Evaluating High Performing Female Colleagues: The Roles of Race, Gender, and Task Performance

by

Jonathan Hochberg

Thesis Advisor: Professor Scott Page, Ph.D.

John Seely Brown Distinguished University Professor of

Complexity, Social Science, and Management

A thesis proposal submitted in fulfillment of the requirements of the
Michigan Ross Senior Thesis Seminar (BA 480), April 22.

Abstract

Understanding how high-performing female colleagues are evaluated has significant implications for improving group cohesion and productivity in the workplace. While a great deal of research has been done on identifying high performers and supporting them, little focus has been lent to studying these phenomena in conjunction with the unique issues faced by women and people of color in the workplace. Across two studies, the impact of a participant's demographics, partner's race, and performance on a task are measured in regards to their evaluation of a high performing female teammate. Results of this study show that female participants rated their partners higher than male participants did, black participants gave higher ratings when they had a white partner, and men gave higher ratings to their partner when they solved both tasks, whereas women gave higher ratings to their partner when they did not solve either task. These results provide valuable insight into the factors that affect how female high performers are perceived, and add to the growing literature around supporting high performers with marginalized identities.

Contents

Introduction and Literature Review

A chief goal of workplace research has been to increase the happiness and productivity of employees, especially high-performing ones, so-called *superstars*, who generate outsized value relative to their peers. However, the concept of employee superstardom has proven difficult to study, because its definition is highly variable, as it depends on one's role, company, and even industry, as well as subjective, because workplace success is often qualitative. Additionally, there are biases, both conscious and unconscious, that affect the identification, promotion, and even the incentivization of these high-performing employees, based on their demographic characteristics, like race and gender. This literature review will draw from organizational behavior, sociology, economics, and psychology to present an eclectic view of female high performers at work, define the conditions that allow them to be most successful when joining a new team, and describe the specific barriers facing them based on their gender and race.

Defining Superstars: A History of High-Performance

The first research into superstar performance began with Alfred Marshall in the 19th century. Far ahead of his time, he studied several industries and observed that only certain ones had individuals who commanded higher prices or achieved an outsized market share. He then posited that these differences were due to the potential size of the audience, so a painter could command a higher price for a painting than a singer could for a single performance, because the latter has a naturally limited audience (Marshall 2009). In fact, the lower per unit costs made possible by mass production and the larger audiences made possible through technological innovations encouraged larger so-called *superstar effects* (Franck & Nuesch 2012). Marshall's insights into these high performers and the industry conditions necessary to create them formed the basis of the next generation of economic thought on superstars, especially Sherwin Rosen.

Marshall's research was expanded in the late 20th century, by Rosen and Adler, who endeavored to provide a generalizable definition for an individuals' financial dominance in a large market. An important caveat is that these economists focused on a visible, monetary form of market dominance and primarily used historical data, so their findings focus on a quantitative, objective forms of success that have already been achieved. The most famous explanation for superstars comes from Rosen, who suggested that superstardom results from reliant on a hierarchy of talent and near perfect reproducibility of art. Essentially, his perspective suggests that one individual will dominate a market when inferior talent is an inefficient replacement for superior talent, like how several mediocre singers singing together sound worse than a single competent singer (Rosen 1981). Adler expanded upon this perspective, by crediting conformity and knowledge capital for superstar performance rather than talent (Adler 1985). Under this perspective, a famous singer has no more talent than a relatively unknown singer, but achieves dominance through network effects and social support (Gergaud & Verard 2006). This result was supported in another study that simulated how people chose to download music, which found that a stronger degree of social influence increased inequality and decreased the probability of success (Salganik, Dodds, & Watts, 2006). These economic perspectives created generalized definitions of superstar performance, but other fields have their own definitions and streams of research.

In addition to the economic theories surrounding the determinants necessary for superstardom, other fields have taken a more varied approach. One conceptual review summarized three different research streams into stars and broke down the siloes between them. These authors conceptualize star status as being a combination of performance, visibility, and social capital, each of which is studied in a different way by three main fields: economics,

sociology, and organizational behavior (Call, Nyberg, & Thatcher, 2015). The challenge in defining high performance, or superstardom, is how to combine these distinct perspectives and definitions into a cogent, operationalizable definition of superstardom to simulate it experimentally. A common thread between the siloed research of these three fields is that of exceptional performance, how a superstar performs better than their peers at work, a trend that is consistent across companies and industries.

One feature of the economics conditions that promoted superstardom is that performance in the industry follows a power law distribution rather than a bell curve one, such that the most successful products are orders of magnitudes more successful than their peers (Kreuger 2019). Similarly, the highest performers at work perform order of magnitudes higher than their peers (O'Boyle Jr & Aguinis 2012). This finding is massively important to the field of organizational behavior, because it demonstrates the outsized benefits of high-performers in organizations, across industries like sports, entertainment, and academic research. This has broad implications for how to manage employees in organizations, because if the expected value of high performers is so large, then logically, more resources should be invested into the identification, support, and compensation of high performers. However, the validity of the power law and such high-performers is more common under conditions of autonomous and complex jobs with a high productivity ceiling, so this law does not apply to every role (Aguinis, O'Boyle Jr, Gonzalez-Mule, & Joo 2016). Additionally, the performance element of superstardom, like percentage return on investment or amount of new business acquired is a quantifiable, easily comparable metric. Superstars and the high performance they bring with them are an obvious boon to the team and company they join, but they sometimes even improve their teammates' experience.

Superstar's Positive Impact on Others

Many studies have analyzed superstars across academia, the sports industry, and finance, as they have clear metrics, performance is externally visible, and the outcomes are collaborative. A focus has been placed on the additive impact of superstars, to better understand the positive and negative externalities associated with them joining a team, so-called *superstar effects*. There is much research on both the positive and negative impacts of superstars on their teammates and their competitors, and it is important to combine the specific claims about superstardom to present a unified theory.

Many of the studies that find a positive superstar effect find that they mostly benefit the larger entity they are a part of, rather than their specific team. Within the domain of academia, one study found that hiring superstars did not affect the productivity of other research scholars, but had a significant positive impact on subsequent recruiting efforts, so they had a positive impact at the department-level, that is otherwise not apparent to existing team members (Agrawal, McHale, & Oettl 2014). In tennis, star players have been shown to have a disproportionate positive impact on ticket sales for an event, impacting demand (Chmait, Roberston, Westerbeek, Eime, Sellitto, & Reid 2019). In basketball, superstars are shown to have an outsized positive impact on television ratings, which benefit the owners of the team, but not necessarily their teammates (Hausman & Leonard 1997). In the medical field, hospitals with star performers were much faster to adopt the coronary stent, suggesting that high performers within an organization spur the adoption of new technology (Burke, Fournier, & Prasad 2007). Within these few examples, the existence of superstars is highly beneficial to the larger organization, but not to their peers or direct managers, which could create a principal-agent issue (Grossman & Hart 1992). Superstars have been demonstrated to benefit their team members, and have even been shown to have a similar effect on their competition.

In addition to superstar's impact on their broader organization, they have even been shown to improve the performance of their competitors. In track and field, competing against a well-known, highly-successful competitor, like Usain Bolt, has been shown to boost performance of other runners (Hill 2014). So, superstars can improve the performance of those around them, even those competing against them. This result could be possible through a positive competition effect, whereby their teammates or competitors train or work harder to defeat them. Or, it could be the result of sharing the superstar's improved methods, in which their teammate make use of a new technique created by the star. A positive peer effect has been observed in swimming, where having a good swimmer on one's team improves the team's performance (Jane, Yao, & Wang 2018). So, in the case of a winner-take-all sporting event, a superstar would likely increase the performance of their competitors, a negative result. However, in the context of a competitive group within a company, improving the performance of one's competitors would overall benefit the company, so this effect could in fact be a positive one. However, despite the performance advantages of stars and their potential positive impact on the overall company, they also sometimes cause negative externalities to those around them.

Superstar's Negative Impact on Others

However, there are several notable counterexamples to the finding that superstars improve the performance of competitors, suggesting a mixed result. In golf, contrary to the several other studies about other sports, the existence of a superstar, in this case Tiger Woods, had a strong decrease on competitor's performance (Brown 2011). Another study looked at the Japan Golf Tour and found that the existence of superstar golfers decreases the performance of other golfers (Tanaka & Ishino 2012). An important consideration is that golf is an individual sport without any semblance of a team, while swimming and track involve individual

competitors on a team. This demonstrates that it is difficult to distinct superstar's effect on their competitors, and is likely context-dependent. In addition to the existence of superstars causing a decrement in performance of their competitors, there is also evidence that they have such an impact on their team members and even the broader organization of which they are a part.

The dark side of hiring stars is apparent when they detract from the performance of their peers, through crowding out, negative externalities, and negative spillover effects. One study looked at academia as a kind of zero-sum game, based on limited space in top journals, and found that when superstar scientists passed away, non-collaborators published over 8% more (Azoulay, Fons-Rosen, & Graff Zivin 2019). However, there is also a 5-8% decrease in the quality-adjusted publishing rates of the superstar's collaborators, suggesting there may be a trade-off to the success of superstars (Azoulay, Graff Zivin, & Wang 2010). Another study distincted highly helpful high-performing academics from non-helpful ones and found that after they die, their coauthors' work decreases in quality, but not quantity (Oettl 2012).

A related study on the biotechnology industry found that stars improved their firm's performance, but crowded out the development of other leaders in their companies (Kehoe & Tzabbar 2015). When superstars are unhelpful or crowd others out from success, their existence on one's team is maladaptive, and brings up the question of how this effect functions as the number of superstars on a team grows.

A related topic is at what point the amount of talent or number of superstars becomes excessive, essentially the diminishing marginal returns of superstars. Groysberg studied this phenomenon and found that for sell-side equity researchers, groups benefitted from star performers, but only to a point. Then, adding more talented employees became less and less beneficial, especially for sectors with high star concentration (Groysberg, Polzer, & Elfenbein

2011). Another work looked at five studies relating talent and team performance and found that more talent is only beneficial to a point, unlike participant's assumptions that more talent is always better. Additionally, they brought up the level of task interdependence as a mediating factor, such that the more teams that have to work together on a collaborative task, the more this "too-much-talent effect" is observed (Swaab, Schaerer, Anicich, Ronay, & Galinsky 2014). The number of superstars on a team clearly has a diminishing impact, but there are factors that affect the success of superstars, like their environment, colleagues, and attitude.

