Is white always the standard? Using replication to revisit and extend what we know about the leadership prototype

William G. Obenauer

Michael J. Kalsher

William G. Obenauer
Corresponding Author
University of Maine
Maine Business School
Management Department
168 College Ave, Orono, ME 04469
william.obenauer@maine.edu

The proper citation for this article is:

Obenauer, W. G., & Kalsher, M. J. (2022). Is white always the standard? Using replication to revisit and extend what we know about the leadership prototype. The Leadership Quarterly, 101633.

The published journal article can be accessed at:

https://doi.org/10.1016/j.leaqua.2022.101633

Replication data are available at:

https://data.mendeley.com/datasets/zzcwjwxcmt

Is White Always the Standard?

Using Replication to Revisit and Extend What We Know About the Leadership Prototype

ABSTRACT

This research is a pre-registered replication of Rosette, Leonardelli, and Phillips' (2008) seminal work in leadership categorization theory. Their work established race as a component to the business leader prototype and found evidence that when a leader was given credit for successful organizational performance, White leaders were evaluated more favorably than non-White leaders. As leadership exemplars are evolving, however, a need to reexamine these relationships has emerged. Results from our replications of their first and third studies showed minimal support for the argument that being White is a component of the business leader prototype. Additionally, across six separate studies, we found no conditions in which White leaders received more favorable evaluations than their non-White counterparts. Contrary to our expectations, we found that non-White leaders received marginally more favorable ratings than White leaders in four of our studies.

Keywords: leadership categorization, leadership prototype, leadership evaluation, experimental design, leadership attributions

INTRODUCTION

In January of 2015, Vincent R. Stewart made history in becoming the first Black Director of the U.S. Defense Intelligence Agency, a post where he provided regular briefings to the president and was responsible for more than 16,000 employees. Reflecting on his stellar leadership career, Lt. Gen. Stewart wrote, "It's hard for me to explain and help you understand the pain of being described as the best black officer in a unit, never able to be described as the best officer in the unit; never the first choice for visible prominent assignments in spite of a record of performance that was superior to my colleagues" (Stewart, 2020). Lt. Gen. Stewart's observation is consistent with extant research showing that minorities in leadership positions experience negative outcomes in terms of achieving promotions (James, 2000; Powell & Butterfield, 1997), promotion quality (Cook & Glass, 2014a), evaluations (Cox Jr. & Nkomo, 1986; Greenhaus et al., 1990; Greenhaus & Parasuraman, 1993) and likelihood of employment termination (*citation removed*). Whereas there is extensive documentation of these negative outcomes, there is still an urgent need for researchers to identify the *causes* of said outcomes in order to facilitate the development of effective solutions to this problem.

Rosette et al. (2008) offered leadership categorization theory (LCT) as one potential explanation for the negative outcomes experienced by racial minorities in leadership positions. In a series of experimental vignette studies, they found support for this explanation as White leaders were evaluated more favorably than non-White leaders, but only when receiving credit for successful firm performance. A close examination of the summary statistics for several recent experimental vignette studies, however, indicates that these findings of differential evaluations may not be universally generalizable (e.g. Gündemir et al., 2019; Hekman et al., 2017; Reynolds et al., 2021; Salerno et al., 2019; Ubaka et al., 2021). The current research seeks to address this

tension in the literature and contribute to this research stream through a series of replications and extensions of Rosette et al.'s (2008) experimental vignette research on LCT.

Leadership Categorization Theory (LCT)

Although not originally conceived with the sole intent of explaining racial discrimination in leadership positions, LCT has demonstrated the potential to provide some insight into this phenomenon. LCT draws upon theories of human information processing which state that people engage in two types of information processing, *controlled* and *automatic*, due to their limited information processing capabilities. Tasks that involve controlled processing require deliberate attention (e.g. performing surgery, learning new activities) whereas tasks that involve automatic processing can often be completed without much focus on the actual task (e.g. walking, performing routine activities). Although people can only engage in controlled processing for one activity at a time, they often toggle between controlled and automatic processing while completing tasks in everyday life. For example, a person who is walking while performing another task (e.g. talking, texting) will primarily rely on automatic processing for walking, but must engage in controlled processing as they periodically avoid obstacles in their path, adjust steps to account for unsure footing, etc. (Lord & Maher, 1993).

According to LCT, to preserve cognitive resources, as information is processed in memory, stimuli are categorized based upon perceptions of the environment. This categorization process results in a coarsening of information such that information that is perceived to be important is retained in memory while that which is perceived as less important is discarded. The retained information contributes to the development of prototypes, or mental models, that represent various environmental stimuli (Rosch, 1978). Lord and Maher (1993) reasoned that when engaged in automatic processing, individuals draw upon prototypes in developing quick

and efficient responses to stimuli. As individuals develop prototypes for leaders, these prototypes influence how they evaluate and respond to different leaders moving forward.

Whereas research examining discrimination against racial minorities in leadership positions has often focused on the effects of negative attitudes towards racial minorities (e.g. Dovidio & Gaertner, 2000; Greenhaus & Parasuraman, 1993; James, 2000), Rosette, Leonardelli, and Phillips (2008; hereafter referred to as RLP) drew upon LCT to propose that racial discrepancies in leadership outcomes could more appropriately be explained by a preference for White leaders than by negative attitudes towards minorities. Recognizing that prototypes are influenced by perceptions of correlation (Rosch, 1978), RLP argued that the disproportionate representation of Whites in leadership positions contributed to "being White [serving as] a central characteristic of the business leader prototype" (p. 760). They argued that this phenomenon caused racial minorities in leadership positions to face challenges that are unique to those faced by racial minorities within other employment contexts.

RLP's first study (RLP1) established race as a component of the business leader prototype by demonstrating that in the absence of racial cues, undergraduate participants were more likely to perceive a business leader as White than they were to perceive a non-leader as White. This study also found that while participants perceived the leader to be White at a higher rate than the population base rate would suggest, such an effect was not present for the non-leader. In a sample of MBA students, the paper's second study (RLP2) found additional support for these findings while also finding evidence that the racial component to the business leader prototype was independent of industry.

RLP then examined the conditions in which a White leader would be evaluated more favorably than a non-White leader. Study 3 (RLP3) examined the effects of leader race,

organizational performance, and external attributions for performance on leadership evaluations in a sample of undergraduate students. The core contribution of this study was the finding that "White leaders were evaluated more favorably than non-White leaders, but only when successful organizational performance was attributed to the leader" (Rosette et al., 2008, p. 769). The paper's fourth study (RLP4) replicated this finding in a sample of graduate students while measuring perceptions of leadership *potential* for various racial and ethnic minorities, thus providing robustness to RLP's theoretical contribution.

RLP argued that findings showing the presence of a White leadership prototype (RLP1), considered in conjunction with findings indicating that the success of non-White leaders is discounted when success is attributed to the leader (RLP3), provided support for their theoretical argument that leadership categorization causes non-White *leaders* to face unique challenges that are derived from racial expectations for leaders, rather than negative racial attitudes towards specific groups of minorities. RLP made an important contribution to LCT as evidenced by the 501 Google Scholar citations, or average of 35.79 citations per year, that the paper had amassed as of October 2021. Accordingly, subsequent research has used RLP to further develop LCT and our understanding of how leadership categorization influences outcomes for non-White leaders.

For example, Rosette and Livingston (2012) contributed to LCT through their finding that the relationship between race and perceptions of leadership prototypicality was influenced by organizational success. This effect emerged such that when an organization was successful, White leaders were perceived as more prototypical than Black leaders, but no such difference existed during times of organizational failure. They also found that prototypicality mediated the relationship between a leader's race and perceptions of effectiveness.

Other LCT research has investigated how followers' perceptions of leadership prototypicality influence outcomes for leaders. For example, van Quaquebeke et al. (2011) found that when followers perceived a leader as prototypical, they rated the leader highly and held the leader in high regard. When the leader was not perceived as prototypical, however, followers had less respect for the leader, particularly when the followers saw themselves as having prototypical leadership traits. Additionally, when leaders and followers share perceptions of what leadership prototypicality looks like, they have higher levels of LMX (Riggs & Porter, 2017).

Recognizing the impact that perceptions of prototypicality have on leadership outcomes, research on LCT has also focused on increasing our understanding of factors that influence these perceptions. Trichas et al. (2017) found that leaders with happy facial expression scored highest on perceptions of prototypical traits such as dedication, intelligence, and dynamism. Braun et al. (2018) found that demographic characteristics associated with an authentic leadership prototype differed from those associated with a more general leadership prototype. Kocoglu and Mithani (2020) found that in addition to being influenced by a leader's gender, perceptions of leadership prototypicality were influenced by the presence and attractiveness of a leader's romantic partner. Additionally, Sy et al. (2010) found that the relationship between a leader's race and impressions of their leadership typicality was moderated by industry. The current research builds on this growing stream of literature on LCT by recognizing the need to replicate the foundational research that the extensions described above have been built upon.

The Need for Replication and its Contribution to Theory

The replication crisis. Although the application of LCT within the context of management has developed considerably since the publication of RLP, replications of seminal works constitute critical scientific contributions. Failure to carry out and publish replications can

bias information that is provided to the scientific community (Schwab et al., 2011). Moreover, the growing number of replication results that have failed to mirror those of their target studies has given rise to largescale dialogue concerning what has been termed a "replication crisis" (e.g. Bishop, 2020; Loken & Gelman, 2017; Maxwell et al., 2015; Shrout & Rodgers, 2018). Failed replications do not necessarily invalidate prior research findings, but instead force us to consider the boundary conditions for established behavioral theory, thereby providing deeper insight, and motivating new research.

Whereas replication work in the area of leadership is limited (Clapp-Smith et al., 2018), the literature does provide some motivation to replicate research focused on discrimination in leadership outcomes. For example, while some research into Fortune 500 CEO transitions has found that minority leaders were promoted into lower quality positions than their White peers (Cook & Glass, 2014a), other research looking at this same population suggests that racial minorities may actually receive *higher quality* leadership positions (Cook & Glass, 2014b). Furthermore, while the findings of RLP showed no evidence of differential outcomes for White and non-White leaders in times of poor organizational performance, archival data has shown evidence that non-White leaders receive more criticism than their White counterparts in times of organizational failure (Carton & Rosette, 2011; Park & Westphal, 2013) and have an increased risk of employment separation (Obenauer & Langer, 2019). These examples of inconsistent findings suggest that research offering important contributions to our understanding of leadership theories, such as RLP, should be prime targets for replication.

The evolution of minority representation in leadership. Another factor that makes RLP prime for replication is that the contextual factors theorized to influence the development of the business leader prototype have been evolving. RLP argued that persistent exposure to prominent

White leaders primes the racial component of the leadership prototype. This means that changes in the overrepresentation of Whites in leadership could alter how individuals develop business leader prototypes. Although the current racial characteristics of individuals in chief executive officer positions (U.S. Bureau of Labor Statistics, 2017; White, 2017) and *Fortune* 100 Boards of Directors (Deloitte & Alliance for Board Diversity, 2017) largely reflect what was reported by RLP, this overrepresentation of Whites in executive positions has not been as consistent in other leadership positions.

Beneath the C-Suite, there have been notable changes in minority leadership representation. As shown in *Appendix AI and AII*[1], minority representation in positions such as lodging managers and food service managers is now comparable to that which would be predicted by the proportion of minorities in the overall population. Even in business leadership positions where racial minorities are still underrepresented, their growth in representation has been considerably greater than their overall population growth. Minority representation in visible university leadership positions has also increased since the publication of RLP (Lapchick, 2008, 2017). Furthermore, racial diversity in salient political leadership positions has also increased as minority representation has grown in both the U.S. Senate and the House of Representatives (Manning, 2018; Rosette et al., 2008; *see Appendix AI*). Additionally, the United States elected its first Black president in 2008. The above data signals an evolution in the exemplar of leaders that could influence perceptions of the leadership prototype.

We posit that these changes may also be reflective of a transformation in the leadership categorization process. Current understanding of leadership categorization would suggest that if being White is an attribute of the business leader prototype, categorization will result in non-

[1] All appendices are available in the online supplemental materials

White leadership candidates being evaluated less favorably than their White counterparts (e.g. RLP4). Such biased evaluations should preclude the increased minority representation in leadership positions described above. As racial minorities are still underrepresented in most leadership positions, it is likely that non-White leadership candidates are still subjected to biased evaluation processes, but their growth in representation suggests that the effect of said biases may be decreasing. This raises questions as to the continued pervasiveness of the White business leader prototype, thus contributing to the motivation to replicate RLP.

