hypothesized that participants would shift the criteria they used to make their hiring decision based on target gender/race and performance. When interview performance was poor, I hypothesized that the importance of the interview would be overemphasized for the female/Black applicants (as stereotype confirmation), whereas when interview performance was high, it should be seen as less important for the female/Black applicant (a fluke). The results from both studies revealed no significant differences in criteria ranking and no interactions with target gender or race.

An additional purpose of the study was to generate a potential feeling of "surprise" for the participant when the applicant's gender or race was revealed at the interview portion of the study. The job description and resume were created to conjure in the participants' mind an image of a White man in Study 1 or a White woman in Study 2 (pretesting confirmed this was the case in both studies) so that when the applicant was actually a Black man or Black woman, there

would be a sense of surprise and incongruence (if only unconsciously) with the job description and resume. Although impossible to truly measure participants' surprise, I did ask if anything surprised the participants about the candidate (after they had answered all dependent variables) as an open-ended question. In Study 1, only about a quarter of participants who saw the White female applicant reported being surprised that they were a female (e.g., "I thought Charlie was going to be a guy because of the name," "I was surprised to find out that I was evaluating a female," etc.). One person even went out of their way to say that they were *not* surprised that the applicant was a woman: "Nope. If I'm supposed to be surprised that the big reveal is that Charlie is a female... come on. Not really a shocker…. Am I really expected to have my mind blown that a woman can be an engineer?" This also indicates that this particular participant, at least, was aware of the purpose of the study. Any other surprise was about the applicant's (usually bad) performance (e.g., "Lack of preparedness in his interview"). In Study 2, perhaps unsurprisingly, only a few people commented on the applicant's race (e.g., "She is a Black female. Not what I would expect for an engineering student," "Frankly, I've never met a Black person named Cassie," etc.). Interestingly, a few people were surprised by her gender (e.g., "That she was a woman," "I thought Cassie was a White male," "I am not sure I would say that Cassie being a female surprised me, but generally women seem to be less interested in engineering than men"), supporting the idea that the job is perceived as one primarily for White men. Any other surprises that participants felt were performance related.[5]

---

[5] Of course, it is quite possible that critiques of the Black applicant's performance (e.g., "That she did not speak well verbally," "She was very bad with speaking and writing," etc.) are racially-coded, but there is no way to directly measure this in the study.

Real patterns of change?

One potential, but—I will argue—improbable, explanation for the (lack of) findings is that discriminatory patterns in the real world are changing. Although Quillian and colleagues' 2017 meta-analysis of hiring discrimination rates over time indicates otherwise, the past three or so years (after the murders of George Floyd, Brianna Taylor, Ahmaud Arbery, and other innocent people of color at the hands of police) have seen heightened conversations around race and bias, as people have made increased calls for a systematic reckoning of the role of institutional and structural racism in the U.S (e.g., Chang et al., 2020). As a response, there has been a rise of "cancel culture," which has been both celebrated and created outrage among various groups. Cancel culture has resulted in many controversies, as 49% of Americans (and 59% of Liberal/Lean Democrats) define it as "actions taken to hold others accountable," but 14% of Americans (and 26% of Conservative/Lean Republicans) define it as "censorship of speech or history" and another 12% define it as "mean-spirited actions taken to cause others harm" (Vogels et al., 2021). Whether an individual thinks of cancel culture as accountability or censorship of speech or history, I do believe people are on 'high alert' for being perceived as doing any behavior that could be deemed discriminatory; this may include participants on platforms like Prolific, who, ideally, are not aware of the purpose of the study but may nonetheless have an idea of what the researchers are trying to do (especially if assigned to the female or Black condition of a study). If non-discriminatory patterns like the ones found in this study do represent actual changes in hiring practices, then these findings should be celebrated. Yet, as research continues to show bias against women and people of color (e.g., Quillian et al., 2017; Hardy et al., 2022; Wilson et al., 2021; Ballakrishnen et al., 2019), I remain unconvinced that these deeply embedded biases are not continuing to affect discrimination in the workplace and other domains.

Limitations

There are also potential methodological limitations that could have limited the ability to detect race and gender effects on judgment. The two studies were conducted online using the Prolific crowdsourcing platform with general population participants. As such, the participants most likely to be included are not particularly experienced in making hiring decisions (I did not ask a question about hiring experience). Additionally, because Prolific was developed as a crowdsourcing platform particularly for research, there is concern that participants are not naïve to the aims of social psychological research (Cho & Lumsden, 2019). Cho and Lumsden (2019) showed that within the week they analyzed Prolific data, their most active 5% of participants completed 20% of *all* Prolific responses; this certainly is not ideal, but it is better than Amazon's Mechanical Turk (MTurk), in which the top 5% of MTurk workers completed 40% of available HITs (Human Intelligence Tasks; Cho & Lumsden, 2019; Robinson et al., 2019). Relatedly, Chandler and colleagues (2013) conducted a study among Amazon MTurk workers to assess how they choose the studies/HITs they want to complete and how they share information about them. The authors found that many MTurk workers reported participating in more well-known experimental paradigms (like the "trolley problem," etc.), and, perhaps unsurprisingly, workers who are more productive are more likely to have participated in those tasks. They conclude that the evidence "stands in contrast to earlier claims that MTurk offered a participant pool of naïve participants," and suggest that researchers take precautions when using more well-known paradigms, and, in any case, do what they can to minimize duplicate respondents (Chandler et al., 2013).

Although that research was conducted among MTurk workers, many of the same issues would likely appear for Prolific workers as well. One particular quote from a participant in my

study in part of their response to an open-ended question seemed concerning for potential problems of participant naïvete. The participant stated, “I think the picture was in several other studies (pretty sure the person in it was - stock photo?),” indicating that they have probably been a part of studies that have also utilized photos from the Chicago Face Database (CFD, Ma et al., 2015); however, given the widespread use of the CFD in psychological research, this is not particularly surprising. Prolific does, however, institute what they term “adaptive rate limiting,” in which priority access to studies is given to participants who have recently spent less time taking studies, during times when the ratio of active participants to number of study places is relatively high (Cho & Lumsden, 2019). Prolific also allows researchers to easily exclude participants from participating in multiple of their own studies (which I ensured for the studies presented here), but there is no way to ensure that participants are not participating in similar studies by other researchers.

Additionally, it is also a possibility that the photo used to manipulate gender and race was not salient or strong enough. Participants only saw the photo a single time – when they read the applicant’s interview transcription. While I did ask participants the gender or race of the applicant as a manipulation check, and excluded those who failed, there were quite a few participants who skipped the question. Although there was an “I don’t know” option, it is possible that skipping the question indicates that the participant was unsure of the target’s gender or race, which would be indicative of a failed manipulation. While I do not think this is the case (as demonstrated by participant’s engagement with the applicant’s gender or race in the open-ended question about surprise), it remains a possibility.

Future Directions for Research

Reflecting on these findings (and previous research I have conducted) where there are no meaningful patterns of discrimination based on race or gender, I believe that the questions we ask and the methods we use to study discrimination (particularly in the hiring process) may have to change in order to understand how bias presently operates in the workplace. Because of social desirability bias, fears of being a target of "cancel culture," and a lack of participant naïvete, it may no longer be as feasible to study workplace discrimination through experiments that manipulate race and/or gender of a target resume and examine dichotomous hiring decisions. People of color, women, or other historically marginalized group members clearly also experience discrimination in settings beyond the hiring decision.