Considerations for Supporting Superstars

The question of whether a superstar joining one's team or competing against someone will improve their performance is one that depends in part on the environment and colleagues of the high-performer. One study found that an important factor that predicted the portability of high performance was the nature of the position at the company a star joined. So, when a company hires a star to support existing activities (exploitative work) rather than hiring them to start new activities (exploratory work), stars experience a smaller decrease in performance at the new firm (Groysberg & Lee 2009). This suggests an additional characteristic that should be considered in hiring a star is whether the work they will be expected to complete is exploratory in that it requires starting new activities. Bringing in a superstar to do work they are already familiar with logically helps them transition more successfully, but their colleagues also play a big role in that process.

An additional consideration is how to treat the high performer at work, because they tend to have specific increased risk factors for certain workplace issues. One study found that high status employees perform worse after a loss of status than their lower performing peers, suggesting companies should be place more of an emphasis on diminishing these moments.

Interestingly, self-affirmation acted as a counterbalance and restored the status of high performers, so the star has options to sustain their performance in the face of a status loss (Marr & Thau 2014). In line with this finding, employees in organizations have a more favorable impression of people who ascend to a role rather than those who descend to it, suggesting any fall from grace would be especially damaging to a star performer (Pettit, Sivanathan, Gladstone, & Marr 2013). In addition to the valid concerns around how to support superstars at work, there is also the complication of how to support superstars that come from another team, or even another company.

Stars have also been demonstrated to have low portability of their performance (Groysberg 2010), but are still at a greater risk for being poached by another firm because they have high performance (Nyberg 2010). In fact, this fear of poaching is reflected even in the hiring process, where high performers are thought to have lower levels of interest in the organization because they are thought to have desirable other opportunities (Galperin, Hahl, Sterling, & Guo, 2019). One retention mechanism, in the case of security analysts, is high quality colleagues, where having such peers provided social support that acted as a buffer against exit to another firm (Groysberg & Lee 2010). While the social capital of a star is certainly beneficial, they however act as a kind of keystone species (Mills, Soule, & Doak 1993), so they are more at risk of information overload, so there are risk factors to increased social capital as well (Oldroyd & Morris 2012). In line with avoiding over-reliance on stars, one study found that organizations can become over-reliant on their stars to disseminate knowledge to their peers, and found that stars temporary absence in the NBA led to improved organizational performance (Chen & Garg 2018). Clearly, superstars are at risk of several workplace issues because of their high performance, but they can also take actionable steps to improve their reception on a new team.

While the colleagues of the superstar should strive to support them, it is also the responsibility of the superstar to work in a way that benefits others to ensure maximum overall productivity. Existing research provides a solid platform for analyzing high performers and understanding the biases faced by people of different races and gender, however there is a gap in the literature on the intersection of the two. Research into superstardom has demonstrated the effects of superstars on their peers and how to best support them. Work on race and gender has demonstrated the different conscious and unconscious biases peers and superiors have against women and people of color. But there is a dearth of research into superstars of color and female superstars, likely because much of the work draws from real-world data, wherein women and people of color are already underrepresented. One important consideration is whether the positive attributes associated with superstardom will mediate these biases, or whether they will in fact be strengthened by possible resentment from lesser performing peers. How characteristics of the superstar affect perceptions of them is an important question, but the superstar's attitude toward their work and even their team will also affect their subsequent impact, so they too have a role in ensuring their team works together well.

A notable study looked at employee victimization based on their performance and found that high performers experienced more victimization at work, especially a passive kind. However, the benevolence of star performers worked to shield them from such resentment from other employees, suggesting that high performers should focus on avoiding entitled behavior and work to help others (Jensen, Patel, & Raver 2014). There is also evidence that superstars attain status through superiority over others, which leads others to perceive them as inauthentic and self-interested (Hahl & Zuckerman, 2014). Another study corroborated this finding and used envy as an explanatory mechanism for why high performers are more victimized. In this study,

they used work group identification, the feeling of working as a team with colleagues and supervisors, as a mechanism to mitigate this victimization (Kim & Glomb 2014).

A related concept is the "tall poppy syndrome," a cultural term describing the denigration of high achievers. In line with the work on victimization, there is work demonstrating that employees are more pleased by a high-performers failure and more punitive to a high-performing cheater (Feather 1989). One study looked at New Zealand athletes and found that how one perceived the impact of such criticisms is more important than the experience of being a target (Pierce, Hodge, Taylor, & Button 2017). These findings are particularly important because high performers are at a higher risk of turning over, because of their higher performance, though there is some work demonstrating that high performance is inverse to voluntary turnover, suggesting they are less likely to leave (Nyberg 2010). How the superstar comports themselves on a new team can affect how successful their transition is, likely because it impacts how others perceive them.

The success of a superstar joining a new environment is also reliant on how they are perceived and the skill levels of their group members. While it is beneficial for an environment to be welcoming and for a superstar to be benevolent, a team's success is also based on the competence of its members. One study found that the superstar effect has been shown to moderate high-performance of small teams, but only when the difference between the superstar's score and the average scores of their teammates are close (Nihalani, Wilson, Thomas, & Robinson 2010). This suggests that superstars should join teams with mostly competent peers, lest they spend too much time explaining things. Another important element of a successful transition is the perception of the superstar. Lockwood demonstrated that superstars are inspiring only when they are relevant and their success appears attainable, suggesting that the perception

of a superstar is a crucial element of how they will impact the team that they join (Lockwood & Kunda 1997). For these high-performers to inspire the rest of the team, it is important that their success look related to one's own pursuits and seem to be achievable.

Additionally, one study looking at tall poppy syndrome found that people who are low in self-esteem, care less about achievement and social power, and lean left politically are more likely to have negative attitudes towards high performers (Feather 1989). Another consideration is the degree to which the high performer has control over their high performance, and to what degree noise affects the value of high performance. One study suggests that high performance is more likely for companies that engage in variable activities, and is subject to much noise because of its rarity (Denrell, 2005). Another paper found a weak link between ability and performance because of noise and self-reinforcing dynamics, so an important question in evaluating the importance of high performance is whether it is the result of skill or luck (Denrell & Liu, 2012). Understanding how to support stars effectively is increasingly important as more stars exist because of the nature of work in the 21st century, with increased access to technology and focus on social capital (Aguinis & O'Boyle Jr 2014). Considering the other factors that impact how a superstar is received on a team, especially race and gender, it is important that how others perceive and engage with them, two questions that form the basis of this paper.

Demographics Influences on Superstardom

Superstars come from all sorts of different demographic backgrounds, but a special consideration should be made to the gender of the superstar. In the swimming study that found positive intra-team peer effects, the researchers also determined that competing against a superstar was shown to improve men's performance, but decrease women's performance, suggesting gender differences in a teammate's performance change in response to the addition of

a superstar (Jane et al. 2018). Stars in general are not portable between companies, meaning they are usually unable to recreate their superior performance and often decrease the value of the team they join, because they lack social support and company-specific knowledge (Grosyberg, Nanda, & Nohria 2004). However, women in fact performed equally as well at new firms, because they had superior external relationships and took more time to select the right employer to which they would transfer (Groysberg 2008). This suggests that women's star performance is more sustainable and they can leverage certain skills they have to adapt better to a new workplace environment. Female superstars are more likely to face additional barriers at work because of unconscious bias (Pollard, 1999) and tend to be more portable, and their gender feeds into how others perceive them, an important factor to how superstars will impact team performance.

An important consideration of how gender intersects with superstar effects is how other people will perceive the superstar, a perception that is open to bias. First, even the term "superstar" is one that faces biases, because even children as young as 6 associated "brilliance" with white men, so there is a systemic issue at hand (Jaxon, Lei, Shachnai, Chestnut, & Cimpian 2019). Similarly, students of color are underrepresented in gifted programs even when they have high standardized test scores, suggesting a bias against associated people of color with being gifted (Grissom & Redding 2015). A related study looked at professor evaluation of black, female professors and found that they are significantly less likely to be called "brilliant" or "genius" (Storage, Horne, Cimpian, & Leslie 2016). Another study evaluated the quantitative metrics used to evaluate performance and found that going from a 10 to a 6-point scale decreased the gender gap in evaluations of performance (Rivera & Tilcsik, 2019). Because these terms of high intelligence and performance and even the metrics used to identify them carry with them a

bias against people of color and women, special care must be given when identifying and labeling high performers.

Problem Statement and Justification

There is a legacy of bias and discrimination in the workplace, especially against underrepresented minorities, due to a history of discriminative practices. Underrepresented minorities must face the biases, both conscious (Deitch, Barsky, Butz, Chan, Brief, & Bradley 2003) and unconscious (Pollard 1999), of their colleagues in an environment where they are already a small percentage of the population. These unfair practices are often subtle (Van Laer & Janssens 2011) and systematic (Agocs & Jain 2001), meaning they are entrenched practices that are difficult to eradicate. These impacted groups must face these negative environmental factors across a wide variety of contexts, even when they are high performers themselves. In addition to the barriers that minorities and superstars face at work, there are also difficulties in transitioning them onto a new team, because of how their team members will respond.

Individual high performance is lauded in individual performance contexts, like sports matches, yet superstars can carry some downsides in group performance contexts, like the workplace, so it is important to consider when to hire one and how to onboard them onto a new team. Firstly, the issue of defining superstar performance is far from solved (Call et al. 2015), so defining high performance and by extension, identifying high performers, is quite difficult, and requires contextualization. Additionally, individual high performance often has a detrimental effect on overall team performance, because of power struggles (Greer & Chu 2019) and resentment (Groysberg & Lee 2009). Even once high performance is defined and identified, it is not clear whether this individual instance of talent is applicable to other contexts, internal and external alike, so the extensibility of high performance must also be considered (Groysberg et al.

2004). Though having high status individuals in a group unsurprisingly benefits performance, there appears to be a diminishing marginal return as more of them join a group (Groysberg et al. 2011). Though the success of an individual is desirable, considerations must be made about the impact made by such a person joining a team, especially when they are an underrepresented minority.

While there has been much research into high performance at work and how women are evaluated, there is limited research that explore the intersection of the two, especially under different demographic and task conditions. This research will address the gap in the literature by considering the demographics of female high performers and applying a collective intelligence lens to the design of the tasks to understand how perceptions of them vary.