Tension in perceptions of racial discrimination. Consistent with the increased minority representation in leadership positions described above, racially-driven charges of employment discrimination have decreased in the decade since the publication of RLP (U.S. Equal Employment Opportunity Commission, 2019). Conversely, however, multiple polls have shown that *perceptions* of racial discrimination have actually increased during this same time period (e.g. Jones & Saad, 2016; Pew Research Center, 2016). This apparent conflict between reports of behavior and perceptions may be influenced by changes in how discrimination is expressed (Swim et al., 1995), along with the context dependence of its manifestation (e.g. Brief, Dietz, Cohen, Pugh, & Vaslow, 2000; Hekman, Johnson, Foo, & Yang, 2016; Zapata, Carton, & Liu, 2016), further highlighting the need for a deeper understanding as to how RLP's findings regarding the business leader prototype apply to current social contexts. Such an understanding can only be gleaned through replications that test the boundaries of the research's theoretical contribution (Whetten, 1989).

Examining boundary conditions. RLP3's research design makes it an appropriate target of replication that is designed to examine the boundary conditions of research findings. Like many organizational studies, RLP3 manipulated a leader's race through the use of names (e.g.

Bertrand & Mullainathan, 2004; Lee et al., 2015; Zapata et al., 2016) and photographs (e.g. Hekman et al., 2017; Younkin & Kuppuswamy, 2017; Zhu et al., 2016). The study's design also allowed us to address concerns regarding manipulation checks biasing results (Singleton Jr. & Straits, 2005) and the potential for research conducted in student samples to yield results that are not reflective of working professionals (Barr & Hitt, 1986; Singer & Bruhns, 1991). Addressing all four of these components of experimental design through replication contributes to the literature by informing us as to how choices in experimental design are influencing findings and thus shaping the way that we understand leadership theories such as LCT.

The Current Research

The current research involved 5,728 participants from eight different samples. This project began as a registered research report, which involved submitting proposed research questions, data collection methods, and data analysis procedures for peer-review *before* data were collected. The research plan was then refined through the peer-review process. Once a research plan is accepted by the editorial team, provided that the researchers follow the procedures described in the plan, work derived from registered research reports is generally accepted for publication irrespective of the actual empirical findings (Chambers, 2019; Clapp-Smith et al., 2018). Preregistering research helps address potential methodological concerns before data are collected, while also reducing the likelihood of researchers engaging in questionable research practices (Oberauer & Lewandowsky, 2019).

In the first phase of this research (Study 1A), we examined if being White is still considered a component of the business leader prototype by conducting a close replication of RLP1. A close replication attempts to use methods and procedures that are as close as possible to the original study. They are referred to as close, rather than exact or direct, because they are

typically conducted by different researchers and often involve minor, unavoidable (often undetectable) deviations from the original research (Brandt et al., 2014; Hüffmeier et al., 2016).

RLP1 was chosen for replication because unlike RLP2, this study included a condition in which the representation of Whites in the workforce was similar to that of the United States population. Additionally, RLP2 was designed to test the boundary conditions of LCT by manipulating industry and the representation of various minority groups. Because RLP1 held these variables constant, RLP1 is better positioned than RLP2 for identifying the initial presence of a business leader prototype. We then conducted a conceptual replication of RLP1 using a sample recruited through Amazon's Mechanical Turk (mTurk; Study 1B). Conceptual replications test the boundary conditions of theory by intentionally deviating from the methodology of the target research (Makel et al., 2012; Watts et al., 2018).

In the second phase of this research, we conducted two separate, close replications (Studies 2A and 2B) and one conceptual (Study 2C) replication of RLP3. RLP3 was chosen for replication because the study's manipulations had higher external validity than the manipulations used in RLP4 and the design of RLP3 allowed us to examine the impact of various aspects of experimental design on findings of differential outcomes for White and non-White leaders.

The final phase of this research, conducted using mTurk, involved a series of conceptual replications of RLP3 that allowed us to examine boundary conditions related to experimental design. In this series, we examined the influence of leader names (Study 3A), profile photographs (Study 3B), manipulation checks (Study 3C), and sample types (Study 3D) on findings. These replications differed from prior work that had modified these elements of experimental design (e.g. RLP4) in that our design isolated these variables in such a way that we were able to identify the specific influence that each design element had on our findings.

Studies 1A and 1B showed minimal support for the argument that White is a component of the business leader prototype. Studies 2 and 3 did not detect evidence that White leaders were evaluated more favorably than non-White leaders during times of organizational success. In fact, the most common finding related to differential evaluations across these studies was that the non-White leader received marginally higher evaluation ratings than the White leader, though the effect size was small. Furthermore, we found little evidence that choices in research design or sample had a meaningful impact on differences in the evaluations received by White and non-White leaders.

Although our findings differed from those of RLP, they should not be interpreted as a rebuke of LCT as extant research has illustrated the role of context in categorization (e.g. Gündemir et al., 2019; Sy et al., 2010). Furthermore, they should not be interpreted as evidence of universal racial equity in terms of leadership outcomes as research in this domain continues to show that is not the case (e.g. Hekman et al., 2017; Obenauer & Langer, 2019; Rosette, Koval, Ma, & Livingston, 2016). Instead, this research represents *a context* in which bias did not manifest into discrimination in favor of Whites in terms of leadership evaluations. Identifying pathways for replicating the contextual factors that contributed to our findings could have important applications for reducing discriminatory outcomes in the workplace.

STUDY 1

The purpose of this study was to examine whether being White was considered a component of the leadership prototype. This was achieved by having participants read an article about a fictional company in which the position of the person being interviewed in the article and the demographic composition of the company described in the article were manipulated. The experiment took on a 2 (interviewee role: leader, employee) X 3 (race base rate: no information,

50% White, 20% White) between-participants design. This phase of the research included both a close replication (Study 1A) and a conceptual replication (Study1B) of RLP1.

Study 1A

Research design. Study 1A represents a close replication of RLP1. To facilitate close replication of RLP1, data collection took place in-person using paper response packets. Participants were instructed to read one of six versions of an article on Selcom, Inc.'s Project Nova that were identical to those used in RLP1 (*see Appendix BI; RLP Appendix A*). The study's dependent variable was derived from a question asking the perceived race of the person interviewed in the article. Consistent with RLP1, participants also responded to two manipulation checks, several distractor questions, (*see Appendix BII*), and a voluntary demographic questionnaire.

Participants. RLP1 utilized data collected from 146 undergraduate participants with the number of participants per condition ranging from 22 to 28. Given the risk that low power poses to replication (Camerer et al., 2018; Open Science Collaboration, 2015), we performed a power analysis based on RLP1's Z-test of proportions using Stata14 (*see Appendix AIII*). Based upon the results of this power analysis, prior to data collection we set a target sample size of 420 participants (70 per condition) with a minimum sample size of 330 (55 per condition). Consistent with prior research (e.g. Zapata et al., 2016), we did not set a maximum sample size in order to ensure that student opportunities to participate in research were equitable.

Our sample consisted of 558 undergraduate students from a mid-sized research university in the northeastern United States. The racial diversity of this sample was similar to that of RLP1, though a greater proportion of RLP1's participants reported that they were actively working (*see Appendix AIV).* Robustness tests indicated that demographic traits did not meaningfully impact

results, thus they will not be discussed further. Data are available at doi: 10.17632/zzcwjwxcmt.1.

Primary analyses. Manipulation checks[2] were used to identify if participants consciously recognized and recalled the components to the study that were manipulated. All participants in RLP1 correctly responded to manipulation checks. This was not the case in the present study as 80 participants (14.3%) in Study 1A did not respond correctly to one or more manipulation checks.

The pre-registration for this study did not specify how to handle manipulation check failures, thus this issue was addressed post-hoc. Discarding participants who fail manipulation checks has the potential to bias results (Aronow et al., 2019) and can result in the reporting of significant findings when no significant findings truly exist (Kotzian et al., 2020). Furthermore, because LCT is based upon exposure to and automatic response to stimuli (Lord & Maher, 1993), not the conscious recognition and recall of stimuli, discarding data because of failed manipulation checks is not consistent with the theory being tested. With this in mind, our primary analyses focus on the full sample. In the interest of full transparency, results excluding participants who incorrectly responded to one or more manipulation checks have been included in the tables and are discussed in robustness tests.

RLP1 tested two hypotheses to identify whether being White was a component of the leadership prototype (*see Table 1*). The first hypothesis stated that "a main effect should occur whereby observers will consider leaders to be White more than non-leaders (e.g. employees), regardless of the base rate information" (p. 761). Consistent with their analysis, this was tested using binary hierarchical logistic regression analysis (Kleinbaum & Klein, 2002) with

[2] See questions 1 and 2 in Appendix BII

interviewee race (White/non-White) as the dependent variable and effect sizes reported in terms of $r_{equivalent}$ (Rosenthal & Rubin, 2003).

Overall, 466 out of 558 participants perceived the interviewee as White with 86.35 percent of participants perceiving the leader as White and 80.34 percent of participants perceiving the non-leader as White. Step 1 in *Table 2, Model 1* shows that the effect of the interviewee's role on perceptions of race had a p-value greater than 0.05 (*p=0.08*), thus failing to reach the threshold of statistical significance[3] and failing to fully replicate the results of RLP1.

The effect of the base rate on perceptions of race was significant (*Wald = 37.13, p<0.0001, r=0.25*). The interviewee was perceived as White less frequently in the 20% White condition (131 of 188 participants = 70%) than in either the no base rate condition (163 of 187 participants = 87%) or the 50% White condition (172 of 183 participants = 94%). This pattern is consistent with that of RLP1, though the overall rates of perceiving the interviewee as White were noticeably higher in the present study. Also consistent with RLP1, the interaction of interviewee role and base rate was not significant (*p=0.70*).

Insert Table 1 & 2 About Here

RLP1's second hypothesis stated that "participants would perceive the leader to be White more frequently than the [base rate] would suggest…but that this would not be the case for non-leaders" (p. 761). Consistent with RLP1, we tested this hypothesis using Z tests for proportions to compare the percentage of interviewees perceived as White to the corresponding base rates. As shown in Model 1 of Table 3, in the 50% White condition, participants perceived both the leader (*94%, Z=8.43, p<0.0001*) and non-leader (*94%, Z=8.40, p<0.0001*) as White more

[3] Earlier versions of this manuscript referred to relationships where 0.05<p<0.10 as "approaching significance." This language has been amended to conform to Neyman-Pearson methodology. We thank the anonymous methods reviewer for this helpful suggestion.

frequently than the base rate would suggest. Similarly, in the 20% White condition, participants perceived the leader (*73%, Z=12.65, p<0.0001*) and non-leader (*66%, Z=11.47, p<0.0001*) as White in the majority of observations. Collectively, these results failed to fully replicate those of RLP1 and provided no support for RLP's second hypothesis.

Insert Table 3 About Here

Robustness. Our findings were robust to model specification with one exception. When restricting the sample to participants who correctly responded to both manipulation checks, the hierarchical logistic regression analysis indicated that the effect of interviewee role on perceptions of race was significant (*see Table 2; B = -0.60, SE = 0.27, p=0.03, r=0.10*), providing some indication that leaders were perceived as being White more often than non-leaders. This model restriction did not impact the results of Z tests for proportion (*see Table 3*).

To address concerns that the use of interaction terms within logistic regressions can result in failure to report significance, false reporting of significance, or incorrect direction on coefficients of interaction terms (Ai & Norton, 2003), we replicated our binary hierarchical logistic regression in a linear probability model (Chatla & Shmueli, 2016). Additionally, to address the possibility that Z tests in the current study could detect effects that were undetectable in RLP1 due to the current study's large sample size, we randomly selected 25 observations per condition and replicated our Z test analyses. Results from these robustness tests were consistent with findings reported in the paper (*see Appendices AVI & AVII*).

Study 1B

Research design. Study 1B was a conceptual replication of RLP1 designed to examine whether being White emerged as a component of the business leader prototype in a sample other than a student sample. Experimental design was similar to that used in Study 1A with the lone

exception being that data were collected electronically through Amazon's Mechanical Turk (mTurk), a platform that has seen increased use in the leadership literature (e.g. Marchiondo, Myers, & Kopelman, 2015; O'Reilly, Doerr, & Chatman, 2017; Schaumberg & Flynn, 2017; Tucker, Ogunfowora, & Ehr, 2016). MTurk is an appropriate venue for a replication of RLP1 as the experimental manipulations and questionnaire were conducive to online formatting.