In order to continue studying workplace bias through the lens of the "perceiver" (i.e., the one *doing* the discrimination), methods may need to get more creative. Perhaps experimental methods like the ones used here need to be more subtle and less reactive, as clearly some participants were partially aware of the goals of the study. Audit studies (Quillian et al., 2017; Zhao & Biernat, 2017; Milkman et al., 2012) continue to show workplace bias on various dimensions. Additionally, using more realistic settings, more nonexperimental and correlational methods, and real-world data (e.g., company hiring outcomes and employee performance evaluations) may prove more fruitful and impactful.

Another way of moving forward with this line of research is to focus on the point of view of the target of the discrimination, rather than the perceiver. In the case of the two studies presented above, that would mean focusing on the point of view of White and Black women, as compared to White men and White women. By surveying and asking questions of the people being discriminated against, we will get a clearer understanding of the types of discrimination they face in the workplace today. This also aligns more with a decolonial approach to

psychological research. A core component of liberation psychology (Martín-Baró, 1994) is "[privileging] the epistemological position of people in oppressed or marginalized conditions" and "a commitment to treat the perspectives and interests of oppressed and impoverished communities as a primary source of insight about everyday truth and social reality" (Adams et al., 2015; Martín-Baró, 1994). Taking a decolonial and liberation psychological approach to research on discriminatory practices in hiring decisions, and the workplace more broadly, may not only yield more interesting results but will most importantly center the lived experiences of people facing discrimination.

Additionally, it may be important to take a more qualitative approach to some of these questions, as dichotomous answer choices and Likert-type items may not be enough to capture to complexity of workplace discrimination today. Qualitative research can produce "rich, in-depth data" (Povee & Roberts, 2014, p. 32), and many psychologists laud the ability of qualitative research to capture the lived experience of its participants; it can also provide "access to information that is typically inaccessible by other research approaches" (p. 32). Qualitative research, either on its own or in combination with quantitative research, from the perspective of the "target" would likely be a fruitful next step in examining the way discrimination occurs in the workplace today.

Overall, these two studies showed little evidence of bias in the interpretation of performance information over time or on hiring decisions and other key objective outcomes, although some evidence of bias did occur in the predicted average performance condition among a sample of only White participants. Future research is needed to understand whether these changes are reflective of real-world patterns or a product of participant awareness and/or a desire to appear unprejudiced, either to themselves or others. Research using more subtle experimental

methods, real-world data, qualitative data, or that focuses on the point of view of applicants may be a fruitful next step.

Table 1

Samples sizes for each Target Gender and Performance condition in both Study 1 and Study 2, among the full and White samples.

Sample Sizes					
			Performance		
			Poor	Average	Very Good
Study 1	Full Sample (*N*=319)	White Female	50	51	55
		White Male	56	54	53
	White Sample (*N*=249)	White Female	42	39	41
		White Male	45	40	42
Study 2	Full Sample (*N*=342)	Black Female	59	55	53
		White Female	58	58	59
	White Sample (*N*=238)	Black Female	41	40	34
		White Female	43	42	38

Table 2

Mean proportion (and standard deviation) of applicants hired by Target Gender and Performance, Study 1, N=319.

	Performance		
Target Gender	Poor	Average	Very Good
White Female	0.32 (0.47)	0.53 (0.50)	0.69 (0.47)
White Male	0.23 (0.43)	0.50 (0.50)	0.58 (0.50)

Note. "Not hired" was coded as 0 and "Hired" was coded as 1, so means indicate proportion hired.

Table 3

Means, standard deviations, and correlations of unstandardized variables with confidence intervals, Study 1, N = 319.

Variable	*M (SD)*	1	2
1. Salary (in dollars)	5738.69 (1807.40)		
2. Projects completed	3.32 (1.14)	.41** [.32, .50]	
3. Projected GRE score	299.11 (16.66)	.46** [.37, .55]	.52** [.43, .59]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Table 4

Means and (standard deviations) for standardized objective ratings as a function of a 2(Target Gender) X 3(Performance), Study 1, Full Sample N = 319, White Sample N = 249.

		Performance		
		Poor	Average	Very Good
	Target Gender	*M (SD)*	*M (SD)*	*M (SD)*
Full Sample	White Female	-0.01 (0.84)	0.06 (0.88)	0.11 (0.70)
	White Male	-0.41 (0.63)	0.05 (0.86)	0.23 (0.77)
White Sample	White Female	-0.13 (0.89)	-0.14 (0.80)	0.16 (0.73)
	White Male	-0.41 (0.67)	0.29 (0.75)	0.28 (0.70)

Note. M and *SD* represent mean and standard deviation, respectively.

Table 5

Means (and standard deviation) of engineering competence ratings, collapsed across time, by Target Gender and Performance, Study 1, Full Sample N=319, White Sample N=249.

		Performance		
		Poor	Average	Very Good
	Target Gender	*M (SD)*	*M (SD)*	*M (SD)*
Full Sample	White Female	4.84 (1.05)	4.61 (0.95)	5.07 (0.98)
	White Male	3.90 (1.09)	4.71 (1.35)	5.18 (1.10)
White Sample	White Female	4.26 (1.23)	4.34 (0.84)	5.18 (0.93)
	White Male	3.86 (1.14)	4.83 (1.13)	5.21 (1.09)

Table 6

Means (and standard deviation) of assertiveness ratings, collapsed across time, by Target Gender and Performance, Study 1, Full Sample N=319, White Sample N=249.

		Performance		
		Poor	Average	Very Good
Sample	Target Gender	*M (SD)*	*M (SD)*	*M (SD)*
Full Sample	White Female	3.71 (1.10)	4.25 (1.08)	4.30 (1.10)
	White Male	3.07 (0.82)	4.38 (1.33)	4.65 (1.04)
White Sample	White Female	3.61 (1.15)	4.13 (1.00)	4.20 (1.10)
	White Male	3.04 (0.85)	4.33 (1.37)	4.68 (0.96)

Table 7

Mean proportion (and standard deviation) of applicants hired by Target Race and Performance, Study 2, N = 342.

		Performance		
		Poor	Average	Very Good
Sample	Target Race	*M (SD)*	*M (SD)*	*M (SD)*
Full Sample	Black Female	0.31 (0.46)	0.51 (0.50)	0.72 (0.45)
	White Female	0.33 (0.47)	0.53 (0.50)	0.80 (0.41)
White Sample	Black Female	0.27 (0.45)	0.53 (0.51)	0.68 (0.47)
	White Female	0.37 (0.49)	0.57 (0.50)	0.82 (0.39)

Note. M and *SD* represent mean and standard deviation, respectively.

Table 8

Means, standard deviations, and correlations of unstandardized variables with confidence intervals, Study 2, N = 342.

Variable	*M (SD)*	1	2
1. Salary (in dollars)	5594.83 (1878.64)		
2. Projects completed	3.34 (1.19)	.47** [.39, .55]	
3. Projective GRE score	299.31 (17.31)	.54** [.46, .61]	.48** [.39, .56]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Table 9

Means and (standard deviations) for standardized objective rating variable as a function of a 2(Target Race) X 3(Performance) design, Study 2, Full Sample N = 342, White Sample N = 238.

		Performance		
		Poor	Average	Very Good
	Target Race	*M (SD)*	*M (SD)*	*M (SD)*
Full Sample	Black Female	-0.11 (0.80)	-0.10 (0.80)	0.35 (0.77)
	White Female	-0.17 (0.90)	-0.06 (0.77)	0.11 (0.76)
White Sample	Black Female	-0.21 (0.83)	-0.02 (0.76)	0.38 (0.77)
	White Female	-0.22 (0.89)	-0.05 (0.84)	0.21 (0.71)

Note. M and *SD* represent mean and standard deviation, respectively.

Table 10

Means and (standard deviations) for warmth ratings as a function of a 2(Target Race) X 3(Time) design, Study 2, Full Sample N = 342, White Sample N = 238.