The purpose of this research is to examine an individual perception of a high performing female peer after they work together on a task. I will work to understand how superstar status is perceived, achieved, and viewed by others. Through the insights this research will provide, I will create a conceptual framework to describe how people perceive superstar status for female high performers, based on their interactions with them and the nature of the task they are working on. Then, I will discuss the implications of research in creating an equitable team environment and promoting more effective teams, which will in turn increase team cohesion and group performance.

Theoretical Framework

Drawing from Call, Nyberg, and Thatcher's 2015 conceptual review, I will focus on the performance component of a star employee. The authors reviewed 75 papers on stars and derived a single explanation of star employees as resulting from performance, relevant social capital, and visibility. They provide an overview of how stars are described in organizations by different

academic fields to create a conceptually integrated framework. Through their research, they see economics defines stars by the changes in markets and their peer effects, sociology sees them in context of their impact on organizational performance and visibility, and management views them as having social capital and high-performance, but lacking portability (Call et al. 2015). These different definitions are silo-ed, but they share certain elements, like the trait of performance.

The focus will be placed on the performance component of star employees, as it is the most visible within an organization and the clearest to demonstrate to others. Within performance, the authors described three subcategories: high performers, who perform in the top 10% of their age group (Gallardo-Gallardo, Dries, & Gonzalez-Cruz 2013), experts who have mastered a subfield of their domain (Ericsson & Lehmann 1996), and one-hit wonders, who achieve high performance for a short term, without sustaining it (Marshall 2014). High performance is evaluated differently based on how skill-dependent it is, rather than luck-dependent, as well as how sustainable it is, so these definitions are important cues of the different kinds of stars. Within this categorization, they describe two sub-dimensions: colleague and organizational effects, the star's impact on their peers and their institution, respectively (Call et al. 2015). Call's conceptual review is an important framework into how academia thinks about superstars, and pairs nicely with other research into the different kind of stars that exist within an organization.

The framework in Appendix A is a valuable tool to summarize the three components of star employees, as it draws from several different academic disciplines, though it lacks an important consideration, namely how these factors are affected by the demographics of said superstar. While the performance of the star may be the same, the visibility of their performance

could be more limited, because it may be dismissed as a one-time performance, or they may be expected to do less visible tasks, known as office housework (Williams & Multhaup, 2018). Additionally, these stars from lesser represented backgrounds may have less access to social capital at work because there are fewer people like them in the workplace. This framework is an important representation of the elements of superstardom, but through this study, one can attain better understanding of the intersectionality of superstardom, and how race and gender affect it.

Like Call's conceptual review, Kehoe, Lepak, and Bentley created a new typology for star employees based on their work performance and perception by others in 2016. They observe three distinct kind of stars, who all create different sources of value. Universal stars have high task performance and broad external status, providing reputation, mentoring, and knowledge, like a star scientist with many citations. Performance stars have exceptional task performance without external status and can provide knowledge transfer, like an associate at a law firm who does excellent work, but only internally. Status stars have external status without exceptional performance and can provide opportunities and sponsorship, like a celebrity.

There are three sub-categories of status stars, affiliation-based, former universal, and networking. Affiliation-based stars achieve star status through their association to another, like an actor who is famous because their parent is a famous director. A former universal is a star who has external status because of their previous exceptional performance and has now shifted their focus, like an academic who published a lot but now focuses on public appearances. A networking star is a star because of their network development and management skill, like a wireless communications professional who have influence on the industry's development of policies. Through these typologies, the different paths to stardom will be elucidated, based on exceptional task performance and broad external status, and the different categories within these

headings (Kehoe, Lepak, & Bentley 2018). This typology presents a different way of conceptualizing superstars and paired with Call's framework, further elucidates the attribute of exceptional performance, doing better than others.

By combining the performance component of the performance component of the star employee framework and the performance star from the star employee typology, one has an empirical definition of star based on their performance, who achieved without garnering external acclaim. This performance star will be presented to external participants in the experiment under different conditions of race to understand how people perceive and evaluate this person.

Based on this research, several hypotheses were created to evaluate understand the impact of participant demographics, the race of a high performing partner, and the performance of the participant on the tasks on the evaluation of a high performing female colleague (and are summarized in Figure1):

> *Hypothesis 1 (Marginalization Bias):* White participants will give lower ratings to the black superstars than the white superstar. And white men will give the lowest ratings to black female superstars, because they do not share a race or a gender and have two marginalized identities.

Much research has been done into unconscious bias and overt discrimination and these factors heavily impact perceptions of women at work. Much work has been done to demonstrate the existence of a white bias in evaluation of expected leadership performance, which is closely tied to perception of high performance (Rosette, Leonardelli, & Phillips 2008). Similarly, there is documentation of a gender bias in perception of leadership behavior as well, so these biases

apply to multiple categories of identity (Scott & Brown 2006). Black female superstars face a kind of double whammy, whereby they hold two marginalized identities and face the judgments about both, so they will likely have lower ratings relative to white female superstars (Epstein 1973).

Hypothesis 2 (Participant Likeness Bias): Participants who share a race or gender with their superstar partner will give higher ratings to them.

One factor that could affect the evaluation of the superstar is the demographics of the participant reviewing them. The experimental design will test for a kind of in-group effect (Crocker & Luhtanen, 1990), whereby participants who share a race or gender with their female high performer will rate them as more likeable and more competent. A result that is expected because people tend to like those who are similar to them and have evolutionary incentives to help in-group members. Participants can relate to the superstar they are matched with based on race and gender and between these two, I would expect the race effect to be more salient because it is a more marginalized one in the workplace.

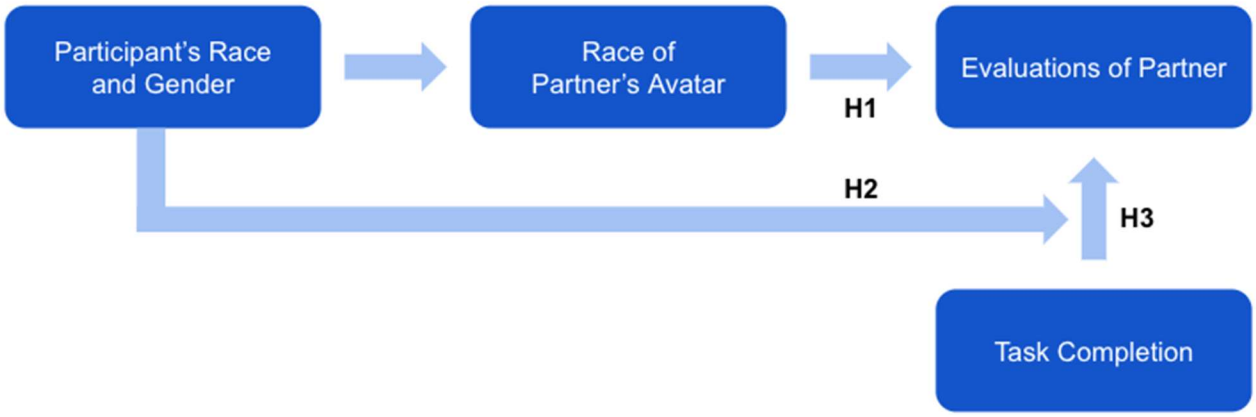
Hypothesis 3 (Participant Performance Bias): Participants who did better on their tasks will rate the superstar higher.

The questions about the superstar's performance will be asked immediately following the experimental tasks, so the relationship between the participant's scores on the tasks and their subsequent evaluations can be measured. Logically, if a participant does well on a task, when

they engaged with help from their superstar partner they are likely to attribute their performance in part to the help they receive, contingent on them finding the help useful. There may also be a carryover effect, whereby if the participant does well on the task, they will have a more favorable perception of it and feel better, causing them to be more generous in their evaluation of their partner. These three hypotheses are summarized in Figure 1 below.

Figure 1

Study Design



Pilot Study

Methods

Study Design. An experiment was constructed in which people's perceptions of female high performers are measured under different conditions of race and task nature. To test certain elements of the hypotheses, a 2 (Participant demographics: male, female) x 2 (Avatar color: white vs. black) x 2 (Task: superadditive vs. emergent) factorial between-subjects experiment design was utilized. Participants were randomly assigned to complete 2 rounds of superadditive tasks or emergent tasks. In the first round, they played their task by themselves. Then, they were matched with a partner, and presented information about that partner's race, gender, and previous task performance. For this study, I chose to restrict the avatars used to represent the female

superstars to white or black, so I could make comparisons between different races, while having sufficient data for both races I chose to include. In the second round, they played a similar task, but now had the option to use clues provided to them by their partner. Lastly, they were asked questions about their evaluation of their partner, the effectiveness of the clues, and their feedback on the study.

Participants. 62 participants consented, passed the screening checks, and completed the entire survey via Amazon MTurk. The participants were 53% male and 77% Caucasian. Of the participants, 29 were assigned to the Emergent task and 33 were assigned to the Superadditive task. They were restricted to people living in the US aged 18-64 who were using a desktop or laptop to fill out the survey. While attention check questions were included in the study, they were not used for purposes of exclusion from the analysis, as this study was exploratory. A breakdown of the participants and the tasks they were assigned to can be seen in Table 1.

Table 1

Participant Demographics for Pilot Study

Game	*Partner Race*	*Gender*	Count
− LSAT	− B	Female	6
		Male	8
	− W	Female	8
		Male	7
LSAT Total			29
− SB	− B	Female	7
		Male	10
	− W	Female	7
		Male	9
SB Total			33
Grand Total			**62**

Procedure. Participants accessed the link via an MTurk bulletin. They were first introduced to the objective and methods of the study, and asked to provide informed consent (see Appendix B). They were then asked screening questions to ensure they did not have previous experience with ordering questions from the LSAT or the Spelling Bee from the New York Times. Next, they were asked to answer demographic questions about their race and gender. Participants were then asked to select the avatar that they felt best represented them based on their gender (see Appendix C) and asked to create a username that would be displayed when they are matched with a partner. After that, they received instructions for round 1 of the first task and were given 5 minutes to complete it. Next, they were presented with a loading animation to represent the matching process for 30 seconds, at which point they received the username, avatar, and relative performance of their partner. Next, they were asked to create 3 clues for their partner to use, and once they did so, they were presented with round 2 of their task, now with the option to make use of clues. Once they completed the task, they would be directed to answer 3 sets of Likert scale questions: judgments of partner (see Appendix D) attribution of partner's ability (see Appendix E), and attribution of partner's performance (see Appendix F). They were then asked to rate the helpfulness of each clue they selected from a scale of 1-7 and tested on the username, avatar, and relative performance of their partner. Finally, they were asked to guess the topic of and provide feedback on the study.