Participants. Participants were recruited in exchange for $0.25 and entry into a drawing for a chance to win a $50 bonus upon successful completion of the HIT. Per the study's preregistration and as approved by our IRB, successful completion was defined as spending at least thirty seconds on the task and correctly responding to at least three out of four attention checks. Attention checks differed from manipulation checks in that while manipulation checks were designed to capture participants' *perceptions* of manipulations, attention checks captured responses to details that were explicitly stated in the case *and present on the screen* at the point that the question was asked (e.g. "What is the name of the PROJECT described in this article?")

Data collected from 21 participants who did not meet these criteria were discarded. These criteria were implemented to reduce the threats to data quality that are associated with online data collection platforms (Hauser & Schwarz, 2016). Discarding low quality data eliminates noise that can mask significant effects (Oppenheimer et al., 2009). The resulting sample consisted of 498 individuals with a mean reported age of 37.38 years and 15.64 years of work experience (*see Appendix AIV*).

Results. Forty-six participants (9.24%) incorrectly responded to one or more manipulation checks. Consistent with Study 1A, the primary discussion of results focuses on the complete sample. Results for both the full sample and a subsample of participants who correctly

responded to manipulation checks are provided in *Tables 2 and 3* with critical differences addressed in the discussion of robustness of findings.

We used binary hierarchical logistic regression analysis to test whether leaders were perceived to be White more frequently than non-leaders. Both the constant and base rate were significant in all reported models (*ps<0.0001; see Table 2, Model 3*). The interviewee was perceived as White less frequently in the 20% White condition (112 of 164 participants = 68%) than in either the no base rate condition (150 of 171 participants = 88%) or the 50% White condition (156 of 163 participants = 96%). The effect of interviewee role was not significant (*p=0.15*), providing no indication a business leader prototype caused the leader (*85.89%*) to be perceived as White more frequently than the non-leader (*82.00%*).

We then used Z tests for proportions to examine RLP's Hypothesis 2. Model 3 of Table 3 shows that participants perceived the interviewee as White more often than the base rate would suggest in the leader / 50% White (*96%, Z=8.27, p<0.0001*), the non-leader / 50% White (*95%, Z=8.23, p<0.0001*), leader / 20% White (*72%, Z=11.93, p<0.0001*), and the non-leader / 20% White (*65%, Z=9.90, p<0.0001*) conditions. The results of Z tests in the current study were consistent with those of Study 1A.

Robustness. We performed the same robustness tests described in Study 1A (*see Tables 2&3, Appendices AVI&AVII*). Results were consistent with those discussed above. Similar to our analyses in the full model, when restricting the sample to participants who correctly responded to both manipulation checks, the effect of interviewee position on perceived race was not significant (*p=0.099*).

STUDY 2

After examining the role of race in the business leader prototype, the purpose of this study was to identify if and how racial components to the business leader prototype influenced outcomes for leaders. RLP3 tested two different hypotheses regarding the relationship between leader race and leadership evaluation. Grounded in theories of negative racial stereotypes, they tested the hypothesis that when a leader was blamed for organizational failure, the non-White leader would be evaluated less favorably than the White leader. Building on LCT, they also hypothesized that when a leader was credited for organizational success, the White leader would be evaluated more favorably than the non-White leader. Their tests showed support for the LCT hypothesis, but not for the negative stereotype hypothesis.

Study 2 included two close replications (Studies 2A and 2B) and one conceptual replication (Study 2C) of RLP3 in which we replicated these tests by having participants evaluate a fictional CEO after reading an article about the CEO's company in which the performance of the company, attributions for performance, and the race of the CEO were manipulated. The experiment took on a 2 (performance: successful, unsuccessful) X 2 (attributions for performance: CEO, marketplace) X 2 (CEO race: White, non-White) between-subjects design.

Study 2A

Research design. Participants completed the "Reading between the Lines" task described in RLP3 using a pen and paper. This task involved reading a fictitious newspaper article describing the performance of a corporation and the role that the CEO has played in the corporation's performance. The instructions in the participant packet stated "In this study, we will be examining the inferences people make after reading a newspaper article. Please review the following newspaper article and accompanying figure. When you are done reviewing these items, please answer the questions that follow." Organizational performance and attributions

were manipulated using an exact replica of the newspaper article shown in Appendix C of RLP (*see Appendix BIII*) along with the unpublished performance graphs used in RLP3.

Following RLP3, CEO race was manipulated using the name of the CEO in the article ("Todd Smith" = White CEO condition, "Tyrone Smith" = non-White CEO condition) along with CEO profile headshots. Pretests (*see Appendices AIX and BIV*) indicated that the headshot images used for manipulations were perceived as the intended races and did not differ in terms of perceptions of age ($p=0.24$), attractiveness ($p=0.38$), or emotional expression ($p=0.29$).

For the dependent variable of this study, we asked participants to evaluate CEOs on perceptions of intelligence, competence, confidence, and competitiveness by rating their level of agreement with statements such as "I think the CEO is intelligent" on a 7-point Likert-type scale. Consistent with RLP3, the mean of these ratings served as a composite leadership evaluation score (*Cronbach's* $\alpha=0.87$*; McDonald's* $\Omega=0.87$*; Guttman's lambda-2=0.87*). Manipulation checks for performance and attributions took place prior to the leadership assessment (*see Appendix BV*) and the manipulation check for CEO race was located "near the end of the post experimental questionnaire" (Rosette et al., 2008, p. 767). Finally, participants were asked to provide voluntary demographic information that included their race/ethnicity, age, gender, employment status, and employment history.

Participants. RLP3 utilized a student sample that had considerable racial/ethnic diversity. RLP's sample was comprised of 479 undergraduates who participated in exchange for either course credit or compensation of $10. The ratio of course credit-to-compensation used to incentivize participants in RLP3 was not published. To estimate our sample size for this replication, we conducted a power analysis using G*Power (*see Appendix AX*). The results of this estimate indicated that using a minimum sample size of 45 participants per condition, our

analysis could detect an effect size as small as *partial eta-squared = 0.05* with a power of 0.90. Using a sample size of 60 participants per condition, our analysis could detect an effect size as small as *partial eta-squared = 0.04* with a power of 0.90, thus this was set as our target sample size. As discussed in Study 1, no maximum sample size was set.

A total of 762 undergraduate students from a mid-sized research university in the northeastern United States participated in exchange for course credit (n=685) or $10 cash (n=77). Although Whites were not overrepresented when compared to population demographics, this sample had a higher proportion of White participants (56.61%) than RLP3 (27.35%).

Results. Our data analysis replicated the three-way analysis of variance (ANOVA) employed by RLP3 with $r_{contrast}$ effect size estimates (Rosnow et al., 2000). As shown in *Model 1 of Table 4,* the main effect of performance was significant, *F(1, 754)=497.63, p<0.0001, r=0.63,* indicating that our performance manipulation was effective. The main effect of leader race was also significant *F(1, 754)=4.21, p=0.04, r=0.07,* though the effect size was small. Surprisingly, however, this effect reflected higher ratings received by non-White leaders (*M=4.56, SD=1.12*) than those received by White leaders (*M=4.41, SD=1.17; see Table 5*).

Insert Tables 4&5 About Here

The foundational contribution of RLP3 came from the significance of a three-way interaction between performance, attributions, and leader race. This interaction was not significant in the current study (*p=0.07*). Interaction contrasts (Keppel, 1991) performed for robustness revealed that when comparing the effect of the two-way interaction of performance and attribution on leadership evaluations for White and non-White leaders, the difference once again, was not significant (*p=0.08).* Considering the absence of a significant three-way interaction, we proceeded cautiously with our planned contrasts.

Planned contrasts indicated that there were no differences in evaluations attributable to race in the high-performance/leader attribution condition (*p=0.95*), thus failing to support RLP3's primary finding and the hypothesis pertaining to LCT. Evaluations did not differ by race in the low-performance/marketplace attribution condition (*p=0.84*) either. Contrasts revealed that the non-White leader was evaluated more favorably than the White leader in the low-performance/leader attribution condition, *F(1, 754)=4.54, p=0.03.* This effect was in the opposite direction of what RLP3 predicted in regard to negative racial stereotypes, thus failing to support the argument that non-White leaders would be evaluated less favorably when the leader was blamed for organizational failure. This effect failed to remain significant after correcting p-values for multiple comparisons using Bonferroni's method (Field, 2012).[4] The difference in evaluation by leader race in the high-performance/marketplace attribution condition failed to meet the threshold for statistical significance (*p=0.09*).

Robustness. Per the research's pre-registration, we conducted several robustness tests using additional covariates or restricted subsamples. The main effect of race was only significant in three out of seven robustness tests, suggesting that the findings discussed above may be sensitive to model specification or loss of power (*see Appendix AXIII*). The White leader was not evaluated more favorably than the non-White leader in any of our robustness tests. Post-hoc analyses indicated that in the external attribution condition, participants who reported that they were employed at the time of the study provided lower evaluation scores to the White leader than

[4] The Bonferroni method adjusts p-values to account for multiple comparisons. The raw p-value is multiplied by the number of comparisons conducted. For example, after adjusting a p-value of 0.09 for four comparisons, the adjusted p-value would be 0.36. Adjusted p-values cannot exceed 1.00. This method for applying Bonferroni adjustments was consistently applied to post-hoc comparisons throughout this series of studies. For transparency, both values have been reported in Appendix AXII.

those provided by other participants. Additionally, participants who did not self-report their race as White provided lower evaluation scores to the White leader (*see Appendix AXII*).

Study 2B

The preregistration for this research study specified that, should the results of Study 2A fail to replicate those of RLP3, we would conduct a second round of data collection at an alternate institution to rule out the possibility that failure to replicate was due to unique characteristics associated with one particular data collection environment. This was done using the same materials as those described above.

Participants. Four-hundred forty-six students from a mid-sized liberal arts college in the northeastern United States participated in the current study. This sample had less racial diversity, but more work experience than that of Study 2A (*see Appendix AIV*). These differences were statistically significant (*ps<0.0001*).

Results. This study also utilized the same three-way ANOVA described above. Once again, the main effect of leader race was significant *F(1, 438)=6.45, p=0.01, r=0.12.* Consistent with the prior sample, this effect reflected higher ratings received by the non-White leader (*M=4.81, SD=1.10*) than those received by the White leader (*M=4.61, SD=1.08*).

The three-way interaction between performance, attributions, and leader race was not significant in the current study (*p=0.70*). Instead, we found a significant leader race X performance attribution interaction effect *F(1, 438)=4.22, p=.04, r=0.10.* Exploratory contrasts indicated that the White and non-White leaders received similar evaluation ratings when performance was attributed to the marketplace (*p=1.00*). When performance was attributed to the leader, however, the non-White leader was evaluated more favorably (*M=4.90, SD=1.25*) than the White leader (*M=4.54, SD=1.19*), *F(1, 438)=10.51, p=0.003.*

Robustness. The significant main effect of race was robust to a variety of model specifications, but the significance of the interaction of race and performance attribution was mildly sensitive to model specification (*see Appendix AXV*).

Study 2C

Research design. In Study 2C, we performed a conceptual replication of RLP3 by replicating RLP3 in a non-student sample. Study 2C incorporated experimental methodology mirroring that of Study 2A with one exception: data were collected using an electronic survey instead of paper packets. Recruitment materials stated that the purpose of the study was to measure how people interpret information.

Participants. Participants were recruited through email distribution lists maintained by multiple non-profit organizations in the northeastern United States (*n=494*). Participants were offered the option to enter a drawing for a $50 gift card upon completion of the study. Just over half (53.24%) of participants opted to participate in the drawing. Compared with the samples of the previous studies, the sample had little racial diversity (91.17% of those self-reporting race identified as White), was older (Mean=48.65 years), and had more work experience (Mean=26.32 years; *see Appendix AIV*).

Results. Our 2X2X2 ANOVA indicated, similar to Studies 2A and 2B, the main effect of leader race was significant *F(1, 486)=8.42, p=0.004, r=0.13* (*see Table 4, Model 3*), as non-White leaders (*M=4.62, SD=1.08*) received higher evaluations than White leaders (*M=4.37, SD=1.00*) in the current study. Neither the three-way interaction of performance, performance attributions, and leader race (*p=0.76*), nor the two-way interaction of performance attribution and leader race (*p=0.07*) were significant. Exploratory interaction contrasts, however, indicated that evaluation ratings given by participants in the White leader, marketplace attribution

condition (*M=4.19, SD=0.70*) were lower than evaluation ratings given by participants in all other conditions (*M=4.61, SD=1.14*), $F(1, 486)=26.20$, $p<0.0001$.