		Time 1	Time 2	Time 3
	Target Race	*M (SD)*	*M (SD)*	*M (SD)*
Full Sample	Black Female	5.02 (1.04)	5.30 (1.24)	5.26 (1.27)
	White Female	5.02 (1.12)	4.91 (1.30)	4.94 (1.29)
White Sample	Black Female	5.03 (1.06)	5.25 (1.23)	5.17 (1.30)
	White Female	4.98(1.15)	4.89 (1.34)	4.99 (1.31)

Note. M and *SD* represent mean and standard deviation, respectively.

Table 11

Summary of Moderation Analyses (3-way interaction coefficients, SEs, and p-values) on all Dependent Measures, Study 2.

	Moderator in Regression and Interaction Term (coefficient, SE, p-value)			
	Bayesian Racism		Structural Racism	
Dependent Variable	TR X PerfD1 (Avg v Poor) X BR	TR X PerfD2 (Avg v VG) X BR	TR X PerfD1 (Avg v Poor) X SR	TR X PerfD2 (Avg v VG) X SR
Hire	-.434 (.23), .0602	-.652 (.27), .0142	.184 (.13), .1446	.167 (.13), .1904
Objective Judgment	-.239 (.09), .0056	-.226 (.09), .0131	.089 (.05), .0579	.011 (.05), .8117
Eng Comp	-.451 (.15), .0032	-.317 (.16), .0498	.215 (.08), .0095	.063 (.08), .4389
Competent	-.510 (.15), .0007	-.309 (.16), .0517	.255 (.08), .0018	.089 (.08), .2617
Assertive	-.314 (.15), .0396	-.1343 (.16), .4051	.162 (.08), .0497	.112 (.08), .1659
Warm	-.375 (.13), .0042	-.306 (.14), .0270	.049 (.07), .4917	-.082 (.07), .2390
Subj Total	-.413 (.12), .0007	-.267 (.13), .0379	.170 (.07), .0098	.046 (.06), .4797

Note. TR = Target Race, Perf = Performance, D1=dummy variable 1, D2=dummy variable 2, BR = Bayesian Racism, SR = Structural Racism, Avg = Average, VG = Very Good.

Figure 1

Target Race X Bayesian Racism interaction at each level of performance for proportion of targets hired (significant in the Average performance condition), Study 2, N=342.

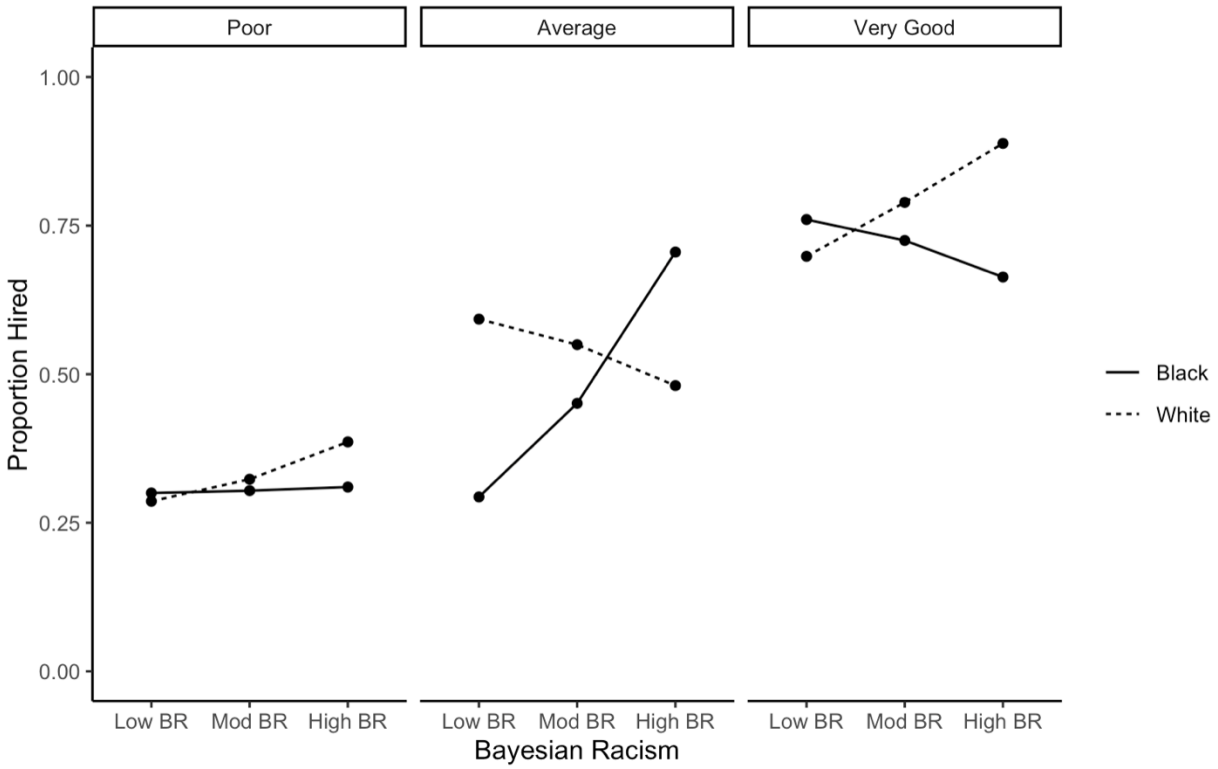


Figure 2

Target Race X Bayesian Racism interaction at each level of performance for objective ratings (significant in the Poor and Average performance conditions), Study 2, N=342.

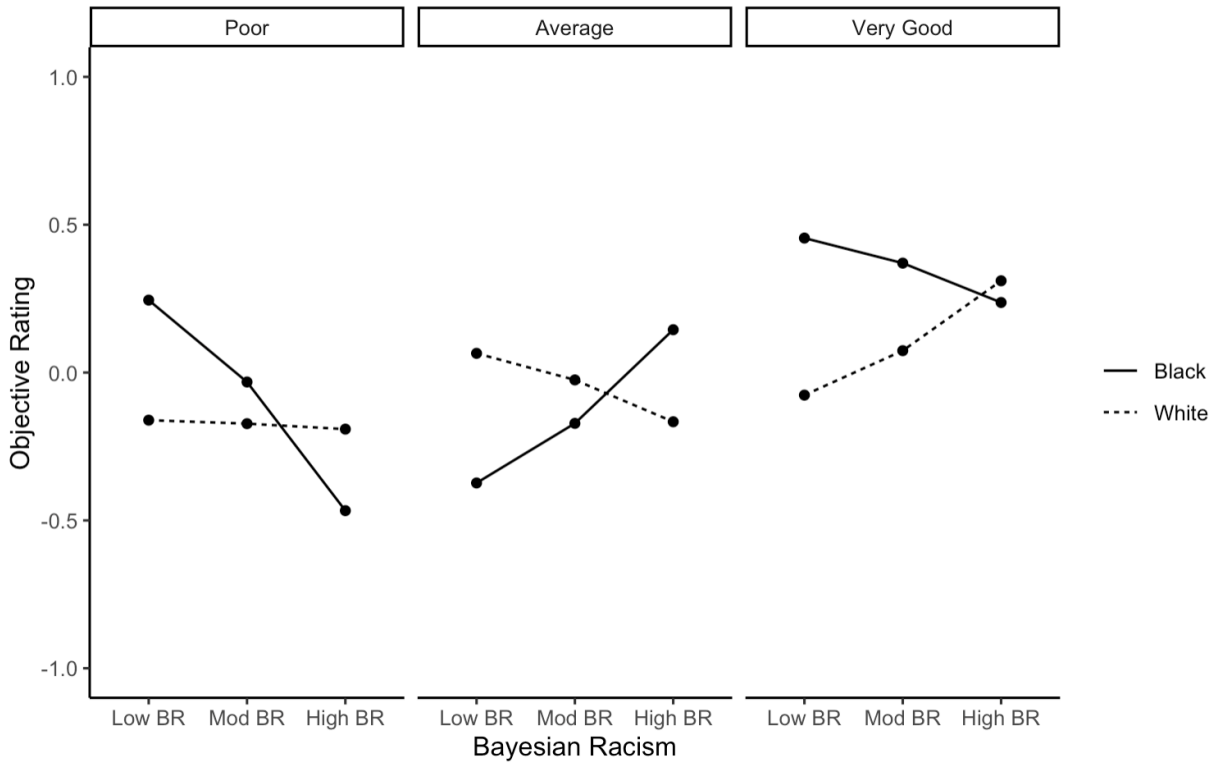


Figure 3

Target Race X Bayesian Racism interaction at each level of performance for total subjective ratings at Time 3 (significant in the Poor and Average performance conditions), Study 2, N=342.