Tasks. The participants will be taking part in two different kinds of tasks, superadditive and emergent ones. These two terms come from the discipline of complex systems and describe two situations where having a group rather than an individual attempt a problem are beneficial. Whether they are superadditive, when the group outperforms the average individual at a creative task (Page, 2007), or emergent, when the group demonstrates a capability not possessed by the

individuals that comprise the group to solve a problem (Hong, Page, & Riolo, 2012). By assessing the perception of the female high performer under two different kinds of tasks, one can gain insight into whether one judges high performers differently based on the nature of the problem they are working together to solve.

For the superadditive task, the New York Times' *Spelling Bee Game* (Ezersky 2019) design was adapted to fit the study as a creative activity where participants would have to generate a list of words using letters from a 7-letter word. First, they were provided with the rules of the game: words must contain at least 4 letters, words must include the center letter, and letters can be used more than once. Then, they were informed of the scoring system: every 4-letter or longer word earns 1 point. For round 1 they were presented with a honeycomb of a 7-letter word (see Appendix G) and told to list as many words as they could think of using letters from that word, then repeated the same task for round 2, using a different honeycomb of a 7-letter word (see Appendix H).

For the emergent task, a sample LSAT ordering question was adapted to fit the study as a logic-based question with a single correct answer. First, participants were provided with the rules of the game: each element has one correct position, there is one overall correct order, and if an element must be directly before an element then there must be no elements in between them. Then, they were informed of the scoring system: every element that is in the correct position will earn 1 point, even if the overall order is incorrect. For round 1 they were presented with a ordering question (see Appendix I) and told to place the elements in the correct order, then received the same instructions for round 2, using a different ordering question (see Appendix E).

Clue Mechanic. Because the partner was not real, there needed to be the option for engagement in a pre-established way during round 2 of the tasks. The clue mechanic sought to

serve this function, because the clues were said to have been created by one's partner based on their experiences from round 1, so using them would represent a form of interaction. The clues were set to appear in a random order, drawing from a list of 3 for the superadditive (See Appendix O) and emergent (see Appendix P) tasks. The participant would have the chance to select one of these clues every minute for the first 3 minutes of the task, but using a clue would subtract 30 seconds from the 5 minutes allotted for the task. Participants were free to choose anywhere from 0 to 3 clues, so clue usage was completely voluntary. If a participant chose to reveal a clue, it would appear and stay on the screen for the duration of the task.

Partner Simulation. Because this was an online study, the high-performing female partner that the participant was matched up with had to be simulated. To do this, participants were presented with several pieces of information about the partner they were matched up with. They were shown an avatar the partner was said to have selected (see Appendices K and L), that corresponded with the race and gender of their superstar partner. They were shown a username that the partner was said to have selected, which was generated by combining a name selected from a list 5 of the top 20 most popular names by race with a "_1995" to make it appear more like a username than just a name (see Appendices M and N). They were also provided information about how the partner performed in the last round relative to them. For the superadditive task, participants were told their partner found 5 more words than they did in round 1. For the emergent task, if the participant completed round 1 in fewer than 2 minutes, they were told their partner was 30 seconds faster, but if they took longer than 2 minutes, they were told their partner was 1 minute faster.

Results

With the limited sample size employed by this study, the goal was to gain insight into which areas demonstrate relevant effects on the evaluation of female high performers. So, the focus of these analyses will be more on refining the study design rather than deriving significant effects. To create a single summary metric to represent the participant's overall evaluation of the superstar, the judgments of partner index was converted to a single metric by averaging the responses, called Likert Average.

First, the data were analyzed by summarizing the Likert average score by gender for each task. A relevant factor to the analysis was whether the participant used clues or not. Logically, if a participant did not use the clues they did not engage with their partner, so it would be difficult for them to rate the strengths of their partner. The data support this, because the Likert scale evaluation for participants who did not use clues had an average of 4.30 with a standard deviation of .38, meaning most participants were effectively rating every element of the high performer as "neither agree nor disagree." Based on this, the when analyzing Likert average scores, only participants who used clues will be included. For the emergent task, men had an average Likert score of 4.47 out of 7 with a standard deviation of 1.38, while women had an average of 5.06, with a standard deviation of 1.22. For the superadditive task, men had an average of 5.15 with a standard deviation of .89, while women had an average of 4.87, with a standard deviation of 1.16. So, the emergent task had more of a gender effect for the Likert average measure. Also, the count for each task is different, because a much larger proportion of participants used clues in the superadditive task, as seen in Table 2.

Table 2

Likert Average by Gender for Pilot Study

Game	Gender	Likert_Avg	Count
− LSAT	Female	5.06	9
	Male	4.47	8
LSAT Total		4.78	17
− SB	Female	4.87	13
	Male	5.15	17
SB Total		5.03	30
Grand Total		**4.94**	**47**

For the emergent task, participants matched up with a black avatar had a Likert average of 4.69 with a standard deviation of 1.49, while participants matched up with a white avatar had a Likert average of 4.88, with a standard deviation of 1.12. Whereas for the superadditive task, participants matched up with a black avatar had a Likert average of 5.00 with a standard deviation of .98, while participants matched up with a white avatar had a Likert average of 5.05, with a standard deviation of 1.07, showing a minimal difference in evaluation based on the superstar. A more detailed output can be seen below in Table 3.

Table 3

Likert Average by Partner Race for Pilot Study

Game	Partner Race	Likert_Avg	Count
− LSAT	B	4.69	9
	W	4.88	8
LSAT Total		4.78	17
− SB	B	5.00	15
	W	5.05	15
SB Total		5.03	30
Grand Total		**4.94**	**47**

Another analysis conducted was measuring the interaction effect between gender and partner's race for each of the tasks. There only appears to be slight differences between each of

the conditions, except in the case of women evaluation black partners, where there is a difference between men and women's Likert averages of 1.25, with a standard deviation of 1.49, as seen in Table 4. An important caveat is the small count for each condition, as these groups represent small subsections of an already small pilot study, but the potential differences between these groups indicate the importance of interaction effects.

Table 4

Likert Average by Conditions for Pilot Study

Game	*Partner Race*	*Gender*	Likert_Avg	Count
− LSAT	− B	Female	5.25	5
		Male	4.00	4
	− W	Female	4.81	4
		Male	4.94	4
LSAT Total			4.78	17
− SB	− B	Female	4.88	6
		Male	5.08	9
	− W	Female	4.86	7
		Male	5.22	8
SB Total			5.03	30
Grand Total			**4.94**	**47**

Participants used at least one clue 58.62% of the time for the emergent task and 90.91% for the superadditive task, so there was dramatically more variation in clue usage for the emergent task. For the emergent task, solution percentage increased from 37.93% to 58.62%, while the average number of words found in the superadditive task increased only from 11.85 to 11.91, demonstrating a minimal difference. Within the emergent task, the clue ratings were 4.10 for the 'Keep' clue, 3.69 for the 'if' clue, and 3.58 for the 'combine' clue (see Appendix P), indicating the 'Keep' clue may be superior to the other two.

Related to solution rates, while the superadditive task has a continuous scoring system, whereby participants can score anywhere from 0 to any number of words, while the emergent task has a binary outcome, as the successful order is either reached or not. Additionally, because there is a minimal difference in performance between the two rounds of superadditive tasks, it is hard to compare participants based on performance. However, when analyzing the emergent task, task performance appears as a potential moderator of partner evaluations, as can be seen in Table 5. Interestingly, the highest ratings of the superstar occurred when the participant was unable to solve either of the emergent rounds, suggesting that participants may be more impressed by a high-performing partner, when they themselves struggle with the task.

Table 5

Likert Average by LSAT Performance for Pilot Study

Game	*# LSAT Rounds Solved*	Likert_Avg	Count
− LSAT	0	5.28	8
	1	4.41	14
	2	4.25	7
LSAT Total		4.61	29

Discussion

Based on the results from the pilot study, there are several possible conclusions about eh intersection of gender and performance on the Likert average. Firstly, the Likert average was much higher when participants were not able to solve a single LSAT round, suggesting that participants give higher ratings to partners when the task is more difficult for them (though there is less of a drop-off between 1 and 2 rounds solved). In terms of Likert average differences by gender and avatar color, females seem to give much higher ratings to black avatars than males

do, specifically for the emergent task. Though there is a gender difference in Likert averages for both tasks, though it is higher for the emergent one than the superadditive one.

Due to the high clue usage and the minimal change in task performance between rounds, the superadditive task appeared to not hold relevant results pertaining to perceptions of one's partner, so it will be removed from the next study. Additionally, there appears to be a hierarchy to the usefulness of the clues, so it was decided to fix the order of them to 'Keep', 'If', 'Combine' (See Appendix O), to ensure the superstar would be perceived as a competent partner because she would always start with a better clue. In addition, the race data was difficult to analyze, because many people identified as multiple races and the participants were mostly Caucasian, so the participant pool would have to either have be vastly increased to account for homogeneity, or it should be more controlled to get a sufficiently diverse sample for analysis.

Additionally, in the open feedback section, multiple participants noted that they were paying attention, but nonetheless had trouble remembering the name of their partner. Because the name of one's partner is less relevant to the study's goal than the race and superior performance of one's partner, it was decided to focus on removing participants who did not try on the tasks or provided the same answer to all the Likert scale questions.

There was also a slight bug with the conditional logic of the attention check for timing of the first round, because if participants did not complete round 1, it would be illogical to state that the partner performed 30 seconds or 1 minute faster than the participant. Based on this, a display logic was added to tell the participant that their partner finished round 1 in 4 minutes if the participant could not complete round 1. If the participant did not complete round 1, they were told their partner completed round 1 in 4 minutes (these participants were also asked about how long it took their partner as an attention time check instead).

Study 1

Methods

Study Design. Based on the results of the pilot study, the survey was updated to only include the emergent task. Additionally, because race and gender interactions are crucial to the understanding of perceptions of high performers, the participants mix was controlled to have an equal balance of participants across white males, black males, white females, and black females, roughly equal subsets of whom would be matched with each condition of the partner avatar, black or white. Also, the clue order was fixed in the order of 'Keep', 'Combine', 'If', based on the results of the pilot study to ensure the participants received the highest rated clue first (see Appendix P). To test the hypotheses from Figure 1, I utilized a 4 (Participant demographics: white male, black male, white female, black female) x 2 (Avatar color: white vs. black) x 1 (Task: emergent) factorial between-subjects experiment. Like the pilot study, participants would be randomly assigned to complete a task, be randomly matched up with either a black or white avatar to represent their partner, then invited to complete a second round of their task, but now with the option to use clues.