Robustness. Robustness tests provided similar results to those described above, though the effect of the interaction of attribution and race was mildly sensitive to model specification (*see Appendix AXVII*).

STUDY 3

Our final set of studies serve as a conceptual replication of RLP3. Their first purpose was to examine if the relationships identified in our Studies 2A-2C were also present when using participants recruited through mTurk. Examining the replicability of our findings within the context of mTurk provides insight as to how the growing use of mTurk in organizational studies (e.g. Marchiondo, Myers, & Kopelman, 2015; O'Reilly, Doerr, & Chatman, 2017; Schaumberg & Flynn, 2017; Tucker, Ogunfowora, & Ehr, 2016) is influencing our understanding of leadership and discrimination theories.

The secondary purpose of this conceptual replication is to examine the degree to which decisions made in the experimental design process can influence our understanding of leadership models. As discussed above, choices in names and photographs used for racial manipulations, along with the location of manipulation checks, may influence how participants respond in a study. Whereas variations in experimental designs sometimes allow researchers to infer that an effect is robust to design specifications (typically in terms of significance and direction), Study 3 systematically altered components of research design such that we could examine the effects of each modification and provide insight as to how variations in experimental design influence the replicability of leadership research. In studies 3A-3C, we isolated the effects of design choices in

leader name, leader photograph, and manipulation check location. Study 3D utilized data from studies 2A, 2C, and 3A to examine sampling effects.

Participants in Studies 3A-3C were restricted to mTurk users in the United States, received $0.25 for successful completion of the task, were entered into a drawing to win one of two $50 bonuses that were available per study, and were restricted to participating in one study. Successful completion was defined as spending at least thirty seconds on the task and correctly responding to at least three out of four attention checks. Data provided by individuals who did not meet these criteria were discarded (36, 40, and 34 responses in Studies 3A, 3B, and 3C, respectively). Like Study 1B, attention checks differed from manipulation checks in that while manipulation checks were designed to capture participants' perceptions of manipulations, attention checks captured responses to details that were explicitly stated in the case.

Study 3A

This study addressed the possibility that names used as racial manipulations in experiments could be communicating information in addition to race, thus introducing "noise" into the model and that, consequently, choices in names used as experimental manipulations have the potential to influence our understanding of LCT. This study was motivated by the possibility that differences in the population frequency of first names used to manipulate race in experimental studies may introduce the confounding variable of familiarity bias. Young, Kennedy, Newhouse, Browne, and Thiessen (1993) found that individuals with uncommon names were perceived as less intelligent than those with more common names. Building upon this line of research, Cotton, O'Neill, and Griffin (2008) reported that unique, unfamiliar names were less likely to elicit an employment offer than names that were utilized frequently in the overall population.

These findings are particularly relevant to the employment discrimination literature as distinctly White names are more common in the overall population of the United States than distinctly non-White names (e.g. Black names). For example, the name Todd, which is frequently used as a manipulation for White names (e.g. Bertrand & Mullainathan, 2004; Rosette et al., 2008), has appeared in birth records twelve times as often as the name Tyrone (DPH, 1979), which is frequently used as a manipulation for Black names. Consequently, first names used to manipulate race in experiments may also be capturing familiarity bias, potentially amplifying the effects of reported racial bias. This potential confounding variable of concern should not be perceived as invalidating findings of studies that use names as racial manipulations, but it could have implications for the applicability of these findings. For example, a Black leader named Tyrone may face biases that differ significantly from those faced by a Black leader named James.

Research design. We addressed the potential confounding factor of familiarity bias by removing that which is familiar through the introduction of a manipulation for the presence of a first name. In the condition where a first name was present, there were no changes to the manipulation script. In the condition where no first name was present, the first instance of "Todd [Tyrone] Smith" was be replaced by "T. Smith." The CEO was simply referred to as Smith in further references to his name. The CEO's profile headshot served as the only racial manipulation in this condition.

The addition of the condition described above created a 2 (performance: successful, unsuccessful) X 2 (attributions for performance: CEO, marketplace) X 2 (CEO race: White, non-White) X 2 (research design: first name present, first name not present) between-subjects experimental design. Our power analysis (*see Appendix AX*) indicated that by maintaining our

target sample size of 60 participants per condition, we could detect an effect size as small as partial eta-squared = 0.03 with a power of 0.90.

Participants. This sample consisted of 965 participants. Most participants identified as White (73.13%). The mean reported age of participants was 36.61 years with average work experience of 15.75 years (*see Appendix AIV*).

Results. Consistent with prior studies, the main effect of leader race was significant, $F(1, 949)=28.84$, $p<0.0001$, $r=0.17$, though the effect size was small. Once again, the non-White leader was evaluated more favorably ($M=5.32$, $SD=1.38$) than the White leader ($M=4.95$, $SD=1.44$). As shown in *Model 4* of *Table 4*, other results were consistent with Study 2 as well.

The main effect of our research design manipulation was not significant ($p=0.91$). Because the research design manipulation did not have any significant interactions with leader race, these findings suggest that whether a first name was used in the experimental manipulation did not influence findings relevant to our research question.

Robustness tests. Per the study's registered report, we conducted robustness tests consistent with those of the previous studies. These tests indicated that the results reported for Study 3A were robust to model specification (*see Appendix AXIX*).

Study 3B

The design concern addressed in Study 3B was the impact of additional information that is communicated by photographs used for racial manipulations on research findings that influence our understanding of LCT. Although researchers tend to concur that photographs have to the potential to communicate information beyond that of demographic characteristics, the process of addressing these concerns has not yet been standardized. For example, Hekman et al. (2017) did not specify pretests used to select photos for their study; Zhu et al. (2016) did not

pretest photos used in their first study, but did pretest another set on perceptions of physical attractiveness; RLP pretested photos on measures of physical attractiveness, emotional expression, and perceptions of age; Younkin and Kuppuswamy (2017) pretested photos on perceptions of age, attractiveness, emotion, trustworthiness, and credibility; and multiple studies have used photographs created through software such as FaceGen (Gladstone & O'Connor, 2014; K. L. Johnson et al., 2012). Whereas it is clear that researchers have incorporated a variety of methods for addressing factors that can confound the racial effects captured by photographs, there is still a lack of clarity regarding the *impact* of choosing one solution over another in experimental design.

Research design. We began to explore the impact of experimental design choices regarding photographs used for racial manipulations through the comparison of pre-tested unmodified photographs and morphed photographs. In the pre-tested photograph condition, the race of the leader was manipulated using the same photographs that were used in Studies 2 and 3A. In the morphed photograph condition, we manipulated the race of the leader using FaceGen to modify the race of an individual in a photograph that was distinct from those used in previous studies (*see Appendix BVI for further details*). FaceGen is a software designed to modify demographic characteristics of individuals in photographs without modifying facial expressions or other aspects of facial construction. Its use in laboratory research has grown in recent years (e.g. Gladstone & O'Connor, 2014; Johnson, Freeman, & Pauker, 2012).

Participant recruitment and data analysis followed the same procedures outlined in Study 3A. This study took on a 2 (performance: successful, unsuccessful) X 2 (attributions for performance: CEO, marketplace) X 2 (CEO race: White, non-White) X 2 (research design: pre-tested photo; FaceGen photo) between-subjects experimental design.

Participants. One thousand individuals with demographic characteristics similar to those of the other mTurk samples (*see Appendix AIV)* participated in this study.

Results. Unlike the prior studies, the main effect of leader race did not have a significant relationship with leadership evaluations (*p=0.25*). The main effect of research design did not have a significant relationship with the dependent variable (*p=0.74*). The interaction of interest (design manipulation X leader race) was not significant (*p=0.056*) either. Exploratory post-hoc analyses using Bonferroni corrections indicated that the difference between the evaluation of White and non-White leaders was neither significant in the pretested pictures condition (*p=0.06*), nor was it significant in the FaceGen condition (*p=1.00*). Similarly, interactions between research design and other variables did not have a significant relationship with the dependent variable (*ps ranged from 0.34 to 0.97*).

Robustness tests. These findings were fairly robust to model specification although there was limited evidence that when the sample was restricted to participants who correctly responded to all three manipulation checks, the type of picture used may have influenced the relationship between leader race and evaluations (*see Appendices AXII & AXXI*). There was also some evidence that participant race influenced the relationship between a leader's race and evaluation, though the interaction effect sizes were small and none of the exploratory contrasts related to our variables of interest yielded significant differences (*ps ranged from 0.69 to 1.00*).

Study 3C

Study 3C investigated how a participant's ability to return to a questionnaire after completing manipulation checks would influence our understanding of LCT. This is an important question as there is not a universal standard regarding how demographic manipulation checks should be included in experimental materials. Demographic manipulation checks can occur

before participants engage with the dependent variable (e.g. Brescoll, 2011; Motro & Ellis, 2017; Rosette et al., 2008), in a location that is not disclosed in the manuscript (e.g. Biernat & Kobrynowicz, 1997; S. K. Johnson et al., 2008; Kaufmann et al., 2016; Ziegert & Hanges, 2005), or near the end of subjects' participation in a study, although even in these cases, there can be ambiguity regarding whether or not participants can alter questionnaire response after completing the manipulation check (e.g. Biernat & Kobrynowicz, 1997; Hernandez et al., 2016; Rosette et al., 2008).

Although the inconsistent placement of manipulation checks is not unique to this domain, it does pose challenges regarding a fundamental interpretation of this body of research as the location of a manipulation check can influence participants' responses to variables of interest by informing them of the true purpose of the study (Singleton Jr. & Straits, 2005). There is a need, however, to more clearly understand *how* a participant's ability to alter their responses after being exposed to manipulation checks will influence responses. Whereas there is evidence that awareness of a study's true purpose can result in demand characteristics (Nichols & Maner, 2008; providing responses that support research hypotheses), awareness of a study's true purpose could also motivate participants to behave in an unbiased or pro-minority manner (E. B. King et al., 2013; Stone et al., 2008). Building on this argument, research designs that allow participants to alter their data after being exposed to the manipulation check may inform participants of a study's true intent, resulting in more favorable ratings of minority leaders. This concern is highly relevant to the current research, given the more favorable ratings reported for the non-White leader in previous studies.

Research design. To examine the effect of manipulation location on results, this study incorporated two different manipulation configurations. The first configuration mirrored that of

RLP3, such that the race manipulation check was present in the end of the study, but participants were not restricted from returning to the questionnaire after seeing the manipulation check[5]. The second condition did not include a manipulation check for race. From a theoretical standpoint, this condition is identical to completing a manipulation check at the end of the experiment, when all variables in the study have been recorded and cannot be changed. This resulted in a 2 (performance: successful, unsuccessful) X 2 (attributions for performance: CEO, marketplace) X 2 (CEO race: White, non-White) X 2 (research design: race manipulation check present, race manipulation check not present) between-subjects experimental design. Participant recruitment and data analysis followed the same procedures outlined in Studies 3A and 3B.

Participants. This study included 1,005 participants who reported similar demographic traits to the previous samples collected through mTurk (*see Appendix AIV*).

Results. Similar to Study 3B, the main effect of leader race was not significant (*p=0.14*), though leader race did have significant interactions with performance, *F(1, 989)=8.20, p=0.004, r=0.09*, and performance attributions, *F(1, 989)=4.50, p=0.03, r=0.07*. Neither our design manipulation nor any of its interactions with leader race influenced leadership evaluations (*ps ranged from 0.42 to 0.89*), indicating that the presence of a manipulation check did not influence our findings related to the relationship between a leader's race and leadership evaluations.

Exploratory post hoc analyses using Bonferroni corrections indicated that when performance was attributed to the marketplace, the non-White leader received higher evaluation ratings (*M=5.39, SD=1.22*) than the White leader (*M=5.14, SD=1.14*), *F(1, 989)=6.55, p=0.04,* but that no such difference was present when performance was attributed to the leader (*p=1.00*). Additionally, in the higher performance condition, the non-White leader received higher

[5] This was not specified in the paper but was verified with the original authors

performance ratings (*M=6.17, SD=0.88*) than the White leader (*M=5.89, SD=0.93*), *F(1,989)=9.41, p=0.009*. There was no significant difference in ratings between White and non-White leaders in the lower performance condition (*p=1.00*).

Robustness tests. Primary findings were robust to model specification, though robustness checks indicated that the model was somewhat sensitive to whether the participant was currently employed (*see Appendix AXXIII*).