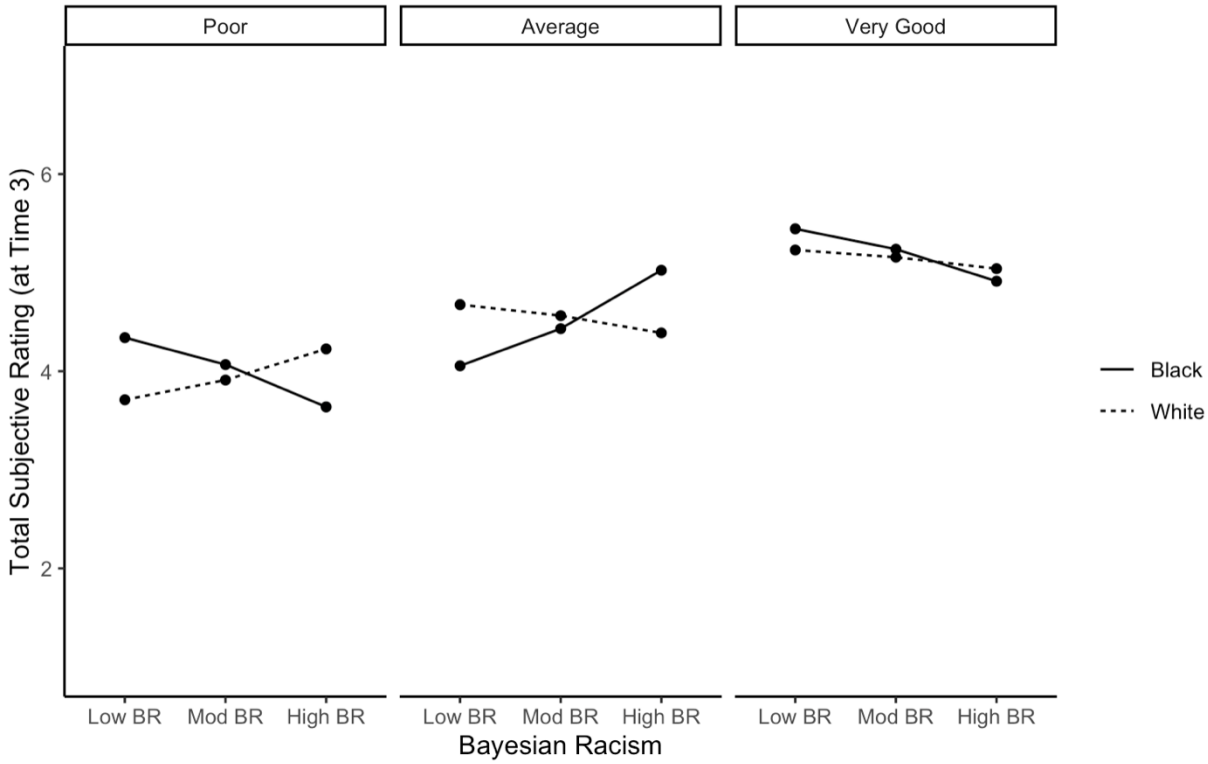
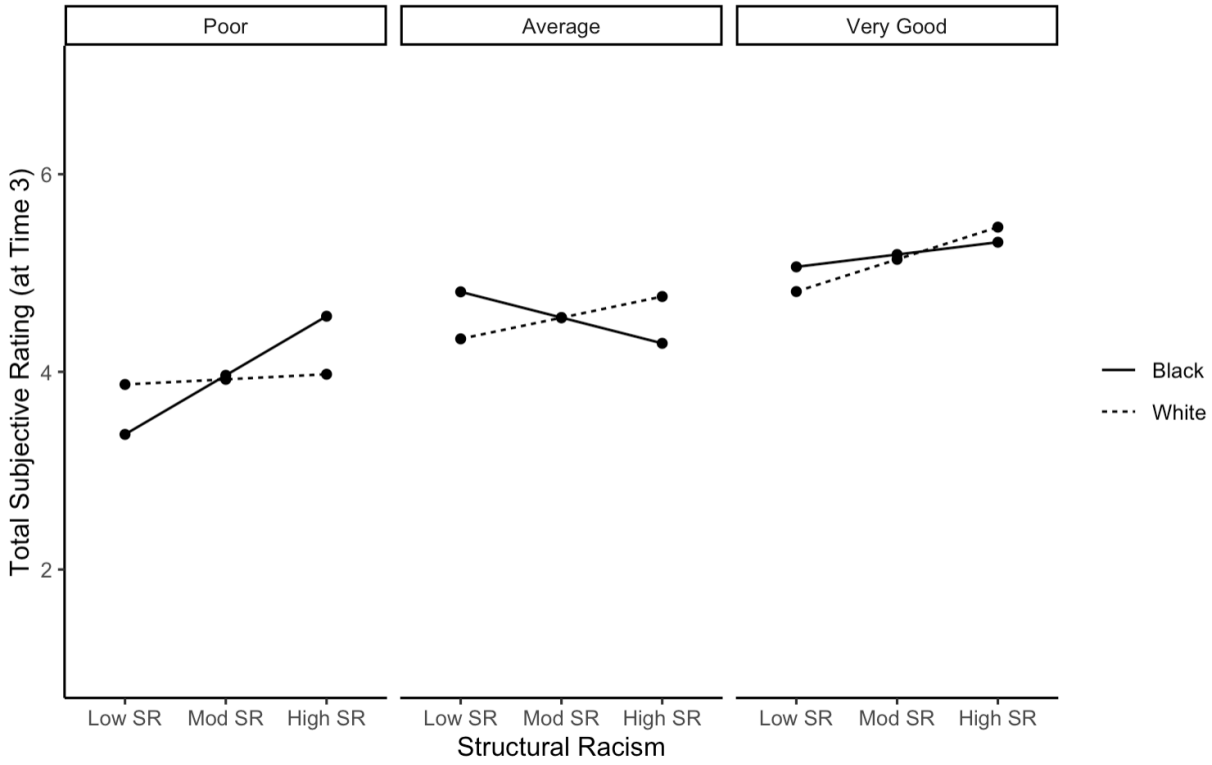


Figure 4

Target Race X Structural Racism interaction at each level of performance for total subjective ratings at Time 3 (marginally significant in the Poor and Average performance conditions), Study 2, N=342.



References

Adams, G., Dobles, I., Gómez, L. H., Kurtiş, Tuğçe, & Molina, L. E. (2015). Decolonizing psychological science: Introduction to the special thematic section. *Journal of Social and Political Psychology, 3*(1), 213-236.