Participants. 202 participants consented, passed the screening checks, passed the attention checks, and completed the entire survey via Amazon MTurk. The participants were 50% male and 50% female, 51.49% black and 48.51% white. Of the participants, 102 (50.50%) were assigned to a black avatar and 100 (49.50%) were assigned to a white avatar. They were restricted to people living in the US aged 18-64 who were using a desktop or laptop to fill out the survey. Of note, this study was conducted amid the Covid-19 pandemic. So, with the stay at home orders and other containment measures, the demographics of the participants who

comprised the study may reflect a larger proportion of people who normally would not have been home.

11 Participants from the 213 that completed the survey were excluded from the analysis for reasons of quality, suspicion, and attention, based on reasons that were decided before the analyses were conducted. The quality criteria was that the participants all faced the same conditions. For quality, 6 (2.82%) participants were removed because they received the clues in an order other than the 'Keep', 'Combine', 'If' order that the rest of the participants did. The criteria for suspicion was whether the participant expressed doubt as to the validity of the partner, because if they did not believe the partner was real, it would affect how they evaluate the partner. For suspicion, 3 (1.41%) participants were removed because they expressed doubt that the participant was real or belief that the participant was a robot. For the attention criteria, it was decided to remove anyone who failed to complete either of the rounds and gave the same Likert score for every question in the judgments of partner index (the Likert average questions). For attention, 2 (0.94%) participants were removed because they failed to complete either round and provided the same response for the Likert average questions. A breakdown of the participants and the condition of avatar color they were assigned to can be seen in Table 6.

Table 6

Participants by Condition for Study 1

Participant_Genc	Participant_Race	Avatar_Color	Count
− female	− black	B	27
		W	27
	black Total		54
	− white	B	24
		W	23
	white Total		47
female Total			101
− male	− black	B	24
		W	26
	black Total		50
	− white	B	27
		W	24
	white Total		51
male Total			101
Grand Total			**202**

Procedure. Participants accessed the link via an MTurk bulletin. They were first introduced to the objective and methods of the study, and asked to provide informed consent (see Appendix B). They were then asked screening questions to ensure they did not have previous experience with ordering questions from the LSAT. Next, they were asked to answer demographic questions about their race and gender. Participants were then asked to select the avatar that they felt best represented them based on their gender (see Appendix C) and asked to create a username that would be displayed when they are matched with a partner. After that, they received instructions for round 1 of the emergent task and were given 5 minutes to complete it. Next, they were presented with a loading animation to represent the matching process for 30 seconds, at which point they received the username, avatar, and relative performance of their

partner from round 1. Next, they were asked to create 3 clues for their partner to use, and once they did so, they were presented with round 2 of their task, now with the option to make use of the clues created by their partner. Once they completed the task, they would be directed to answer 3 sets of Likert scale questions: judgments of partner (see Appendix D) attribution of partner's ability (see Appendix E), and attribution of partner's performance (see Appendix F). They were then asked to rate the helpfulness of each clue they selected, if any, from a scale of 1-7 and tested on the username, avatar, and relative performance of their partner. Finally, they were asked to guess the topic of and provide feedback on the study. Additionally, they were asked to provide what they thought the differences were between the two ordering games they played, as an additional attention check.

Tasks. Participants were presented instructions for the emergent task (the LSAT ordering question): each element has one correct position, there is one overall correct order, and if an element must be directly before an element then there must be no elements in between them. Then, they were informed of the scoring system: every element that is in the correct position will earn 1 point, even if the overall order is incorrect. For round 1 they were presented with an ordering question (see Appendix I) and told to place the elements in the correct order, then received the same instructions for round 2, using a different ordering question (see Appendix J). These questions were identical to the pilot study, and appeared in the same order.

Partner Simulation. Participants were shown an avatar the partner was said to have selected (see Appendices K and L), that corresponded with the race and gender of their superstar partner. They were shown a username that the partner was said to have selected, which was generated by combining a name selected from a list 5 of the top 20 most popular names by race with a "_1995" to make it appear more like a username than just a name (see Appendices M and

N). They were also provided information about how the partner performed in the last round relative to them. If the participant completed round 1 in fewer than 2 minutes, they were told their partner was 30 seconds faster, but if they took longer than 2 minutes, they were told their partner was 1 minute faster. If the participant failed to complete round 1, they were told their partner completed round 1 in 4 minutes.

Clue Mechanic. The clues were fixed to appear in the order of 'Keep', 'Combine', then 'If' (see Appendix P). The participant would have the chance to select one of these clues every minute for the first 3 minutes of the task, but using a clue would subtract 30 seconds from the 5 minutes allotted for the task. Participants were free to choose anywhere from 0 to 3 clues, so clue usage was completely voluntary. If a participant chose to reveal a clue, it would appear and stay on the screen for the duration of the task.

Partner Simulation. Like the pilot study, participants were presented with several pieces of information about the partner they were matched up with. They were shown an avatar the partner was said to have selected (see Appendices K and L), that corresponded with the race and gender of their superstar partner. They were shown a username that the partner was said to have selected, which was generated by combining a name selected from a list 5 of the top 20 most popular names by race with a "_1995" to make it appear more like a username than just a name (see Appendices M and N). They were also provided information about how the partner performed in the last round relative to them. For the emergent task, if the participant completed round 1 in fewer than 2 minutes, they were told their partner was 30 seconds faster, but if they took longer than 2 minutes, they were told their partner was 1 minute faster. If the participant was unable to complete round 1, they were told that their partner had completed round 1 in 4 minutes.

Results

Primary Analysis. The main focus of the data analysis was to measure how participants reviewed their superstar partners, represented by the Likert average metric. An ANOVA for Likert Average presented, and its main effects are summarized. From there, an in-depth analysis of the 3 statistically significant simple effects: participant gender x participant race, participant race x avatar color, and participant gender x task performance is conducted. Next, the three hypotheses from the theoretical framework will be evaluated based on the data collected.

A four-way ANOVA was conducted on the Likert Average metric, based on participant gender (male or female), participant race (black or white), avatar color (black or white), and task performance (0, 1, or 2 rounds completed successfully) for participants who used clues. This output can be seen below in Table 7.

Table 7

4-way ANOVA for Likert Average by Conditions for Study 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Participant_Gender	1	0.09	0.091	0.128	0.72130	
Participant_Race	1	0.88	0.879	1.238	0.26812	
Avatar_Color	1	0.36	0.365	0.514	0.47496	
Task_Performance	1	0.07	0.068	0.096	0.75762	
Participant_Gender:Participant_Race	1	7.52	7.518	10.583	0.00149	**
Participant_Gender:Avatar_Color	1	0.50	0.499	0.702	0.40387	
Participant_Race:Avatar_Color	1	3.01	3.012	4.240	0.04170	*
Participant_Gender:Task_Performance	1	3.65	3.649	5.136	0.02527	*
Participant_Race:Task_Performance	1	0.67	0.673	0.948	0.33233	
Avatar_Color:Task_Performance	1	0.32	0.320	0.450	0.50346	
Participant_Gender:Participant_Race:Avatar_Color	1	0.03	0.034	0.048	0.82784	
Participant_Gender:Participant_Race:Task_Performance	1	0.00	0.000	0.000	0.99572	
Participant_Gender:Avatar_Color:Task_Performance	1	0.11	0.109	0.153	0.69635	
Participant_Race:Avatar_Color:Task_Performance	1	1.36	1.358	1.912	0.16938	
Participant_Gender:Participant_Race:Avatar_Color:Task_Performance	1	0.03	0.031	0.043	0.83608	
Residuals	117	83.11	0.710			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

None of the main effects were statistically significant, suggesting that none of the variables included in the model had a large impact on the overall Likert average by themselves. This means neither the race nor the gender of the participant is individually responsible for

statistically significant changes in the Likert average. This is significant because avatar color by itself did not have a significant impact on partner evaluation, so there does not appear to be a simple bias against one race in particular. Similarly, task performance is not a significant predictor by itself, demonstrating that doing well or poorly on the test does not have a large impact on evaluations of the partner. There were however, 3 statistically significant simple effects: participant gender x participant race, participant race x avatar color, and participant gender x task performance.

In Table 8, the Likert averages for each condition of participant gender, participant race, and avatar color can be observed for participants who used clues.

Table 8

Likert Average by Conditions for Study 1

Avatar_Color	*Participant_Gender*	*Participant_Race*	Likert_Avg
− B	− female	black	4.94
		white	4.63
	female Total		4.78
	− male	black	4.72
		white	5.23
	male Total		4.99
B Total			4.90
− W	− female	black	5.52
		white	4.48
	female Total		5.07
	− male	black	4.97
		white	4.96
	male Total		4.96
W Total			5.01
Grand Total			**4.96**

Participant Gender x Participant Race Interaction. There was however, a significant interaction effect between participant gender and participant race, with a p-value of .00149 < .01. This suggests that the intersection of one's race and gender impact have a significant impact on how they review the superstar, for both conditions of avatar color. This underscores the importance of intersectionality in this research, as the effects on perception of one's partner are contingent on multiple, intersecting identities of a participant.

Based on the significant interaction between participant gender and participant race, the effect was broken down by the race of the participant. Through a linear regression for Likert average for white participants, a p-value of .01137 < .05 was achieved. However, when the same analysis was applied to black participants, the p-value was .14732, not significant, suggesting the gender effect is stronger for white participants.

To interpret this data better, t tests were performed based on the different subgroups in the data, gender and race. A 2-sample difference of means t-test was performed on the Likert averages by gender, and the p-value was .7401 > .05, so the null hypothesis that there is no gender difference fails to be rejected. A 2-sample difference of means t-test was performed on the Likert averages by race, and the p-value was .2957 > .05, so the null hypothesis that there is no gender difference fails to be rejected.

However, when the gender and race differences are broken down by the different conditions, there are significant differences. A 2-sample difference of means t-test was performed on the Likert averages by gender for white participants, and the p-value was .01833 < .05, so the null hypothesis is rejected. A 2-sample difference of means t-test was performed on the Likert averages by gender for black participants, and the p-value was .04954 < .05, so the

null hypothesis is rejected. These results indicate that gender is a relevant factor when evaluating partners, but only for within race measurements.