Study 3D

Our final analysis explored the possibility that the type of sample used to collect data could influence our understanding of LCT. Organizational diversity research derived from experimental studies frequently utilizes student participants (e.g. S. K. Johnson et al., 2008; Livingston & Pearce, 2009; Rosette et al., 2008; Zapata et al., 2016; Ziegert & Hanges, 2005). Although the use of student samples in a laboratory setting provides researchers with some benefits, such as tightly controlled conditions where treatment effects can be more effectively isolated, it may also reduce the generalizability and replicability of findings (Singleton Jr. & Straits, 2005). Barr and Hitt (1986) found evidence that student participants evaluated job candidates more favorably than working professionals did. These findings were supported by those of Singer and Bruhns (1991) who found differences between student and professional samples even after controlling for work experience. Differences between student and professional samples may be particularly relevant when experimental designs manipulate assessments provided by industry experts, as done in RLP and the current replication studies. Specifically, social science research has shown that college students have malleable opinions and are likely to concur with perceived authority figures such as industry experts (Sears, 1986).

In addition to having the potential to influence the outcomes of organizational studies, sample selection is relevant to diversity and inclusion research. Because university environments tend to be relatively liberal in nature, diversity and inclusion research conducted with student samples may be influenced by increased pressures for participants to control displays of prejudice (Henry, 2008). This statement, however, is contradicted by findings from Koch, D'Mello, and Sackett's (2015) meta-analysis which indicated that experienced professionals exhibited less bias in organizational decision making than undergraduate students did, thus motivating further tests of the effect of sample selection on findings of discrimination.

Whereas replications in different samples (e.g. RLP4's conceptual replication of RLP3 utilized graduate students instead of undergraduate students) provide some insight into the generalizability of findings, in the absence of statistical comparisons of studies, these replications do not allow us to isolate the effects of using different samples. This is addressed in the current study through statistical analysis that integrates data from a student sample, the professional sample, and an mTurk sample. We recognize that across studies, there may be differences in addition to the composition of the sample. For example, participants in Study 3 were incentivized differently from participants in Study 2, were required to complete multiple attention checks, and were likely to be completing the study in a physical environment that differs greatly from that of the participants in Study 2. Consequently, any differences across samples cannot be solely attributed to sample composition, though they can be reasonably attributed to the method of data collection (e.g. recruiting student samples versus recruiting mTurk samples).

Research design. In order to compare the results from different samples, we integrated and analyzed data from Studies 2 and 3. Per the procedure described in our registered research report, the first step in data integration was to extract mTurk data from conditions in which a first

name (Todd or Tyrone) was used as racial manipulation from Study 3A. The extracted mTurk data was then integrated with the data from Studies 2A and 2C. We then executed the ANOVA described in the previous studies, introducing a variable for research design and interaction terms resulting in a 2 (performance: successful, unsuccessful) X 2 (attributions for performance: CEO, marketplace) X 2 (CEO race: White, non-White) X 3 (research design: student sample, professional sample, mTurk) design. The resulting sample included 1,743 participants which exceeded both our minimum sample size of 1,080 and our target sample size of 1,440. Our primary interest in this analysis was to investigate how sample selection influenced the relationship between race and leadership evaluations.

Results. As shown in Model 7 of *Table 4*, the main effect of leader race was significant *F(1, 1719)=28.43, p<0.0001, r=0.13.* Research design did have a significant effect on leadership evaluations, *F(2, 1719)=87.45, p<0.0001, r=0.22,* as exploratory post hoc analyses indicated that mTurk participants rated the leader more favorably than participants from the student or professional samples *F(1, 1719)=171.46, p<0.0001.* The interaction of leader race and research design was not significant (*p=0.087*). No other interaction terms involving race and research design were significant either (*ps ranged from 0.36 to 0.96*).

DISCUSSION

Summary of findings

The business leadership prototype. In Study 1, we found minimal support for the argument that race is a component of the business leader prototype as the relationship between an individual's employment position (leader vs. non-leader) and perceived race was only significant by conventional standards in one robustness test. Our z-tests of proportion indicated

that both leaders and non-leaders were perceived as White at a higher rate than the base rates provided would suggest.

Evaluations of leaders. All six of our replications failed to replicate RLP3's finding that the White leader was evaluated more favorably than the non-White leader when credit for successful performance was given to the leader. In fact, the mean evaluation rating was higher for the non-White leader than it was for the White leader in all six studies and this difference was statistically significant in four studies. In one study in which the main effect of leader race was not significant, the main effect of performance was qualified by an interaction with leader race, such that the non-White leader was evaluated more favorably than the White leader in the high-performance condition. Collectively, we report that the non-White leader received higher evaluation ratings than the White leader, although the effect size was consistently small.

Replication of prior findings and theoretical contributions

Race as a component of the business leader prototype. RLP's contribution to LCT was built upon the core findings that leaders were more likely to be perceived as White than non-leaders were, and that unlike non-leaders, leaders would be perceived as White more often than demographic base rates would suggest. Our failure to replicate these findings was driven by the fact that the *non-leader* was perceived as White at higher rates than base rates would suggest in the current research. This finding was robust to model specification. In light of the minimal similarities and considerable differences to the findings of RLP1, these results may be seen as providing minimal evidence that being White is a component of the business leader prototype. These findings, however, do not necessarily reflect a change in the business leader prototype as a change in the prototype should have resulted in a lower percentage of leaders being perceived as White rather than a higher percentage of non-leaders being perceived as White.

One possible explanation for the divergence in findings from RLP1, in terms of perceptions of the non-leader, is that there may have been other prototypes at work in this research. Although the vignette provided no deliberate indication of the interviewee's industry, when asked their perception of the industry, the most common responses were education (*Study 1A=28.67%, Study 1B=23.90*) and engineering (*Study 1A=25.63%, Study 1B=31.12%*), two industries in which Whites are overrepresented (U.S. Bureau of Labor Statistics, 2017). Perhaps participants were drawing upon previously formed prototypes for workers in these industries.

Similarly, prior work suggests that demographic traits may play a role in the development of spokesperson prototypes (Ryan et al., 2011). Given the systematic inequalities that have led to Whites holding a disproportionate share of visible positions in organizations (Bell & Hartmann, 2007), it is likely that in addition to being a component of the business leader prototype, being White could be a component of a spokesperson prototype. If participants drew upon this type of spokesperson prototype, it would have increased their probability of perceiving the non-leader as White, thus limiting the potential for the effects of *leadership* categorization to be detectable.

Building on prior work that has identified the contextual nature of prototype relevance (e.g. Braun et al., 2018; Kocoglu & Mithani, 2020; Sy et al., 2010), the findings of Study 1 in the current research serve as strong motivation for researchers to more closely examine not only if being White is still a component of the business leader prototype, but under what conditions this prototype is most influential. Future research could also examine how the activation of alternative prototypes interacts with the activation of leadership prototypes to influence perceptions of leader prototypicality.

Factors influencing evaluations of leaders. Our findings diverged from those of RLP3 in that non-White leaders consistently received marginally higher evaluation ratings than White

leaders in the current research. One possible explanation for these findings can be derived by continuing to consider our outcomes through the lens of LCT. Our failure to replicate the racial effects illustrated in RLP3 is consistent with the summary statistics provided in the recent work of Gündemir et al. (2019) and Reynolds et al. (2021), as well as the primary findings of emerging research (e.g. Ubaka et al., 2020, *citation removed to preserve blind review*). What each of these studies has in common is that participants engaged in experimental vignettes in which they would have the opportunity to carefully consider their responses, thus increasing the likelihood of engaging in controlled processing (Lord & Maher, 1993; Lord & Smith, 1983).

If engaged in controlled processing, participants *may* have drawn upon their desire to behave in an unprejudiced manner (Crandall & Eshleman, 2003; Dovidio & Gaertner, 2004). While this would not have resulted in altered ratings for the White leader, imperfect efforts to control prejudice could have resulted in the non-White leader receiving marginally higher ratings than the White leader. Such an explanation allows us to reconcile these findings with evidence from the field that non-White leaders (e.g. Black leaders) continue to face discrimination in terms of assessment and advancement (Roberts & Mayo, 2019). Unlike participants in experimental vignettes, actors in the real world are faced with multiple competing demands for their time and attention. These competing demands can cause individuals to engage in automatic processes that rely on the application of previously encoded prototypes (Lord & Maher, 1993).

To understand why participants *may* have engaged in controlled processing in the current research, but not in RLP3, we must recognize that increased perceptions of racial discrimination (Jones & Saad, 2016; Pew Research Center, 2016) *may* have increased awareness of racial cues in research studies such that participants recognize that their responses to racial stimuli are being deliberately observed. An LCT-grounded explanation for our divergence in findings from those

of RLP3 might suggest that a recently increased awareness of issues pertaining to racial equity has the potential to cause current research participants to engage in greater levels of controlled processing than participants in the past. Consequently, our findings *may* not reflect a change in the leadership categorization process or its implications. Rather, they *could* be evidence that the conditions in which participants engage in automatic processing and draw upon prototypes have changed. Future research should explore this possibility as data in the current research do not allow us to isolate mechanisms related to automatic and controlled processing.

It is also important to note that failure to replicate the racial effects described in RLP3 may simply be a result of findings having low reliability. Prior work has illustrated patterns whereas effect sizes in replications are half of those found in target studies (Open Science Collaboration, 2015). As effect sizes are negatively correlated with p-values, findings with small effect sizes should be most susceptible to failed replications. The three-way interaction in RLP3 falls into this category as does the main effect of race in the current research (which we failed to replicate in our final two studies). This insight suggests that findings with small effect sizes should undergo repeated replications before being used to draw strong theoretical inferences. Consequently, any inferences about LCT drawn from the current research may be strengthened by additional replications of these findings in new samples.

Reconciling perceptions of interviewee race and leadership evaluations. Next, we reconcile Study 1 with the seemingly incompatible results of Studies 2 and 3. Although the results of Study 1 suggest a bias in favor of White leaders/spokespeople, White leaders did not experience more favorable evaluations in Studies 2 and 3. One *possible* explanation for this is that the design of Study 1 allowed participants to confront bias in a different way from Studies 2and 3. Participants in Study 1 were essentially being asked to provide an estimate as it may be

influenced by contextual factors, such as societal bias. Given current social movements to recognize and respond to racial bias in society (Kendi, 2019), participants *may* have been comfortable estimating that a leader/spokesperson was likely to be White because of societal inequities. Studies 2 and 3, however, required participants to make decisions based upon their own interpretations and potentially confront their own biases. In these cases, they *may* have been motivated to behave in a socially desirable manner (Dovidio & Gaertner, 2000; E. B. King et al., 2013; Stone et al., 2008). Consequently, the differential roles of bias recognition across study formats *may* have influenced the compatibility of results.

Furthermore, as LCT research has advanced, prototypes have become conceptualized as more contextual in nature (Braunn et al., 2018; Gündemir et al., 2019; Sy et al., 2010). This is relevant to the current research as the context of the vignette used in Study 1 was portrayed as stable while the context of the vignette used in Studies 2 and 3 was portrayed as unstable. Given the contextual nature of prototypes, it is *possible* that any prototypes identified in the stable context of Study 1 were not relevant to the unstable context described in Studies 2 and 3. This highlights the need for researchers to consider the contextual compatibility of vignettes in future LCT research.

Methodological contributions

Name manipulations. Study 3A indicated that the presence of a first name in the racial manipulation did not significantly influence findings. This should temper concerns that research designs relying on first names to manipulate race (e.g. Bertrand & Mullainathan, 2004; Zapata et al., 2016; Zhu et al., 2016) could be confounded by factors such as familiarity bias.

Photograph manipulations. Prior research has not established a uniform process for selecting photographs used as racial manipulations (e.g. Hekman et al., 2017; Rosette et al.,

2008; Zhu et al., 2016). As photographs have the potential to communicate additional information beyond that which is intended by researchers, Study 3B was designed to examine the effect that the process of selecting photographs used for racial manipulation has on research findings. Our results yielded no meaningful differences between the responses of participants exposed to pretested photos and those exposed to morphed (FaceGen) images.

Manipulation checks. Recognizing that manipulation checks can inform participants of a study's true purpose (Singleton Jr. & Straits, 2005) and potentially lead to demand effects (Nichols & Maner, 2008), Study 3C was designed to examine how the presence of a manipulation check influenced participant responses. Results of Study 3C indicated that the presence of a manipulation check had little influence on our findings. At face value, this would suggest including a manipulation check at the end of an experiment's questionnaire, even if the participant still has the ability to return to the survey and modify other responses, poses minimal risk to the study's validity.