Ajilore, O. (2020, February 24). *On the persistence of the Black-White wage gap.* Center for American Progress. https://www.americanprogress.org/issues/economy/reports/2020/02/24/480743/persistence-black-white-unemployment-gap/

Allport, G. *The nature of prejudice.* Cambridge, Massachusetts: Addison-Wesley Pub. Co., 1954.

Anderson, N. (1965). Averaging vs. adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, 70, 394-400.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 3030-314.

Ballakrishnen, S., Fielding-Singh, P., Magliozzi, D. (2019). Intentional invisibility: Professional women and the navigation of workplace constraints. *Sociological Perspectives, 62*(1), 23-41.

Barriers and bias: The status of women in leadership. American Association of University women. https://www.aauw.org/resources/research/barrier-bias/.

Bertocchi, G., & Dimico, A. (2010). Slavery, education, and inequality. [Discussion Paper No. 5329]. *Institute for the Study of Labor (IZA), Bonn, Germany.*

Biernat, M., Collins, E. C., Katzarska-Miller, I., & Thompson, E. R. (2009). Race-based shifting standards and racial discrimination. *Personality and Social Psychology Bulletin, 35*(1), 16-28. Doi: 10.1177/0146167208325195.

Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology, 72*(3), 544-557.

Biernat, M., & Manis, M. (1994). Shifting Standards and Stereotype-Based Judgments. *Journal of Personality and Social Psychology, 66*(1), 5-20. http://doi:10.1037/0022-3514.66.1.5

Biernat, M., Manis, M., & Nelson, T. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology, 60*(4), 485-499. http://doi:10.1037/0022-3514.60.4.485.

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *A dual process model of impression formation* (pp. 1–36). Lawrence Erlbaum Associates, Inc.

Brooks, K. J. (2019, December 10). *Why so many black business professionals are missing from the C-suite.* CBS. https://www.cbsnews.com/news/black-professionals-hold-only-3-percent-of-executive-jobs-1-percent-of-ceo-jobs-at-fortune-500-firms-new-report-says/

Castilla, E.J. (2008). Gender, race, and meritocracy in organizational careers. *American Journal of Sociology, 113,* 1479-1526.

Chandler, J., Mueller, P., & Paolacci, G. (2013). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods.*

Chang, A., Martin, R., & Marrapodi, E. (2020, August 16). *Summer of racial reckoning.* National Public Radio. https://www.npr.org/2020/08/16/902179773/summer-of-racial-reckoning-the-match-lit

Cho, B., & Lumsden, J. (2019, December 12). *Data quality at Prolific – Part 2: Naivety and engagement.* Prolific. https://www.prolific.co/blog/data-quality-at-prolific-part-2-naivety-and-engagement

Deloitte & Datawheel (n.d.). Data USA: Mechanical Engineers. Retrieved from https://datausa.io/profile/soc/mechanical-engineers

DeSante, C. D. (2013). Working twice as hard to get half as far: Race, work ethic, and America's deserving poor. *American Journal of Political Science, 57*(2), 342-356.

Devine, P. G. (1989). Stereotypes and prejudice: their automatic and controlled components. Journal of Personality and Social Psychology, 56(1), 5-18.

Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes *really* fading? The Princeton Trilogy revisited. *Personality and Social Psychology Bulletin, 21*(11), 1139-1150.

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

Eagly, A. H. (2004). Prejudice: Toward a more inclusive understanding. In A. H. Eagly, R. M. Baron & V. L. Hamilton (Eds.), *The social psychology of group identity and social conflict: Theory, application, and practice* (pp. 45-64). American Psychological Association, Washington, DC. doi:http://dx.doi.org/10.1037/10683-003

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573-598.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

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

Gaddis, M. S. (2017). "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." *Sociological Science, 4,* 469-489.

Gould, E. (2020, February 20). *State of working America: Wages 2019.* Economic Policy Institute. https://www.epi.org/publication/swa-wages-2019/

Edwards. *Engineering summer internship program*. Edwards Careers. https://edwards.wd5.myworkdayjobs.com/en-US/edwardscareers/job/USA---California--Irvine/Engineering-Summer-Internship-Program-2022_Req-10554)

Hanks, A., Solomon, D., & Weller, C. E. (2018, February 21). *How America's structural racism helped create the Black-White wealth gap.* Center for American Progress. https://www.americanprogress.org/issues/race/reports/2018/02/21/447051/systematic-inequality/

Hardy III, J. H., Sion Tey, K., Cyrus-Lai, W., Martell, R. F., Olstad, A., & Uhlmann, E. L. (2021). Bias in context: Small biases in hiring evaluations have big consequences. *Journal of Management, 48*(3), 657-692.

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd edition). (T. D. Little, Ed.). Guilford Press.

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

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior, 32,* 112-135.

Heilman, M. E. (1983). Sex bias in work settings: The lack of fit model. *Research in Organization Behavior, 5,* 269-497.

Heilman, M. E., & Caleo, S. (2018). Combatting gender bias: A lack of fit framework. *Group Processes and Intergroup Relations, 21,* 725-744.

Heilman, M. E. , Manzi, F., & Caleo, S. (2019). Updating impressions: The differential effects of new performance information on evaluations of women and men. *Organizational Behavior and Human Decision Processes, 152,* 105-121.

Hodson, G., Dovidio, J. F., & Gaertner, S. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin, 28*(4), 460-471.

Kang, S. K., DeCelles, K. A., Tilesik, A., & Jun, S. (2016). Whitened Résumés: Race and Self-Presentation in the Labor Market. *Administrative Science Quarterly, 61*(3), 469-502.

Kravitz, D. A., & Platania, J. (1993). Attitudes and beliefs about affirmative action: Effects of target and of respondent sex and ethnicity. *Journal of Applied Psychology, 78*(6), 928-938.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review, 103*(2), 284-308.

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science, 347*, 262-265.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data. *Behavior Research Methods, 47*, 1122- 1135.

Martín-Baró, I. (1994). *Writings for a liberation psychology* (A. Aron & S. Corne, Trans.). Cambridge, MA, USA: Harvard University Press.

Meyer, M., Cimpian, A., & Leslie, S.-J. (2015). Women are underrepresented in fields where success is believed to require brilliance. *Frontiers in Psychology, 6*(235), 1-12.

Milkman, K. L., Akinola, M., & Chugh, D. (2012). Temporal distance and discrimination: An audit study in academia. *Psychological Science, 23*(7), 710-717. https://doi.org/10.1177/0956797611434539

Moskowitz, G. B. (2005). Dual-Process Models. In *Social Cognition: Understanding Self and Others*, (pp. 193-232). Guilford Publications.

Moss-Racusin, C. A., Phelan, J. E., & Rudman, L. A. (2010). When men break the gender rules: Status incongruity and backlash against modest men. *Psychology of Men & Masculinity*, *11*(2), 140–151. https://doi.org/10.1037/a0018093

Olson, J. M., Roese, N. J., & Zanna, M. P. (1996). Expectancies. In E. T. Higgins & A. W. Kruglanski (Eds.) *Social psychology handbook of basic principles* (pp. 211-238). The Guilford Press.

Pew Research Center, "2016 Racial Attitudes in America III Survey [Data and Codebook]" (Pew, 2016); www.pewsocialtrends.org/2016/06/27/on-views-of-race-andinequality-blacks-and-whites-are-worlds-apart/.

Phelan, J. E., Moss-Racusin, C. A., & Rudman, L. A. (2008). Competent yet out in the cold: Shifting Criteria for hiring reflect backlash toward agentic women. *Psychology of Women Quarterly, 32*, 406-413.