In addition, gender's impact for within-race differences were measured. A 2-sample difference of means t-test was performed on the Likert averages by race for male participants, and the p-value was .145 < .05, so the null hypothesis fails to be rejected. A 2-sample difference of means t-test was performed on the Likert averages by race for female participants, and the p-value was .00661 < .01, so the null hypothesis is rejected. For within-race comparisons, there are significant differences between white and black female participants, but not for male participants. This could be because black women themselves are more familiar with the experience of the "double whammy," so they give higher ratings to their partners.

Participant Race x Avatar Color Interaction. There was also a significant interaction effect between participant race and avatar color, with a p-value of .0417 < .05. Interestingly, there was no significant interaction effect between participant gender and avatar color, suggesting the race interaction is stronger than the gender interaction in regards to avatar color. This demonstrates that the relationship between the race of the avatar and the race of the participant affects the Likert average.

First, the effect of participant race was measured when participants were presented with the same avatar color condition to observe the resulting differences in the Likert average. A 2-sample difference of means t-test was performed on the Likert averages by race for participants who had a white avatar, and the p-value was .03283 < .05, so the null hypothesis is rejected. A 2-sample difference of means t-test was performed on the Likert averages by race for participants who had a black avatar, and the p-value was .40690 > .05, so the null hypothesis fails to be

rejected. The race of the participants has a statistically significant effect on the Likert average, but only for the white avatar.

Next, the effect of the avatar color was measured when participants were the same race to observe the resulting differences in the Likert average. A 2-sample difference of means t-test was performed on the Likert averages by avatar color for black participants, and the p-value was .03509 < .05, so the null hypothesis is rejected. A 2-sample difference of means t-test was performed on the Likert averages by avatar color for white participants, and the p-value was .32040 > .05, so the null hypothesis fails to be rejected. The color of the avatar has an impact on the Likert average for the partner, but only for black participants.

Participant Gender x Task Performance Interaction. The interaction effect between participant gender and task performance is also significant, with a p-value of .02527 < .05. In this case, the gender interaction is more impactful than the race interaction in regards to task performance. This finding suggests that how there is a relationship between how one performs on the task and one's gender. While the ANOVA model used three levels of the Task performance metric, for the sake of simplicity, the t-tests will consider only two levels: 0 for participants who solved neither task and 2 for participants who solved both tasks.

To begin, the effect of participant gender was measured when participants performed equally well on the emergent tasks to observe the resulting differences in the Likert average. A 2-sample difference of means t-test was performed on the Likert averages by gender for participants who did not solve either task, and the p-value was .10450 > .05, so the null hypothesis fails to be rejected. A 2-sample difference of means t-test was performed on the Likert averages by gender for participants who solved both tasks, and the p-value was .01057 <

.05, so the null hypothesis is rejected. The participant's gender impacted the evaluation of one's partner, but only for participants who solved both tasks.

Next, the effect of task performance was measured when participants had the same gender to observe the resulting differences in the Likert average. A 2-sample difference of means t-test was performed on the Likert averages based on whether both tasks were completed for male participants, and the p-value was .06736 > .05, so the null hypothesis fails to be rejected, though the p-value is approaching significance. A 2-sample difference of means t-test was performed on the Likert averages based on whether both tasks were completed for female participants, and the p-value was .01392 < .05, so the null hypothesis fails is rejected. Completing both tasks is important to the Likert average for female participants, but slightly less so for male participants.

Task Performance. In Table 9, the percentage of participants who were able to solve the emergent task for each condition of participant gender, participant race, and avatar color can be observed for participants who used clues.

Table 9

Solution Percentages by Condition for Study 1

Participant_Race	*Participant_Gender*	*Avatar_Color*	Task 1 Solve %	Task 2 Solve %
− black	− female	B	28.57%	35.71%
		W	11.76%	11.76%
	female Total		19.35%	22.58%
	− male	B	27.78%	44.44%
		W	11.76%	23.53%
	male Total		20.00%	34.29%
black Total			19.70%	28.79%
− white	− female	B	33.33%	73.33%
		W	38.46%	38.46%
	female Total		35.71%	57.14%
	− male	B	71.43%	52.38%
		W	55.56%	66.67%
	male Total		64.10%	58.97%
white Total			52.24%	58.21%
Grand Total			**36.09%**	**43.61%**

While the overall improvement in percentage of participants who were able to solve the task is a modest 7.52% between round 1 and round 2, there exists significant variation between the different conditions. Improvement rates by race were 5.97% for white participants and 9.09% for black participants. Improvement rates by gender were 4.05% for male participants and 11.86% for female participants. Improvement rates by avatar color were 6.15% for white participants and 8.82% for black participants.

Notably, white females assigned a black partner had a change of 0% in their solution percentages, but when assigned to a white partner, they increased by 40%. Similarly, white males

assigned to a black partner had a change of -19.05% in their solution percentages, but when assigned to a white partner, they increased by 11.11%. These findings indicate that white participants performed much better on the second round when matched with a white participant, but this same performance improvement for a racial in-group member did not exist for black participants.

Clues. In addition to task performance and Likert average, another metric by which to analyze the relationship between the participant and their partner is how often they used clues and what rating they gave to them. In Table 10, the percentage of participants who used at least one clue for each condition of participant gender, participant race, and avatar color can be observed.

Table 10

Clue Usage by Conditions for Study 1

Participant_Gender	*Participant_Race*	*Avatar_Color*	Average Clue Usage
− female	− black	B	51.85%
		W	62.96%
	black Total		57.41%
	− white	B	62.50%
		W	56.52%
	white Total		59.57%
female Total			58.42%
− male	− black	B	75.00%
		W	65.38%
	black Total		70.00%
	− white	B	77.78%
		W	75.00%
	white Total		76.47%
male Total			73.27%
Grand Total			**65.84%**

Clues act as a proxy for engagement with the partner, so whether participants used them is an indication of their willingness to accept help from their partner. Men used clues 14.85% more than women, and there were not considerable differences based on participant race or avatar color. In addition to using clues, participants were also asked to rate the clues they received, as Table 11 displays the ratings for each clue for each condition of participant gender, participant race, and avatar color can be observed.

Table 11

Clue Ratings by Condition for Study 1

Participant_Race	*Participant_Gender*	*Avatar_Color*	Keep Clue	If Clue	Combine Clue
− black	− female	B	3.29	3.57	3.67
		W	4.56	4.38	4.45
	female Total		3.97	4.10	4.18
	− male	B	4.19	3.75	3.70
		W	4.53	5.00	5.00
	male Total		4.35	4.33	4.13
black Total			4.16	4.20	4.16
− white	− female	B	3.00	3.13	4.00
		W	2.83	3.22	3.25
	female Total		2.93	3.18	3.63
	− male	B	4.05	4.36	4.63
		W	3.61	4.90	4.17
	male Total		3.84	4.58	4.43
white Total			3.46	4.00	4.14
Grand Total			**3.80**	**4.09**	**4.15**

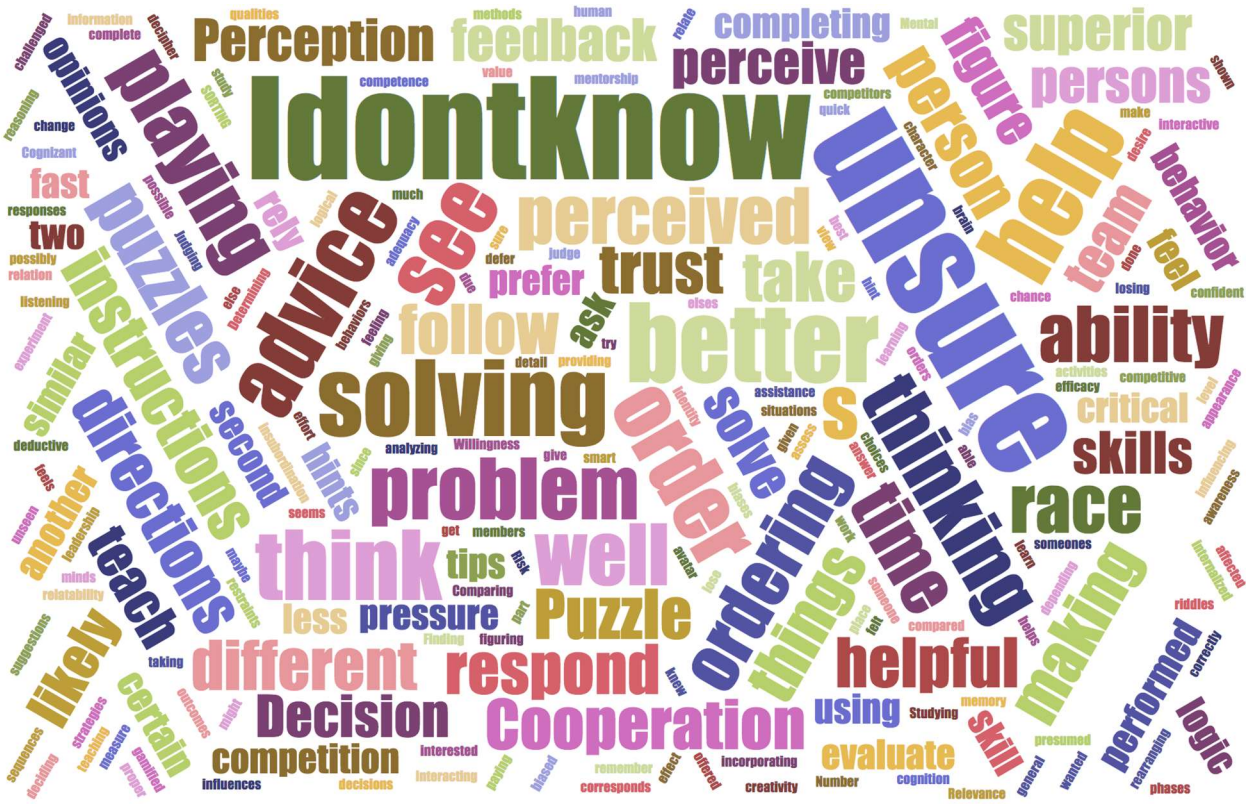
Interestingly, the clue averages are all centered around 4 out of 7, a stark departure from the pilot study, where the 'Keep' clue was vastly superior. For white participants of both genders, there is a minimal difference between clue ratings for each avatar color condition, with an average clue rating .20 higher for black avatars. However, for black participants of both

gender, considerably high ratings were given to white avatars than black ones, with an average clue rating .96 lower for black avatars.