Our robustness tests illustrated, however, that decisions regarding how researchers *respond to* manipulation checks can influence our understanding of theory. In both replications of RLP1, we failed to fully replicate the significant relationship between interviewee position and perceived race of the interviewee that was present in the original study. When restricting the sample to participants who responded to manipulation checks correctly, this effect was significant in the student sample. Consequently, our interpretation as to the role of race in the business leader prototype was directly influenced by how we, as researchers, responded to manipulation checks.

Sample selection. One of the strengths of the current research is that it involved three different sample types: student, professional, and mTurk. Our analyses indicated that there were

no significant differences in our primary findings based upon the type of sample used. The most notable difference was that regardless of leader race, participants from all three mTurk samples evaluated the leader more favorably than participants from the other samples. Whereas this finding does not have direct implications for LCT, it suggests that researchers should account for sample source when comparing data collected through mTurk with data from other sources.

Limitations

Although post-hoc power analyses for Studies 1A and 1B did not indicate that either study was underpowered, they did indicate that detecting a significant relationship between interviewee position and perceived race would require approximately 90 percent of participants to have perceived the leader as White (*see Appendix BVII*). This requirement could raise some concerns that our findings may have been influenced by ceiling effects. If a ceiling effect was preventing detection of significant relationships, however, this should have also prevented the base rate manipulation from having a significant relationship with perceived race. Additionally, a ceiling effect should not have been applicable the 20 percent White base rate condition where only 70 percent of participants perceived the interviewee as White. Our analyses, however, showed no evidence that the effect of the interviewee's employment position on perceived race differed in this condition, thus assuaging concerns regarding potential ceiling effects.

Another potential limitation could be derived from evidence that the non-White CEO image used in Studies 2 and 3 may have conveyed higher levels of trustworthiness than the White CEO image (*see Appendices AIX & BIV*) and that the small sample size used in our photograph t-tests preclude us from completely ruling out that other differences were present. These concerns, however, were somewhat assuaged by the fact that, as discussed above, we did not find strong evidence that the image selection process was a driving factor behind our results.

We also considered the possibility that participants recruited through college campuses (Henry, 2008), and mTurk (Huff & Tingley, 2015) have a high likelihood of embracing liberal ideology. These concerns should have been addressed by Study 2C, however, as the majority of participants in this sample reported working in traditionally conservative (e.g. management) positions (Cheng & Groysberg, 2016; Depillis, 2019). Future research could further address this concern by examining role of political ideology on automatic processes that influence differential outcomes.

Finally, one may argue that our findings could have been influenced by the fact that our participants did not have a vested interest in outcomes related to their decisions. Researchers have argued that questionnaires may lack external validity (Aguinis & Bradley, 2014; M. F. King & Bruner, 2000) and that field experiments may be the most appropriate methodology for examining certain sensitive topics (E. B. King et al., 2013; Wulff & Villadsen, 2020). Because our methodology replicated RLP, however, this lack of a vested interest in outcomes is not unique to the current research. To better understand the impact of participants' interests on research outcomes, future research may investigate whether experimental work using consequential decisions leads to similar findings.

Conclusion

To find no evidence of racial discrimination in leadership evaluations in six different U.S. samples during a time period in which racial inequities in the United States have been highly salient is counterintuitive to say the least. Accordingly, it is quite possible that this research raises more questions than it answers. Although consistent results across six studies in three different types of samples indicate experimental reliability, evidence of continued racial discrimination in the real world suggests that our findings may have limited generalizability

outside of the experimental setting. It is plausible that changes in the social climate and an increased awareness of racial inequities may cause participants to behave differently in contexts where they are carefully analyzing their choices, such as when completing a questionnaire for scientific research. If this is the case, it has the potential to challenge the external validity of experiments that rely on salient racial manipulations moving forward. Perhaps the most promising application of this research would be to identify the contextual factors that influenced null findings and small effect sizes in this research so that they can be replicated in organizations, thus leading to more equitable outcomes in real-world applications.

REFERENCES

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, *17*(4), 351–371. https://doi.org/10.1177/1094428114547952

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economic Letters*, *80*(1), 123–129.

Aronow, P. M., Baron, J., & Pinson, L. (2019). A note on dropping experimental subjects who fail a manipulation check. *Political Analysis*, *27*(4), 572–589. https://doi.org/10.1017/pan.2019.5

Barr, S. H., & Hitt, M. A. (1986). A comparison of selection decision models in manager versus student samples. *Personnel Psychology*, *39*(3), 599–617. https://doi.org/10.1111/j.1744-6570.1986.tb00955.x

Bell, J. M., & Hartmann, D. (2007). Diversity in everyday discourse: The cultural ambiguities and consequences of "happy talk." *American Sociological Review*, *72*(6), 895–914. https://doi.org/10.1177/000312240707200603

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

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. https://doi.org/10.1037/0022-3514.72.3.544

Bishop, D. V. M. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, *73*(1), 1–19. https://doi.org/10.1177/1747021819886519

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*(1), 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Braunn, S., Peus, C., & Frey, D. (2018). Connectionism in action: Exploring the links between leader prototypes, leader gender, and perceptions of authentic leadership. *Organizational Behavior and Human Decision Processes*, *149*, 129–144.

Brescoll, V. L. (2011). Who takes the floor and why: Gender, power, and volubility in organizations. *Administrative Science Quarterly*, *56*(4), 622–641. https://doi.org/10.1177/0001839212439994

Brief, A. P., Dietz, J., Cohen, R. R., Pugh, S. D., & Vaslow, J. B. (2000). Just doing business : Modern racism and obedience to authority as explanations for employment discrimination. *Organizational Behavior and Human Decision Processes*, *81*(1), 72–97. https://doi.org/10.1006/obhd.1999.2867

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Supplementary information for evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, *2*(9), 637–644. https://doi.org/https://www.nature.com/articles/s41562-018-0399-z

Carton, A. M., & Rosette, A. S. (2011). Explaining bias against black leaders: Integrating theory on information processing and goal-based stereotyping. *Academy of Management Journal*, *54*(6), 1141–1158. https://doi.org/10.5465/amj.2009.0745

Chambers, C. (2019). What's next for registered reports? *Nature*, *573*, 187–189.

Chatla, S., & Shmueli, G. (2016). Linear probability models (LPM) and big data : The good, the bad, and the ugly. *Indian School of Business Research Paper Series*. http://ssrn.com/abstract=2353841

Cheng, J. Y.-J., & Groysberg, B. (2016). 7 charts show how political affiliation shapes U.S. boards. *Harvard Business Review*, *23*(7).

Clapp-Smith, R., Carsten, M., Gooty, J., Connelly, S., Haslam, A., Bastardoz, N., & Spain, S. (2018). Call for papers - Special registered report issue on replication and rigorous retesting of leadership models. *The Leadership Quarterly*, *29*(2), V–VI. https://doi.org/10.1016/s1048-9843(18)30146-2

Cook, A., & Glass, C. (2014a). Above the glass ceiling: When are women and racial/ethnic minorities promoted to CEO. *Strategic Management Journal*, *35*(7), 1080–1089. https://doi.org/10.1002/smj

Cook, A., & Glass, C. (2014b). Analyzing promotions of racial/ethnic minority CEOs. *Journal of Managerial Psychology*, *29*(4), 440–454. https://doi.org/10.1108/JMP-02-2012-0066

Cotton, J. L., O'Neill, B. S., & Griffin, A. (2008). The "Name Game": Affective and hiring reactions to first names. *Journal of Managerial Psychology*, *23*(1), 18–39. https://doi.org/10.1108/02683940810849648

Cox Jr., T., & Nkomo, S. M. (1986). Differential performance appraisal criteria: A field study of black and white managers. *Group and Organization Studies*, *11*(1–2), 101–119.

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin*, *129*(3), 414–446. https://doi.org/10.1037/0033-2909.129.3.414

Deloitte, & Alliance for Board Diversity. (2017). *Missing pieces report: The 2016 board diversity census of women and minorities on Fortune 500 boards*. https://www2.deloitte.com/us/en/pages/center-for-board-effectiveness/articles/board-diversity-census-missing-pieces.html

Depillis, L. (2019). *Way more CEOs are Republicans than Democrats. Here's the proof*. CNN. https://www.cnn.com/2019/05/14/business/republican-democrat-ceos/index.html

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions. *Psychological*

Science, *11*(4), 315–319.

Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (36th ed., pp. 1–52). Elsevier Academic Press.

DPH. (1979). *Birth name count*. State of Massachusetts Department of Public Health.

Field, A. (2012). *Discovering statistics using IBM SPSS Statistics* (M. Carmichael (ed.); Fourth). Sage Publications.

Gladstone, E., & O'Connor, K. M. (2014). A counterpart's feminine face signals cooperativeness and encourages negotiators to compete. *Organizational Behavior and Human Decision Processes*, *125*(1), 18–25. https://doi.org/10.1016/j.obhdp.2014.05.001

Greenhaus, J. H., & Parasuraman, S. (1993). Job performance attributions and career advancement prospects: An examination of gender and race effects. *Organizational Behavior and Human Decision Processes*, *55*, 273–297. https://doi.org/10.1006/obhd.1993.1034

Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of race on organizational experiences, job performance evaluations, and career outcomes. *Academy of Management Journal*, *33*(1), 64–86. https://doi.org/10.2307/256352

Gündemir, S., Carton, A. M., & Homan, A. C. (2019). The impact of organizational performance on the emergence of Asian American leaders. *Journal of Applied Psychology*, *104*(1), 107–122. https://doi.org/10.1037/apl0000347

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers : MTurk participants perform better on online attention checks than subject pool participants. *Behavior Research Methods*, *48*, 400–407. https://doi.org/10.3758/s13428-015-0578-z

Hekman, D. R., Johnson, S. K., Foo, M.-D., & Yang, W. (2017). Does diversity-valuing behavior result in diminished performance ratings for nonwhite and female leaders? *Academy of Management Journal*, *60*(2), 771–797.

Henry, P. . (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, *19*, 49–71. https://doi.org/10.1037/0022-3514.51.3.515

Hernandez, M., Avery, D. R., Tonidandel, S., Hebl, M. R., Smith, A. N., & McKay, P. F. (2016). The role of proximal social contexts: Assessing stigma-by-association effects on leader. *Journal of Applied Psychology*, *101*(1), 68–85. https://doi.org/10.1037/apl0000030

Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research and Politics*, *2*(3). https://doi.org/10.1177/2053168015604648

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, *66*, 81–92. https://doi.org/10.1016/j.jesp.2015.09.009

James, E. H. (2000). Race-related differences in promotions and support: Underlying effects of

human and social capital. *Organization Science*, *11*(5), 493–508. https://doi.org/10.1287/orsc.11.5.493.15202

Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology*, *102*(1), 116–131. https://doi.org/10.1037/a0025335

Johnson, S. K., Murphy, S. E., Zewdie, S., & Reichard, R. J. (2008). The strong, sensitive type: Effects of gender stereotypes and leadership prototypes on the evaluation of male and female leaders. *Organizational Behavior and Human Decision Processes*, *106*(1), 39–60. https://doi.org/10.1016/j.obhdp.2007.12.002

Jones, J., & Saad, L. (2016). Gallup poll series: Minority rights & relations. *Gallup News Service*.

Kaufmann, M. C., Krings, F., & Sczesny, S. (2016). Looking too old? How an older age appearance reduces chances of being hired. *British Journal of Management*, *27*(4), 727–739.

Kendi, I. X. (2019). *How to be an anitracist*. One World.

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Prentice-Hall.

King, E. B., Hebl, M. R., Botsford Morgan, W., & Ahmad, A. S. (2013). Field experiments on sensitive organizational topics. *Organizational Research Methods*, *16*(4), 501–521. https://doi.org/10.1177/1094428112462608

King, M. F., & Bruner, G. C. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychology and Marketing*, *17*(2), 79–103. https://doi.org/10.1002/(SICI)1520-6793(200002)17:2<79::AID-MAR2>3.0.CO;2-0

Kleinbaum, D. G., & Klein, M. (2002). Logistic regression: A self-learning text. In *Survival* (2nd ed.). Springer.

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, *100*(1), 128–161. https://doi.org/10.1037/a0036734

Kocoglu, I., & Mithani, M. A. (2020). Does an attractive partner make you a better leader? Only if you are a male! *Leadership Quarterly*, *31*(2), 101339. https://doi.org/10.1016/j.leaqua.2019.101339

Kotzian, P., Stoeber, T., Hoos, F., & Weissenberger, B. E. (2020). To be or not to be in the sample? On using manipulation checks in experimental accounting research. *Accounting Research Journal*, *33*(3), 469482.