Povee, K., & Roberts, L. D. (2014). Qualitative research in psychology: Attitudes of psychology students and academic staff. *Australian Journal of Psychology, 66*(1), 28-37.

Pratto F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology, 67*(4), 741-763.

Quadlin, N. (2018). The mark of a woman's record: Gender and academic performance in hiring. *American Sociological Review, 83*(2), 331-360.

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences, 114*(41), 10870-10875.

Robinson, J., Rosenzweig, C., Moss, A. J., & Litman, L. (2019). Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PLos ONE, 14*(12), 1-29.

Rucker, J. M, & Richeson, J. A. (2021). Toward an understanding of structural racism: Implications for criminal justice system. *Science, 374*, 286-290.

Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the Implicit Association Test. *Group Processes and Intergroup Relations, 10*(3), 359-372.

Sahadi, J. (2020). *After years of talking about diversity, the number of Black leaders at US companies is still dismal.* CNN Business. https://www.cnn.com/2020/06/02/ success/diversity-and-black-leadership-in-corporate-america/index.html

Sesko, A. K., & Biernat, M. (2010). Prototypes of race and gender: The invisibility of Black women. *Journal of Experimental Social Psychology, 46*: 356-360.

Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression.* New York, NY: Cambridge University Press.

Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles, 41*(7/8), 509-528.

Uhlmann, E. L., Brescoll, V. L., & Machery, E. (2010). The motives underlying stereotype-based discrimination against members of stigmatized groups. *Social Justice Research, 23,* 1-16.

Van Fleet, D. D., & Atwater, L. (1997). Gender neutral names: Don't be so sure! *Sex Roles, 37*(1/2), 111-123.

Vogels, E. A., Anderson, M., Porteus, M., Baronavski, C., Atske, S., McClain, C., Auxier, B., Perrin, A., & Ramshankar, M. (2021, May 19). *Americans and 'Cancel Culture': Where some see calls for accountability, others see censorship, punishment*. Pew Research Center. https://www.pewresearch.org/internet/2021/05/19/americans-and-cancel-culture-where-some-see-calls-for-accountability-others-see-censorship-punishment/

Warner, J., Ellmann, N., & Boesch, D. (2018). *The women's leadership gap: Women's leadership by the numbers.* Center for American Progress. https://www.americanprogress.org/issues/women/reports/2018/11/20/461273/womens-leadership-gap-2/

Wilson, V., Miller, E., & Kassa, M. (2021, June 8). *Racial representation in professional occupations.* Economic Policy Institute. https://www.epi.org/publication/racial-representation-prof-occ/

Zhao, X., & Biernat, M. (2017). "Welcome to the U.S." but "change your name"? Adopting Anglo names and discrimination. *Journal of Experimental Social Psychology, 70*, 59-68.

Appendix A: Job Description

Xenatech Engineer Post-Graduate Internship Program

Xenatech Engineering, Inc. is an engineering firm on the West Coast looking for brilliant soon-to-graduate students or recent graduates with degrees in engineering for our summer internship program. Our internship experience is a 10-week program designed to give recent college graduates real-world engineering experience in a fast-paced learning environment. Dependent upon need/resources at the time, there is potential for further advancement in the company to full-time positions. While candidates with mechanical engineering degrees are preferred, candidates with other engineering specialties are also invited to apply.

Role Expectations

- Be timely, punctual, and responsible for day-to-day attendance while proactively seeking direction from senior engineers
- Conceptualize, develop, and iterate designs in response to design requirements specified
- Test new designs (functional testing, mechanical testing, and simulated use) and brainstorm novel ideas for device improvements using test results
- Assist in documenting production department concerns and aid in troubleshooting and resolving technical problems
- Be proficient in drafting software (Solidworks or Creo Parametric preferred)
- Provide documentation support and complete written reports on testing and designs
- Actively participate in design reviews and provide relevant feedback

Requirements

- **Degree completion by May or early-June 2022**
- **Degree in Mechanical Engineering preferred**, but will accept applicants with degrees in other engineering-related fields
- **Comfortable working in a laboratory environment**. Hands-on and highly motivated. Able to follow procedures, perform testing, and clearly convey data results
- **Strong interest in pursuing relevant industry experience**.
- **Attention to detail**.
- **Proficient in drafting software**. Solidworks a plus.
- **Must be in good scholastic standing.** (3.0+ GPA or equivalent).

This is a paid internship.

Appendix B: Resumes

Study 1

Charlie W[redacted]

[redacted]@gmail.com | [redacted]

EDUCATION

University of Nevada, Reno, Nevada
Bachelor of Science, Mechanical Engineering, Expected Graduation May 2022

- **Cumulative GPA:** 3.03/4.0
- **Honors and Awards:** Honor Roll Spring 2019

WORK EXPERIENCE

Dexco Materials, Inc., *Summer Intern,* Irvine, CA, Summer 2021

- Worked with other interns and senior engineers on the Product Development Team
- Gained hands-on experience using drafting software
- Organized team records and helped keep projects on track

Five Guys Burgers and Fries, *Crewmember*, Boulder City, NV, Summers 2016 – 2019

- Served and interacted with customers
- Maintained cleanliness of kitchen and dining area
- Assembled burgers and manned fry station

COMMUNITY AND EXTRACURRICULAR EXPERIENCE

Intramural Soccer Team, University of Nevada, Reno, NV, Fall 2018 – Spring 2020

Reno Club Soccer Tournament, *Volunteer*, Reno, NV, Spring 2019 – 2020

- Volunteer referee

Study 2

Cassie W[redacted]

[redacted]@gmail.com | [redacted]

EDUCATION

University of Nevada, Reno, Nevada
Bachelor of Science, Mechanical Engineering, Expected Graduation May 2022

- **Cumulative GPA:** 3.03/4.0
- **Honors and Awards:** Honor Roll Spring 2019

WORK EXPERIENCE

Dexco Materials, Inc., *Summer Intern,* Irvine, CA, Summer 2021

- Worked with other interns and senior engineers on the Product Development Team
- Gained hands-on experience using drafting software
- Organized team records and helped keep projects on track

Five Guys Burgers and Fries, *Crewmember*, Boulder City, NV, Summers 2016 – 2019

- Served and interacted with customers
- Maintained cleanliness of kitchen and dining area
- Assembled burgers and manned fry station

COMMUNITY AND EXTRACURRICULAR EXPERIENCE

Intramural Soccer Team, University of Nevada, Reno, NV, Fall 2018 – Spring 2020

Reno Club Soccer Tournament, *Volunteer*, Reno, NV, Spring 2019 – 2020

- Volunteer referee

Appendix C: Interviews

Note: The only differences between interviews in Study 1 and Study 2 are the applicant's name and photo. I have included the photos and full interviews used in Study 1, and just the photos for Study 2.

Study 1

Poor Interview

Charlie's Interview

Interviewer: "Tell me a bit about why you are interested in the internship with our company."

Applicant: "Well… uh, you know… I really like mechanical engineering and I've always wanted to live on the West Coast, so… you know… it just felt like a good fit."

Interviewer: "Okay, and what skills or strengths do you think you bring to the position?"

Applicant: "Uh, well I think I'm a pretty independent worker, and I'm good at math, and…uh… I'm pretty good at working on a team I think, so yeah."

Interviewer: "And, do you feel that you work well in a fast-paced environment? How do you manage multiple deadlines?"

Applicant: "I mean, yeah, I think I'm pretty good at working quickly. I try to just work as fast as I can on one project at a time and finish that before I start anything else so that things aren't just left unfinished."

Average Interview

Charlie's Interview

Interviewer: "Tell me a bit about why you are interested in the internship with our company."