Secondary Analysis. In additional to the above quantitative inventories, participants were also asked open response questions about the differences between rounds, the topic of the study, and for their feedback on the study. The topic of the study response had a variety of answers, mainly centered around problem-solving, cooperation, race perception, and advice taking. I attached a tag cloud diagram, as seen below in Figure 2. For the purpose of visualization, the data were cleaned to remove prepositions, subjects, and words related to the logistics of the experiment. Also, responses expressing doubt about whether they were correct were converted to "unsure" and responses that indicated a complete lack of knowledge were converted to "Idontknow."

Figure 2

Tag Cloud for Predicted Study Topic for Study 1



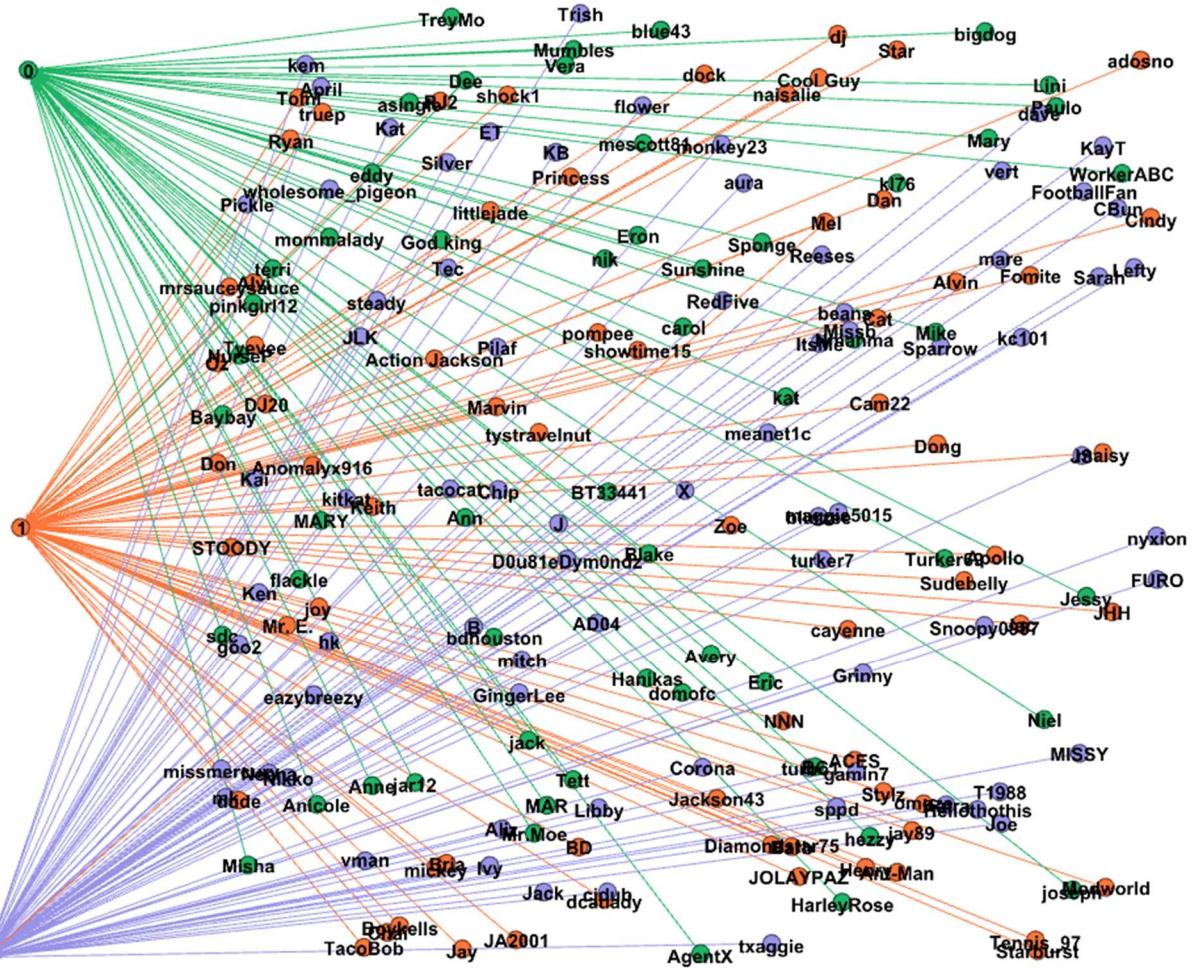
Davies, J. (n.d.). Word Cloud Generator. Retrieved from

https://www.jasondavies.com/wordcloud/

Network visualizations were also created to capture the effects of race and gender on overall task performance. Below in Figure 3 is one such visualization, a network graph plotting the subjects by their overall task performance (0, 1, or 2 rounds solved correctly). It provides a a visual representation of task performance, with each color corresponding to a level of performance, green for 0 tasks solved, orange for 1 task solved, and purple for 2 tasks solved.

Figure 3

Network Diagram of Tasks Solved by Username



Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Discussion

The purpose of these two studies was to investigate the interactions between demographic elements and task performance of people and how it affects their evaluations of high performing female partners. First, the gender differences for clue usage, performance, and Likert evaluations for superadditive and emergent tasks were analyzed. Next, the focus was refined to look at race and gender interactions' impact on Likert average, specifically for an emergent task. These results are relevant because they provide insight into how colleagues form evaluations of their

high performing female peers. This area of research is relevant, because it informs the fight to combat unconscious bias at work. Additionally, it analyzes the effects of demographics and performance, so more holistic models of evaluation can be derived, because it includes complex interaction effects.

The study demonstrated the interaction effects of the gender and race of the participant in evaluating their partners. Gender was a significant moderator, but only for white participants. Additionally, for both races of participant, women gave higher overall ratings than men did. Another significant effect was that black female participants gave higher scores than white women. This suggests that women give higher evaluations to female partners, which may be due to an in-group effect, because women are evaluating other women in this case. Interestingly, black women gave higher average ratings than white women, for both conditions of avatar color, but this could be due to the different base rates of task performance across different demographic conditions.

Another significant interaction was participant race and avatar color, demonstrating the effects of in and out group racial perceptions for evaluating a colleague. For this analysis, black participants rated the white avatar significantly higher than the black avatar. This rating is surprising because it represents a group of participants rating an outgroup higher than an in-group. In line with this, black participants also rated the white avatar higher than white participants did, demonstrating the strength of this effect.

The interaction between the gender of the participant and their performance on the task was also statistically significant in terms of how it affected the Likert average evaluation of the female high performing partner. In this case, men who solved both rounds gave higher ratings than women who solved both rounds. This could be because men base their evaluation of others

more on their own performance, so the better they do, the higher they will rate their partner. Within performance, women who solved neither round rated their partner higher than those who finished both rounds. So, women rate their partner higher when they perform worse, possibly because they have more respect for someone who could complete the tasks successfully.

In terms of overall performance, several trends were noted, depending on race, gender, and avatar color. White male participants were the only demographic that got worse from round one to round two. However, when their performance is uncoupled by avatar color, it is revealed that they do better when they are paired with a white avatar, but do worse when paired with a black avatar. A similar effect is observed for women, with a larger magnitude, demonstrating that when participants have an outgroup avatar, their ratings will not change, but their performance decreases from one round to another. White avatars also received much higher clue rating for identical clues than did their black avatar counterparts from black participants, suggesting that the attribution of a clue's value is dependent in part on the race of one's partner.

From the open-response questions, it was gleaned that people had a variety of ideas about the nature of the study. Most thought it had to do with problem or puzzle solving, with a smaller subsection correctly guessing it was focused on perceptions of race in the context of a game. That said, most people said they were unsure or had no idea as to the topic of the study, so this study design was relatively effective in not giving away the purpose of the study.

Returning to the hypotheses, there is now sufficient data and analysis to address them and evaluate their accuracy:

Hypothesis 1 (Marginalization Bias): White participants will give lower ratings to the black superstars than the white superstar. And white men will give the lowest ratings to black

female superstars, because they do not share a race or a gender and have two marginalized identities.

A two sample t-test was run on the difference for the Likert average among white and black participants, and the p-value was .30 > .05, so I fail to reject the null hypothesis that white and black participants rate their partners equally.

Additionally, the Likert average for white males rating black females was the second highest of any demographic pairing, at 5.23 out of 7, so it would appear that white men rate black women higher. This may be a form of socially desirable responding, where white men would not want to give lower ratings to black women, because they are aware that they may be biased. Interestingly, the other opposite demographic pair (black male participant and white female) had a Likert average of 4.97, just .01 above the overall average of 4.96. So this hypothesis was wrong, and it appears that the white male participant – black female avatar pairing resulted in one of the highest overall ratings.

Hypothesis 2 (Participant Likeness Bias): Participants who share a race or gender with their superstar partner will give higher ratings to them.

Black participants gave Likert averages of 4.82 and 5.25 to black and white avatars respectively. White participants gave Likert averages of 4.98 and 4.76 to black and white avatar respectively. So, in both cases having an opposite-race partner resulted in a higher overall Likert average, particularly for black participants. This is the opposite of the rationale of the hypothesis and this could be because participants are more aware of partners who are a different race, so they then given them higher ratings.

Male participants gave a Likert averages of 4.93 to their female partners, while female participants gave an average of 4.98 to their female partners. While male partners were not

included, such that a true gender interaction effect cannot be calculated, there is only a minimal difference in the overall evaluations of the partners based on gender, so the hypothesis was not true for either demographic condition (race and gender).

Hypothesis 3 (Participant Performance Bias): Participants who did better on their tasks will rate the superstar higher.

The Likert Averages for conditions of 0, 1, and 2 rounds solved are 5.08, 4.79, and 5.00 respectively. There is only a minimal difference between these values, so performance does not seem to have a relevant impact on the evaluation of one's partner.

Limitations

This study investigated the interactions of race, gender, and task performance on evaluations of a high performing female colleague. While many interactions were studied, male partners were not used in the study, to increase the focus on marginalized communities, so conclusions about how different participants would have viewed male superstars cannot be drawn. While this means that the gender effect analysis is incomplete, because it can only represent relationships that involve a female partner, this decision was made to increase the statistical power of the analyses and to focus on a demographic that faces more bias and discrimination at work.