Lapchick, R. E. (2008). *The 2006-07 racial and gender report card: College sport*. http://nebula.wsimg.com/5a6003e760caa0fe3001595d256db788?AccessKeyId=DAC3A56D8FB782449D2A&disposition=0&alloworigin=1

Lapchick, R. E. (2017). *College sport racial & gender report card* (B. Estrella & N. Bredikhina (eds.)).

http://nebula.wsimg.com/5665825afd75728dc0c45b52ae6c412d?AccessKeyId=DAC3A56D8FB782449D2A&disposition=0&alloworigin=1

Lee, S. Y., Pitesa, M., Thau, S., & Pillutla, M. M. (2015). Discrimination in selection decisions: Integrating stereotype fit and interdependence theories. *Academy of Management Journal*, *58*(3), 789–812.

Livingston, R. W., & Pearce, N. A. (2009). The teddy-bear effect: Does having a baby face benefit black chief executive officers? *Psychological Science*, *20*(10), 1229–1236. https://doi.org/10.1111/j.1467-9280.2009.02431.x

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis: The assumption that measurement error always reduces effect sizes is fals. *Science*, *355*(6325), 582–584. https://doi.org/10.1126/science.aam5409

Lord, R. G., & Maher, K. J. (1993). *Leadership and information processing: Linking perceptions and performance*. Routledge.

Lord, R. G., & Smith, J. E. (1983). Theoretical, information processing, and situational factors affecting attribution theory models of organizational behavior. *Academy of Management Review*, *8*(1), 50–60. https://doi.org/10.5465/AMR.1983.4287658

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Manning, J. E. (2018). Membership of the 115th Congress: A profile. In *Congressional Research Service*. https://fas.org/sgp/crs/misc/R44762.pdf

Marchiondo, L. A., Myers, C. G., & Kopelman, S. (2015). The relational nature of leadership identity construction: How and when it influences perceived leadership and decision-making. *Leadership Quarterly*, *26*(5), 892–908. https://doi.org/10.1016/j.leaqua.2015.06.006

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

Motro, D., & Ellis, A. P. J. (2017). Boys, don't cry: Gender and reactions to negative performance feedback. *Journal of Applied Psychology*, *102*(2), 227–235. https://doi.org/10.1037/apl0000175

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *Journal of General Psychology*, *135*(2), 151–165. https://doi.org/10.3200/GENP.135.2.151-166

O'Reilly, C. A., Doerr, B., & Chatman, J. A. (2017). "See you in court": How CEO narcissism increases firms' vulnerability to lawsuits. *Leadership Quarterly*. https://doi.org/10.1016/j.leaqua.2017.08.001

Obenauer, W. G., & Langer, N. (2019). Inclusion is not a slam dunk: A study of differential leadership outcomes in the absence of a glass cliff. *Leadership Quarterly*, *30*(6), 101334.

https://doi.org/10.1016/j.leaqua.2019.101334

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin and Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Park, S. H., & Westphal, J. D. (2013). Social discrimination in the corporate elite: How status affects the propensity for minority CEOs to receive blame for low firm performance. *Administrative Science Quarterly*, *58*(4), 542–586. https://doi.org/10.1177/0001839213509364

Pew Research Center. (2016). Discrimination and racial inequality. In *On views of race and inequality, blacks and whites are worlds apart* (pp. 44–45). http://www.pewsocialtrends.org/2016/06/27/3-discrimination-and-racial-inequality/

Powell, G. N., & Butterfield, D. A. (1997). Effect of race on promotions to top management in a federal department. *Academy of Management Journal*, *40*(1), 112–128. https://doi.org/10.2307/257022

Reynolds, T., Zhu, L., Aquino, K., & Strejcek, B. (2021). Dual pathways to bias: Evaluators' ideology and ressentiment independently predict racial discrimination in hiring contexts. *Journal of Applied Psychology*, *106*(4), 624–641. https://doi.org/10.1037/apl0000804

Riggs, B. S., & Porter, C. O. L. H. (2017). Are there advantages to seeing leadership the same? A test of the mediating effects of LMX on the relationship between ILT congruence and employees' development. *Leadership Quarterly*, *28*(2), 285–299. https://doi.org/10.1016/j.leaqua.2016.10.009

Roberts, L. M., & Mayo, A. J. (2019). *Toward a racially just workplace*. Harvard Business Review. https://hbr.org/2019/11/toward-a-racially-just-workplace

Rosch, E. (1978). Principles of categorization. In E. Rosch & Barbara B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48).

Rosenthal, R., & Rubin, D. B. (2003). r equivalent: A simple effect size indicator. *Psychological Methods*, *8*(4), 492–496. https://doi.org/10.1037/1082-989X.8.4.492

Rosette, A. S., Koval, C. Z., Ma, A., & Livingston, R. (2016). Race matters for women leaders: Intersectional effects on agentic deficiencies and penalties. *Leadership Quarterly*, *27*(3), 429–445. https://doi.org/10.1016/j.leaqua.2016.01.008

Rosette, A. S., Leonardelli, G. J., & Phillips, K. W. (2008). The white standard: Racial bias in leader categorization. *Journal of Applied Psychology*, *93*(4), 758–777. https://doi.org/10.1037/0021-9010.93.4.758

Rosette, A. S., & Livingston, R. W. (2012). Failure is not an option for Black women: Effects of organizational performance on leaders with single versus dual-subordinate identities. *Journal of Experimental Social Psychology*, *48*(5), 1162–1167. https://doi.org/10.1016/j.jesp.2012.05.002

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, *11*(6), 446–453.

Ryan, M. K., Haslam, S. A., Hersby, M. D., & Bongiorno, R. (2011). Think crisis-think female: The glass cliff and contextual variation in the think manager-think male stereotype. *Journal of Applied Psychology*, *96*(3), 470–484. https://doi.org/10.1037/a0022133

Salerno, J. M., Peter-Hagene, L. C., & Jay, A. C. V. (2019). Women and African Americans are less influential when they express anger during group decision making. *Group Processes and Intergroup Relations*, *22*(1), 57–79. https://doi.org/10.1177/1368430217702967

Schaumberg, R. L., & Flynn, F. J. (2017). Self-reliance: A gender perspective on its relationship to communality and leadership evaluations. *Academic Management Journal*, *60*(5), 1859–1881.

Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). Researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science*, *22*(4), 1105–1120. https://doi.org/10.1287/orsc.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*(3), 515–530. https://doi.org/10.1080/10478400802049936

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Singer, M. S., & Bruhns, C. (1991). Relative effect of applicant work experience and academic qualification on selection interview decisions: A study of between-sample generalizability. *Journal of Applied Psychology*, *76*(4), 550–559. https://doi.org/10.1037/0021-9010.76.4.550

Singleton Jr., R. A., & Straits, B. C. (2005). *Approaches to social research* (Fourth). Oxford University Press.

Stewart, V. (2020). *Please, take your knee off our necks so we can breathe*. The Cipher Brief. https://www.thecipherbrief.com/column_article/please-take-your-knee-off-our-necks-so-we-can-breathe

Stone, D. L., Hosoda, M., Lukaszewski, K. M., & Phillips, T. N. (2008). Methodological problems associated with research on unfair discrimination against racial minorities. *Human Resource Management Review*, *18*(4), 243–258. https://doi.org/10.1016/j.hrmr.2008.06.001

Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A. (1995). Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology*, *68*(2), 199–214. https://doi.org/10.1037/0022-3514.68.2.199

Sy, T., Shore, L. M., Strauss, J., Shore, T. H., Tram, S., Whiteley, P., & Ikeda-Muromachi, K. (2010). Leadership perceptions as a function of race-occupation fit: The case of Asian Americans. *Journal of Applied Psychology*, *95*(5), 902–919. https://doi.org/10.1037/a0019501

Trichas, S., Schyns, B., Lord, R., & Hall, R. (2017). "Facing" leaders: Facial expression and leadership perception. *Leadership Quarterly*, *28*(2), 317–333. https://doi.org/10.1016/j.leaqua.2016.10.013

Tucker, S., Ogunfowora, B., & Ehr, D. (2016). Safety in the c-suite: How chief executive officers influence organizational safety climate and employee injuries. *Journal of Applied Psychology*, *101*(9), 1228–1239.

U.S. Bureau of Labor Statistics. (2017). *Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity*. U.S. Department of Labor. https://www.bls.gov/cps/cpsaat11.pdf

U.S. Equal Employment Opportunity Commission. (2019). *Race-based charges (Charges filed with EEOC) FY 1997 - FY 2019*. https://www.eeoc.gov/statistics/race-based-charges-charges-filed-eeoc-fy-1997-fy-2019

Ubaka, A., Lu, X., & Gutierrez, L. (2020). *Taking another look at race and the American business leader prototype: A direct replication of Rosette, Leonardelli, and Phillips (2008)*.

Ubaka, A., Lu, X., & Gutierrez, L. (2021). *Testing the generalizability of the White leadership standard in the post-Obama era*.

van Quaquebeke, N., van Knippenberg, D., & Brodbeck, F. C. (2011). More than meets the eye: The role of subordinates' self-perceptions in leader categorization processes. *Leadership Quarterly*, *22*(2), 367–382. https://doi.org/10.1016/j.leaqua.2011.02.011

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of fratification and later outcomes. *Psychological Science*, *29*(7), 1159–1177. https://doi.org/10.1177/0956797618761661

Whetten, D. a. (1989). What constitutes a theoretical contribution? *Academy of Management Review*, *14*(4), 490–495. https://doi.org/10.5465/AMR.1989.4308371

White, G. B. (2017, October 26). There are currently 4 Black CEOs in the Fortune 500. *The Atlantic*. https://www.theatlantic.com/business/archive/2017/10/black-ceos-fortune-500/543960/

Wulff, J. N., & Villadsen, A. R. (2020). Are survey experiments as valid as field experiments in management research? An empirical comparison using the case of ethnic employment discrimination. *European Management Review*, *17*(1), 347–356. https://doi.org/10.1111/emre.12342

Young, R. K., Kennedy, A. H., Newhouse, A., Browne, P., & Thiessen, D. (1993). The effects of names on perception of intelligence, popularity, and competence. *Journal of Applied Social Psychology*, *23*(21), 1770–1788.

Younkin, P., & Kuppuswamy, V. (2017). The colorblind crowd? Founder race and performance

in crowdfunding. *Management Science*, *64*(7), 3269–3287.

Zapata, C. P., Carton, A. M., & Liu, J. T. (2016). When justice promotes injustice: Why minority leaders experience bias when they adhere to interpersonal justice rules. *Academy of Management Journal*, *59*(4), 1150–1173. https://doi.org/10.5465/amj.2014.0275

Zhu, L. (Lei), Aquino, K., & Vadera, A. K. (2016). What makes professors appear credible: The effect of demographic characteristics and ideological beliefs. *Journal of Applied Psychology*, *101*(6), 862–880. https://doi.org/10.1037/apl0000095

Ziegert, J. C., & Hanges, P. J. (2005). Employment discrimination: the role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology*, *90*(3), 553–562. https://doi.org/10.1037/0021-9010.90.3.553

TABLE 1

Original RLP Hypotheses Results and Replication Results

RLP Hypotheses (Studies 1&3 Only)	RLP 1	Study 1A	Study 1B	RLP3	Study 2A	Study 2B	Study 2C	Study 3A	Study 3B	Study 3C
Hypothesis 1: If being White is associated with the leader prototype, then there should be no interaction between the target person's role (employee, leader) and base rate information; instead, a main effect should occur whereby observers will consider leaders to be White more than nonleaders (e.g., employees), regardless of the base rate information	Full Support	Minimal Support	Minimal Support	--	--	--	--	--	--	--
Hypothesis 2: If being White is associated with the leader prototype, then observers will be more likely to deviate from base rates in the leader than in the nonleader (e.g., employee) conditions when assessing the race of the target	Full Support	No Support	No Support	--	--	--	--	--	--	--
Hypothesis 5A: On the basis of the negative stereotype explanation, we predict a three-way interaction between race, organizational performance, and performance attributions: Non-White leaders will be evaluated less favorably than White leaders, but only when they are blamed for unsuccessful organizational performance	--	--	--	No Support	No Support	No Support	No Support	No Support	No Support	No Support
Hypothesis 5B: On the basis of our proposed White business leader prototype, we predict a competing three-way interaction: White leaders will be evaluated more favorably than non-White leaders, but only when they are given credit for successful organizational performances	--	--	--	Full Support	No Support	No Support	No Support	No Support	No Support	No Support