Applicant: "Well, when I saw the position available at your company, I was excited because I admire what your company does. I've also always wanted to live on the West Coast, so this seemed like a good fit. I'm excited about the potential of the internship turning into a full-time position."

Interviewer: "Okay, and what skills or strengths do you think you bring to the position?"

Applicant: "I think, you know, I work pretty independently, but I also work well on a team. And, math is one of my strong suits, so I will do well in this kind of position."

Interviewer: "And, do you feel that you work well in a fast-paced environment? How do you manage multiple deadlines?"

Applicant: "Yeah, I think I do. I often have multiple assignments due at the same time in college, so I'm used to having to work on multiple things at once."

Very Good Interview

Charlie's Interview

Interviewer: "Tell me a bit about why you are interested in the internship with our company."

Applicant: "Well, when I saw the position for the Internship program online, I was very excited because I have admired the work your company does for a while, and the opportunity to work with the engineers at Xenatech was one I couldn't pass up. I particularly like your company because of the clients you work with and the real impact your work has. I'm really excited to put what I've learned in college to use in the real world."

Interviewer: "Okay, and what skills or strengths do you think you bring to the position?"

Applicant: "I'm a very independent worker and am able to work well on my own, but I also know when to ask questions or for help, and I do really enjoy working on team. I also really enjoy taking complex problems and distilling them into easier to understand reports."

Interviewer: "And, do you feel that you work well in a fast-paced environment? How do you manage multiple deadlines?"

Applicant: "I feel that I thrive in fast-paced environments. I often had multiple deadlines to work on at the same time in college, as well as extracurriculars. I'm good at prioritizing and making sure the most important things get done on time. I like to stay busy – it motivates me to work harder."

Study 2

Appendix D: Writing Sample

Note: The writing sample was the same for Study 1 and Study 2.

Writing Sample

(This is a sample of a paper from a class I took on famous scientists)

In 1616 William Harvey was the first to discovery the circulation of the blood through observation and dissection of animals. At this time medical knowledge was taken straight from textbooks, people didn't challenge it. Although his discoveries and his public publishing of them were spread out and Harvey's theories challenge the medical theories that everybody had blindly believed for centuries. However, Harvey seemed kind of scared to tell people about these findings that would challenge things, for example he did not publish his findings until twelve years later. In this paper, I think that Harvey was a scared "New Scientist" because although some of Harvey's findings and thoughts come from older traditions, his methodology and discoveries were part of the "New Science" tradition that looked at observation and experimentation based on nature.

In this paper I will first explore the medical theories existing both before and during Harvey's time, including the findings and teachings of some other scientists and then I will elaborate on Harvey's discoveries and his presentation of them in his book *De Motu Cordis*.

Appendix E: Questionnaire Items and Answer Scales

Materials manipulation check

How would you rate the quality of the resume you saw, on a scale of 1 (Very poor) to 7 (Very good)?

1 – Very poor
2
3
4
5
6
7 – Very good

How would you rate the quality of the interview, on a scale of 1 (Very poor) to 7 (Very good)?

1 – Very poor
2
3
4
5
6
7 – Very good

How would you rate the quality of the writing sample you read, on a scale of 1 (Very poor) to 7 (Very good)?

1 – Very poor
2
3
4
5
6
7 – Very good

Subjective trait ratings – same at each of 3 time points

How competent do you think [Charlie/Cassie] would be in the Engineering Internship, on a scale of 1 (Not at all competent) to 7 (Very competent)?

1 – Not at all competent
2
3
4
5
6
7 – Very competent

How successful do you think [Charlie/Cassie] would be in the Engineering Internship, on a scale of 1 (Not at all successful) to 7 (Very successful)?

1 – Not at all successful
2
3
4
5
6
7 – Very successful

How effective do you think [Charlie/Cassie] would be in the Engineering Internship, on a scale of 1 (Not at all effective) to 7 (Very effective)?

1 – Not at all effective
2
3
4
5
6
7 – Very effective

How competent, in general, do you think [Charlie/Cassie] is, on a scale of 1 (Not at all competent) to 7 (Very competent)?

1 – Not at all competent
2
3
4
5
6
7 – Very competent

How assertive do you think [Charlie/Cassie] is, on a scale of 1 (Not at all assertive) to 7 (Very assertive)?

1 – Not at all assertive
2
3
4
5
6
7 – Very assertive

How warm do you think [Charlie/Cassie] is, on a scale of 1 (Not at all successful) to 7 (Very successful)?

1 – Not at all warm
2
3

4
5
6
7 – Very warm

Dependent variables

Would you hire [Charlie/Cassie] for the Mechanical Engineering Internship?
Yes, I would hire [Charlie/Cassie] for the internship.
No, I would hire [Charlie/Cassie] for the internship.

[If hired] Why did you decide to hire Charlie? [Open-ended response]

[If not hired] Why did you decide not to hire Charlie? [Open-ended response]

How important were the following criteria in your decision to hire or not hire [Charlie/Cassie] for the Engineering Internship? Please rank the criteria you used to make your decision, from 1 (most important) to 6 (least important), by clicking and dragging each option into its appropriate rank. [ranking]
Undergraduate GPA
Previous work experience
Extracurricular/service activities
Interview quality
Writing sample quality

Internships like this, given the limited time frame, are usually salaried, rather than paid by the hour. This salary/stipend can range from about $3,000 to $10,000 for the 10 weeks. If [Charlie/Cassie] was hired, how much do you think Charlie should be paid for the 10-week internship? [sliding scale]

3000--------4000--------5000--------6000--------7000--------8000---------9000--------10000

An intern in this type of program general completes between 1 and 6 projects during the course of their 10-week internship. If [Charlie/Cassie] was hired, how many projects do you think [Charlie/Cassie] would complete during the internship? [sliding scale]

1----------2----------3----------4----------5---------6

The GRE (Graduate Record Examination) is a standardized test often used for admission to graduate school. Total GRE scores can range from 260 to 340. If [Charlie/Cassie] were to take the GRE, what score do you think [Charlie/Cassie] would get? [sliding scale]

260--------270--------280--------290--------300--------310---------320---------330--------340

Do you think [Charlie/Cassie] should be hired at Xenatech full-time after the Internship is completed?

Yes
No

Did anything surprise you about [Charlie/Cassie]? [Open-ended response]

Before you found out about [Charlie's gender/Cassie's race], you may have had an image of [Charlie/Cassie] in mind. Did you think Charlie was…

A man/White
A woman/Black
I did not think about [Charlie's gender/Cassie's race/ethnicity]

Moderators

Measure of Structural Sexism (Study 1) / Structural Racism (Study2)

Please indicate to what extent you believe sexism/racism is an individual problem (i.e., sexism/racism is individual bias towards another person or group of people) or a structural problem (i.e., sexism/racism is part of the structure and laws of our institutions).

1 – Individual problem
2
3
4
5
6
7 – Structural problem

Measure of Attitudes towards Affirmative Action (Study 1)

1 – Strongly disagree
2
3
4
5 – Strongly agree

- Affirmative action is a good policy.
- I would not like to work at an organization with an affirmative action plan. (reverse-scored)
- Employees should be actively involved in attempts to improve the affirmative action conditions at their place of employment.
- I would be willing to work at an organization with an affirmative action plan.
- All in all, I oppose affirmative action plans in industry for women. (reverse-scored)

Modern Sexism Scale (Study 1)

1 – Strongly agree
2
3
4
5 – Strongly disagree

- Women often miss out on good jobs due to sexual discrimination. (reverse-scored)
- It is rare to see women treated in a sexist manner on television.
- Society as reached the point where women and men have equal opportunities for achievement.
- It is easy to understand the anger of women's groups in America. (reverse-scored)
- Over the past few years, the government and news have been showing more concern about the treatment of women than is warranted by women's actual experiences.
- Discrimination against women is no longer a problem in the United States.
- On average, people in our society treat husbands and wives equally.
- It is easy to understand why women's groups are still concerned about society limitations of women's opportunities. (reverse-scored)

Bayesian Racism Scale (Study 2)

1 – Strongly disagree
2
3
4
5
6
7 – Strongly agree

- If you want to make accurate predictions, you should use information about a person's ethnic group when deciding if they will perform well.
- If your personal safety is at stake, it's sensible to avoid members of ethnic groups known to behave more aggressively.
- When the only thing you know about someone is their race, it makes sense to use your knowledge of their racial group to form an impression of them.
- Law enforcement officers should pay particular attention to those social groups more heavily involved in crime, even if this means focusing on members of particular ethnic groups.
- Law enforcement officers should act as if members of all racial groups are equally likely to commit crimes. (reverse-scored)
- It is always wrong to avoid someone because members of their racial group are more likely to commit violent crimes. (reverse-scored)

Target Manipulation checks

What was [Charlie's/Cassie's] race? If you can't remember, please select the "I don't know" option.

White
Black
Asian American
Hispanic/Latinx
I don't know

What was [Charlie's/Cassie's] gender? If you can't remember, please select the "I don't know" option.

Man
Woman
I don't know

Demographic questions

With which group do you identify?

Male
Female
Gender non-binary or non-conforming
Trans
Other (open-ended)

How old are you? [Open-ended response]

Are you of Hispanic, Latino, or Spanish origin?

Yes
No

With which racial group(s) do you identify? Check all that apply.

White/Caucasian
Black/African American
Latinx
Asian/Pacific Islander
Native American
Middle Eastern
Other (open-ended)

Is English your first language?

Yes
No
Prefer not to answer

How would you describe your political views?

- Very conservative
- Conservative
- Moderate
- Liberal
- Very liberal

In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?

- Republican
- Democrat
- Independent
- Other (open-ended)

What is the highest level of education you have completed?

- High school, GED or less
- Vocational or technical school
- Some college, 2-year college, associate degree
- 4-year college degree, bachelor's degree
- Post-graduate (e.g., master's, doctorate)

Appendix F: Criteria Ranking Analyses

Study 1

GPA

The 2 (Target Gender) X 3 (Performance) ANOVA indicated a marginally significant main effect of performance, $F(2, 313) = 2.976$, $p = .052$. Pairwise comparisons indicated that in the Poor performance condition, GPA was ranked significantly less important ($M = 3.311$, $SD = 1.206$) than in the Average condition ($M = 2.905$, $SD = 1.327$), $p = .017$, and non-significantly less important than in the Very Good ($M = 3.046$, $SD = 1.155$) condition, $p = .118$. There was no main effect of Target Gender, $F(1, 313) = 0.210$, $p = .647$, and no significant interaction, $F(2, 313) = 1.423$, $p = .242$.

Previous Work Experience, Extracurriculars/Service Activities, Writing Sample Quality

No significant main or interaction effects emerged for previous work experience, extracurriculars/service activities, and writing sample quality in the full sample.

In the White sample only, there was a significant main effect of Performance on ranking of the writing sample, $F(2, 243) = 5.494$, $p = 0.005$, $\eta_p^2 = 0.043$.[6] The writing sample was ranked non-significantly more important in the Poor condition ($M = 3.054$, $SD = 1.258$) than in the Average condition ($M = 3.416$, $SD = 1.500$), $p = .086$; the writing sample was non-significantly less important in the Very Good condition ($M = 3.736$, $SD = 1.323$) compared to the Average

[6] When decision to hire or not was included as another factor in the ANOVAs on criteria rankings, those who hired versus did not hire ranked the criteria differently (main effects of hiring decision were significant for all 5 rankings, ps < .002. Those who hired the applicant ranked GPA and Extracurricular activities lower (more important), and ranked the other 3 criteria higher (less important), than those who did not hire the applicant.

condition, $p = .131$. The only significant difference in ranking was between the Poor and Very Good conditions, $p = .001$.

Interview Quality

There was a main effect of Performance, $F(2, 313) = 3.540$, $p = .030$, such that Interview Quality was significantly more important in the Poor condition ($M = 2.003$, $SD = 1.287$) than in the Average condition ($M = 2.400$, $SD = 1.198$), $p = .016$; Interview Quality was significantly more important in the Very Good condition ($M = 2.045$, $SD = 1.062$) than in the Average condition, $p = .030$. The effects of Target Gender, $F(1, 313) = 0.323$, $p = .570$, and the interaction, $F(2, 313) = 0.041$, $p = .959$, were not significant.

Study 2

GPA

A 2 (Target Race: Black, White) X 3 (Performance: Poor, Average, Very Good) ANOVA for the use of GPA as a criterion in the hiring decision resulted in no significant findings. The main effect of performance was not significant, $F(2, 333) = 1.679$, $p = 0.188$, $\eta_p^2 = 0.010$, the main effect of target race was not significant, $F(1, 333) = 0.376$, $p = 0.540$, $\eta_p^2 = 0.001$, and the two-way interaction was also not significant, $F(2, 333) = 0.001$, $p = .999$, $\eta_p^2 = 0.000$.

Previous Work Experience

The 2 (Target Race) X 3 (Performance) ANOVA for previous work experience indicated no significant effects. The main effect of performance was not significant, $F(2, 333) = 0.597$, $p = 0.551$, $\eta_p^2 = 0.004$, the main effect of target race was not significant, $F(1, 333) = 0.010$, $p = 0.920$, $\eta_p^2 = 0.000$. , and the two-way interaction was also not significant, $F(2, 333) = 0.830$, $p = 0.437$, $\eta_p^2 = 0.005$.

Extracurricular/Service Activities

The two-way ANOVA indicated a significant main effect of performance, $F(2, 333) = 6.650$, $p = .001$, $\eta_p^2 = 0.038$. Although extracurricular/service activities were seen as least important in the Poor condition ($M = 4.461$, $SD = 0.836$) and most important in the Very Good condition ($M = 3.994$, $SD = 1.066$), there were only significant comparisons between the Average ($M = 4.296$, $SD = 1.028$) and Very Good conditions, $p = .022$, and the Poor and Very Good conditions, $p < .001$. There was also a marginally significant main effect of target race, $F(1, 333) = 3.511$, $p = .062$, $\eta_p^2 = 0.010$. The interaction was not significant, $F(2, 333) = 1.246$, $p = .289$, $\eta_p^2 = 0.010$.

Interview Quality

The 2 (Target Race) X 3 (Performance) ANOVA for the use of interview quality as a criterion for hiring did not result in any significant findings. The main effect of performance was not significant, $F(2, 333) = 1.993$, $p = 0.138$, $\eta_p^2 = 0.012$, the main effect of target race was not significant, $F(1, 333) = 0.001$, $p = 0.976$, $\eta_p^2 = 0.000$, and the two-way interaction was also not significant, $F(2, 333) = 0.632$, $p = 0.532$, $\eta_p^2 = 0.004$.

Writing Sample Quality

The 2 (Target Race) X 3 (Performance) ANOVA for writing sample quality resulted in a marginally significant main effect of performance, $F(2, 333) = 2.943$, $p = 0.054$, $\eta_p^2 = 0.017$. The main effect of target race, $F(1, 333) = 0.532$, $p = 0.466$, $\eta_p^2 = 0.002$, and the two-way interaction, $F(2, 333) = 0.118$, $p = 0.889$, $\eta_p^2 = 0.001$, were not significant.