TABLE 2

Binary Hierarchical Logistic Regression for DV = Perceptions of Leader Race (Study 1)

Target Study	RLP1, N = 146											
	Step 0				Step 1				Step 2			
	B	*SE*	*Wald*	*r*	*B*	*SE*	*Wald*	*r*	*B*	*SE*	*Wald*	*r*
Constant	0.65	0.17	13.99 ***	0.30	0.11	0.33	0.11	0.03	0.00	0.38	0.00	0.00
Interviewee role (IR)					-0.76	0.38	3.90 *	0.17	-0.51	0.57	0.81	0.07
Base rate (BR)							17.01 **	0.33			10.47 **	0.26
IR X BR											0.34	0.05
Model 1	Study 1A Analysis for Full Sample, N = 558											
	Step 0				Step 1				Step 2			
	B	*SE*	*Wald*	*r*	*B*	*SE*	*Wald*	*r*	*B*	*SE*	*Wald*	*r*
Constant	1.62	0.11	202.23 ***	0.52	2.98	0.34	75.48 ***	0.35	2.83	0.46	37.91 ***	0.25
Interviewee role (IR)					-0.42	0.24	3.05 t	0.07	-0.16	0.62	0.07	0.01
Base rate (BR)							37.13 ***	0.25			17.24 ***	0.17
IR X BR											0.70	0.04
Model 2	Study 1A: Analysis for Only Participants Correctly Responding to Manipulation Checks, N = 478											
	Step 0				Step 1				Step 2			
	B	*SE*	*Wald*	*r*	*B*	*SE*	*Wald*	*r*	*B*	*SE*	*Wald*	*r*
Constant	1.67	0.13	177.34 ***	0.52	3.36	0.42	64.22 ***	0.34	3.28	0.59	31.17 ***	0.25
Interviewee role (IR)					-0.60	0.27	4.94 *	0.10	-0.47	0.78	0.35	0.03
Base rate (BR)							40.28 ***	0.28			17.72 ***	0.19
IR X BR											0.08	0.01
Model 3	Study 1B Analysis for Full Sample, N = 498											
	Step 0				Step 1				Step 2			
	B	*SE*	*Wald*	*r*	*B*	*SE*	*Wald*	*r*	*B*	*SE*	*Wald*	*r*
Constant	1.65	0.12	183.58 ***	0.52	3.31	0.42	63.23 ***	0.34	3.25	0.59	30.41 ***	0.24
Interviewee role (IR)					-0.37	0.26	2.04	0.06	-0.26	0.78	0.11	0.02
Base rate (BR)							40.09 ***	0.27			18.26 ***	0.19
IR X BR											0.09	0.01
Model 4	Study 1B: Analysis for Only Participants Correctly Responding to Manipulation Checks, N = 452											
	Step 0				Step 1				Step 2			
	B	*SE*	*Wald*	*r*	*B*	*SE*	*Wald*	*r*	*B*	*SE*	*Wald*	*r*
Constant	1.70	0.13	170.35 ***	0.52	4.11	0.61	45.75 ***	0.30	3.58	0.72	24.99 ***	0.23
Interviewee role (IR)					-0.47	0.28	2.72 t	0.08	0.67	1.24	0.29	0.03
Base rate (BR)							40.28 ***	0.29			19.07 ***	0.20
IR X BR											1.20	0.05

$^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$, $^{t}p<0.10$

Note: Because the base rate had three conditions, there is no referent point from which coefficients can be derived, hence there are no coefficients reported for BR or IR X BR. "r" refers to effect sizes reported in terms of $r_{equivalent}$ (Rosenthal & Rubin, 2003).

TABLE 3

Proportion Perceived White by Condition and Z-tests of Proportion (Study 1)

			Interviewee Role: Employee				Interviewee Role: Leader			
	Study	Base Rate	N	Perceived White	SE	Z	N	Perceived White	SE	Z
Target Study	RLP1	20%	24	37.50%	--	0.54	28	50.00%	--	2.35 *
	RLP1	50%	24	62.50%	--	0.88	22	81.82%	--	2.23 *
Model 1	Study 1A[a]	None	96	83.33%	--	--	91	91.21%	--	--
	Study 1A[a]	20%	98	66.33%	0.05	11.47 ***	90	73.33%	0.05	12.65 ***
	Study 1A[a]	50%	93	93.55%	0.03	8.40 ***	90	94.44%	0.02	8.43 ***
Model 2	Study 1A[b]	None	81	85.19%	--	--	88	92.05%	--	--
	Study 1A[b]	20%	75	61.33%	0.06	8.95 ***	80	73.75%	0.05	12.02 ***
	Study 1A[b]	50%	71	94.37%	0.03	7.48 ***	83	96.39%	0.02	8.45 ***
Model 3	Study 1B[a]	None	88	85.23%	--	--	83	90.36%	--	--
	Study 1B[a]	20%	79	64.56%	0.05	9.901 ***	85	71.76%	0.05	11.93 ***
	Study 1B[a]	50%	83	95.18%	0.02	8.23 ***	80	96.25%	0.02	8.27 ***
Model 4	Study 1B[b]	None	85	84.71%	--	--	78	92.31%	--	--
	Study 1B[b]	20%	71	61.97%	0.06	8.84 ***	73	71.23%	0.05	10.94 ***
	Study 1B[b]	50%	71	98.59%	0.01	8.19 ***	74	97.30%	0.02	8.14 ***

$^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$, $^{t}p<0.10$

[a]Full sample, [b]Sample restricted to participants who responded to manipulation checks correctly

Note: Perceived refers to the proportion of participants who perceived the employee/leader as White

TABLE 4

ANOVA for Dependent Variable = Leader Evaluation Rating (Studies 2 & 3)

Variable	RLP3 *N = 479* *F(1, 471)*	*r*	Model 1 Study 2A *N = 762* *F(1, 754)*	*r*	Model 2 Study 2B *N = 446* *F(1, 438)*	*r*	Model 3 Study 2C *N = 494* *F(1, 486)*	*r*	Model 4 Study 3A *N = 965* *F(1, 949)*	*r*	Model 5 Study 3B *N = 1000* *F(1, 984)*	*r*	Model 6 Study 3C *N = 1005* *F(1, 989)*	*r*	Model 7 Study 3D *N = 1743* *F(1, 1719)*[#]	*r*
Performance	270.86 ***	0.60	497.63 ***	0.63	230.95 ***	0.58	186.93 ***	0.52	711.36 ***	0.65	784.65 ***	0.66	853.75 ***	0.68	928.82 ***	0.59
Attribution	2.94	0.08	3.29 t	0.07	0.03	0.01	12.81 ***	0.16	6.14 *	0.08	38.49 ***	0.19	38.58 ***	0.19	0.33	0.01
Leader's race	3.92 *	0.09	4.21 *	0.07	6.45 *	0.12	8.42 **	0.13	28.84 ***	0.17	1.35	0.04	2.21	0.05	28.43 ***	0.13
Performance X Attribution	39.92 ***	0.28	96.89 ***	0.34	28.06 ***	0.24	82.68 ***	0.38	163.36 ***	0.38	154.15 ***	0.37	198.62 ***	0.41	297.00 ***	0.38
Performance X Leader's Race	0.16	0.00	0.08	0.01	1.40	0.06	0.01	0.00	1.93	0.04	2.77 t	0.05	8.20 **	0.09	0.08	0.01
Attribution X Leader's Race	3.95 *	0.09	0.02	0.01	4.22 *	0.10	3.33 t	0.08	1.04	0.03	2.11	0.05	4.50 *	0.07	1.32	0.03
Performance X Attribution X Leader's Race	6.68 **	0.12	3.20 t	0.06	0.15	0.02	0.09	0.01	0.54	0.02	2.29	0.05	0.56	0.02	1.03	0.02
Design Manipulation									0.01	0.00	0.11	0.01	0.13	0.01	87.45 ***	0.22
Design X Performance									4.04 *	0.06	0.42	0.02	0.03	0.01	12.74 ***	0.09
Design X Attribution									5.05 *	0.07	0.27	0.02	4.36 *	0.07	13.79 ***	0.09
Design X Leader's Race									0.01	0.00	3.67 t	0.06	0.68	0.03	2.44 t	0.04
Design X Performance X Attribution									4.97 *	0.07	0.91	0.03	0.02	0.00	6.43 **	0.06
Design X Performance X Leader's Race									0.96	0.03	0.00	0.00	0.65	0.03	0.04	0.00
Design X Attribution X Leader's Race									0.44	0.02	0.00	0.00	0.02	0.00	1.04	0.02
Design X Performance X Attribution X Leader's Race									0.54	0.02	0.59	0.02	0.60	0.02	0.68	0.02

$^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$, $^{t}p<0.10$

[#] degrees of freedom were 2 and 1719 for all terms in model below the dashed line

Note: "r" refers to effect sizes reported in terms of $r_{contrast}$ (Rosnow et al., 2000)

TABLE 5

Mean Leadership Evaluation Ratings by Condition and Study

Race	Perf	Attr	RLP3 N = 479	Study 2A N = 762	Study 2B N = 446	Study 2C N = 494	Study 3A[#] N = 965	Study 3B[#] N = 1000	Study 3C[#] N = 1005
White	All	All	4.78	4.41	4.61	4.37	4.93	4.96	5.03
			(1.13)	(1.17)	(1.08)	(1.00)	(1.41)	(1.46)	(1.46)
Non-White	All	All	4.67	4.56	4.81	4.62	5.32	5.17	5.05
			(1.13)	(1.12)	(1.10)	(1.08)	(1.38)	(1.50)	(1.60)
White	High	All		5.13	5.27	4.84	5.77	5.84	5.88
				(1.04)	(0.85)	(1.08)	(1.04)	(0.95)	(0.96)
White	Low	All		3.72	3.93	3.89	4.12	4.12	4.19
				(0.84)	(0.84)	(0.62)	(1.24)	(1.36)	(1.37)
Non-White	High	All		5.23	5.38	5.13	6.10	6.16	6.17
				(0.96)	(1.02)	(0.98)	(0.90)	(0.78)	(0.94)
Non-White	Low	All		3.87	4.23	4.11	4.51	4.17	3.97
				(0.80)	(0.85)	(0.92)	(1.32)	(1.39)	(1.35)
White	All	Leader		4.46	4.54	4.60	4.78	5.25	5.10
				(1.38)	(1.19)	(1.25)	(1.62)	(1.13)	(1.12)
White	All	Market		4.36	4.68	4.19	5.09	4.68	4.96
				(0.92)	(0.96)	(0.70)	(1.14)	(1.68)	(1.74)
Non-White	All	Leader		4.62	4.90	4.72	5.16	5.40	5.28
				(1.28)	(1.25)	(1.28)	(1.66)	(1.24)	(1.26)
Non-White	All	Market		4.49	4.72	4.54	5.48	4.95	4.82
				(0.93)	(0.93)	(0.85)	(1.00)	(1.69)	(1.86)
White	High	Leader	6.04	5.54	5.43	5.47	6.11	6.03	6.33
			(0.62)	(0.90)	(0.79)	(0.98)	(0.83)	(0.88)	(0.75)
White	High	Market	5.01	4.71	5.11	4.35	5.42	5.65	5.44
			(0.98)	(1.00)	(0.89)	(0.89)	(1.11)	(0.99)	(0.95)
White	Low	Leader	3.97	3.41	3.61	3.73	3.50	3.38	3.59
			(0.89)	(0.87)	(0.76)	(0.79)	(1.06)	(1.17)	(1.31)
White	Low	Market	4.37	4.03	4.23	4.03	4.76	4.85	4.76
			(0.79)	(0.69)	(0.82)	(0.39)	(1.08)	(1.12)	(1.17)
Non-White	High	Leader	5.47	5.55	5.67	5.53	6.42	6.29	6.38
			(0.95)	(0.90)	(1.06)	(0.96)	(0.69)	(0.71)	(0.87)
Non-White	High	Market	5.19	4.92	5.09	4.73	5.77	6.03	5.96
			(0.94)	(0.93)	(0.90)	(0.83)	(0.98)	(0.82)	(0.96)
Non-White	Low	Leader	3.87	3.67	4.12	3.82	3.86	3.72	3.36
			(0.92)	(0.86)	(0.88)	(0.95)	(1.33)	(1.36)	(1.25)
Non-White	Low	Market	4.20	4.05	4.34	4.36	5.17	4.69	4.59
			(0.90)	(0.70)	(0.81)	(0.84)	(0.93)	(1.25)	(1.15)

Standard deviations in parenthesis

[#] *Sample restricted to participants in conditions comparable to Studies 2 and 3*