Another limitation to the study is the fact that the partners were simulated, and the study was conducted online. It is possible that the participants were not conscious of their partner's race while they were completing the activity. Participants may have responded differently if they were interacting with a live person, where many more judgments could have been made about their partner, rather than simply seeing a picture of an avatar meant to represent their race and gender. It is also possible that participants did not try as hard as they could have on the task

because they were not being watched and there was no monetary incentive for performing well. It would have been better to have conducted this in person under the watchful eye of a research administrator, however, it would have been prohibitively expensive to secure lab space and get a sufficiently diverse sample.

The emergent task used for this study was meant to represent an emergent task that required problem solving and would benefit for clues. However, it may have been too hard for many participants, and there were large differences in the percentage of participants who were able to solve round 1, based on race and gender, so it may have been a biased task. The task was selected because it had a single right answer, it could not be cheated on (because the answer was not online), and there are known strategies for tackling it.

Additionally, only the data of participants who used clues were used for analysis of Likert average ratings, which restricts the range of conclusions that can be drawn from the study. It is possible that there were more people who did not use clues because they were either very competent or biased against the avatar they were paired with, both of which would not be accounted for by the analysis. By selecting for participants who used clues, this study was also inadvertently selecting for all of the moderating factors that affected clue usage.

In the case of the Likert scale evaluations, there may be some bias as to how participants evaluate their partners because they are using a simple quantitative inventory. Prior research has shown that having a greater number of elements on an inventory can decrease women's average score, so adding more qualitative response options, like open response or verbal recordings, can be another way to measure how participants perceived their partners (Rivera & Tilcsik, 2019).

The sample of participants only looked at white and black races in the context of evaluating a partner. This was done to standardize the results and streamline the output.

However, this means there are many demographics that were not included, like Asian, Hispanic or Latino, or Native American. These races have many idiosyncratic elements, so only using two races narrows the group of people for whom this study accurately represents. Additionally, this study only looked at US residents, so it does not represent perceptions of female colleagues in other countries. The limited focus on races other than white or black and inclusion of only US residents limits the generalizability of the study, as its results only accurately represent certain subsections of the population.

Considerations for Future Research

This study's findings contribute to the growing field of research on female high performers, at the intersection of superstardom and diversity and inclusion. However, the approach to task design for this area is far from finalized. Future research directions could use a different kind of emergent task, or operationalize the superadditive task in a different way to make it more conducive for an online study. In terms of simulating a partner, future studies could artificially generate two identical faces with features from different races, which might increase how much the participant perceived the avatar as real. Future studies might analyze the evaluations of the partner on the basis on the warmth / competence trade off, rather than on a single average metric.

Future studies could take place in person, with real participants. This could be a way of verifying the results found in this study and comparing perceptions for an online avatar meant to represent a person as compared to a live person. This design would allow for superior forms of engagement, like working through the problem together or being allotted time to chat directly with one's partner.

If male partners were included, conclusions could be drawn about how male and female high performers are perceived differently. A related line of inquiry is the usage of "superstar" or other high-performing vocabulary when applied to people of different demographics, as typically people associate these terms with men rather than women, so a comparison could be made if the partner pool was expanded to include both genders (Storage et al., 2016.

A related study design could have provided feedback about the participant's performance between rounds as a way to allow for a deeper analysis of the impact of task performance. Whether the information they were told was accurate or manipulated, it would provide insight into task performance's impact on how a high performing partner is perceived.

There are also other factors that could mediate the relationship between race, gender, task performance, like the age, education level, and socioeconomic status of the participants. Additional research could be conducted on the different sub-attributes of the participants to see if their attitudes shift based on non-visible factors. This research could provide insight into how other demographic elements interact with race and gender to affect how high performers are evaluated.

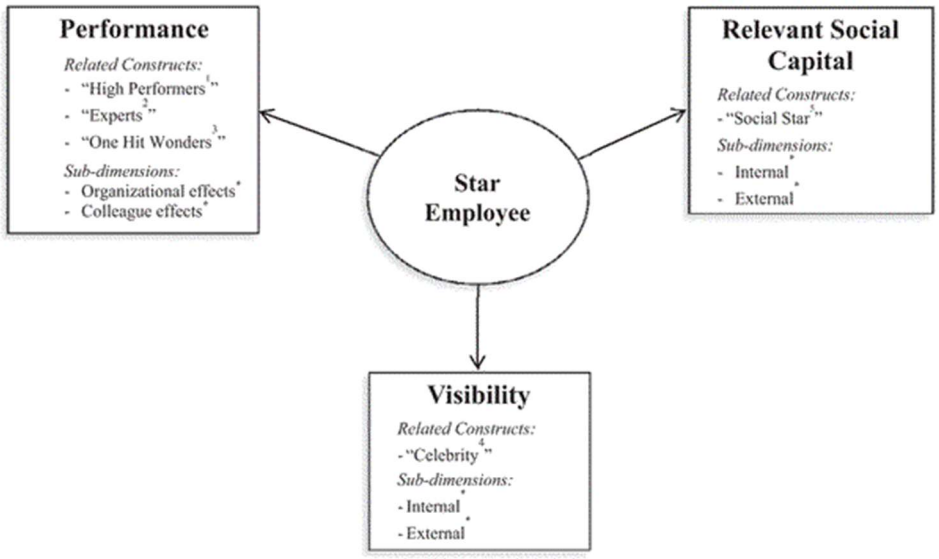
Conclusion

The questions inherent to the superstar and diversity and inclusion literature are numerous, based on different elements like the race, gender, and task performance of people and how these elements affect how a high-performing female partner is evaluated. The volume of potential demographics that can be affect these relationships are large in number, leading to many possible interaction effects that can explain how high performing females are evaluated. This study, however, provides a valuable insight into how demographic characteristics and task performance of people can affect how they perceive their high performing female colleagues for an emergent task. Much research has been done to quantity the barriers that women and people of color face at work, but there is still much work to be done, particularly for the domain of high performers. In regards to this, this study provides valuable insights about how these different factors intersect for people and partners from different backgrounds. These insights are valuable to understand the specific challenges faced by high performing females of different races. Workplaces in the US are continuing to prioritize diversity and inclusion, and beginning to better integrate data-driven practices for identifying and supporting high performers. As this trend continues, the challenge of understanding how high performers, particularly ones with marginalized identities, are evaluated differently is of increasing importance.

Appendix

Appendix A

Call's Definitional Dimensions of Star Employees Framework (2015)



Call, M. L., Nyberg, A. J., & Thatcher, S. (2015). Stargazing: An integrative conceptual review, theoretical reconciliation, and extension for star employee research. Journal of Applied Psychology, 100(3), 623.

Appendix B

Informed Consent Question:

Thank you for your interest in participating in this research study! Your responses will help University of Michigan scholars continue their important contributions to management education and research.

To participate, please review the following information and indicate your consent by clicking the "I consent to this study" button below.

Study Objective: This study is intended to help understand how we respond to someone who is about to teach us something. Results will be used for academic purposes (e.g., publication in academic journals, conference presentations, teaching, etc.).

Participation Details: You will be asked to fill out a survey. The survey will take approximately 10-15 minutes to complete.

Data Collection: Surveys will be distributed through Qualtrics, the online survey program approved by the Ross School of Business at the University of Michigan – Ann Arbor. The website is behind a firewall, and data will only be accessible by members of the University of Michigan research team who must provide password and user ID. Following the study, data will be downloaded and all individual identifiers will be removed.

Benefits from Results: Results will be used to help understand what influences employees' evaluations of their peers in the workplace.

Payment: Participants who successfully complete the study will receive payment. Successful completion is determined at the sole discretion of the study author(s) using common methods to identify non-genuine responses. Examples of non-genuine responses include nonsense answers, responses completed in an extremely short or long amount of time, failure to respond to instructions provided in the survey, and/or otherwise clearly failing to offer genuine responses. Participants should complete the study in one sitting without interruptions to help ensure their response is not considered as non-genuine due to time length. Only submissions considered genuine will receive payment. There is an eligibility screener on the beginning page. Duplicate responses will not receive payment.

Risks: We do not anticipate more than a minimal level of risk to participating in this survey. Some individuals may find a few questions to be difficult or sensitive in nature; however, you are welcome to not answer these questions or to terminate the survey if you wish. You will not know everything about the study until after your participation.

Confidentiality: Beyond the study team, no one will be made aware of your decision to participate in this research (whether you complete the questionnaire), nor of your responses. Your responses will be kept completely confidential by the University of Michigan study team, and will only be shared after being averaged with responses from others (for example, "43% of participants said"). Your individual responses will NOT be made available to anyone.

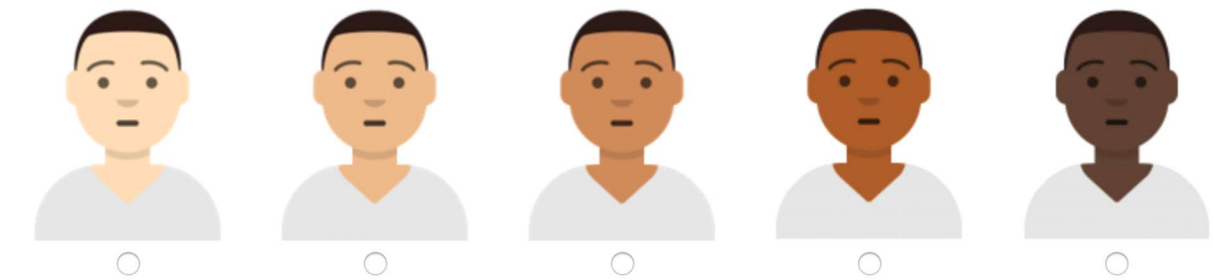
Voluntary nature of participation: Your participation in this project is voluntary, and if you wish, you do not have to participate. Some questions may be sensitive and you may decide to leave the study at any time by simply closing the Qualtrics website.

Contact information: If you have questions about your rights as a research participant, or wish to obtain information, ask questions or discuss any concerns about this study with someone other than the researcher(s), please contact the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board, 2800 Plymouth Rd. Building 520, Room 1169, Ann Arbor, MI 48109-2800, (734) 936-0933, or toll free, (866) 936-0933, irbhsbs@umich.edu. HUM00177613. IMPORTANT: THERE IS NO PENALTY TO YOU OR TO ANYONE ELSE IF YOU DECIDE NOT TO PARTICIPATE OR IF YOU CHOOSE TO END THE SURVEY EARLY.

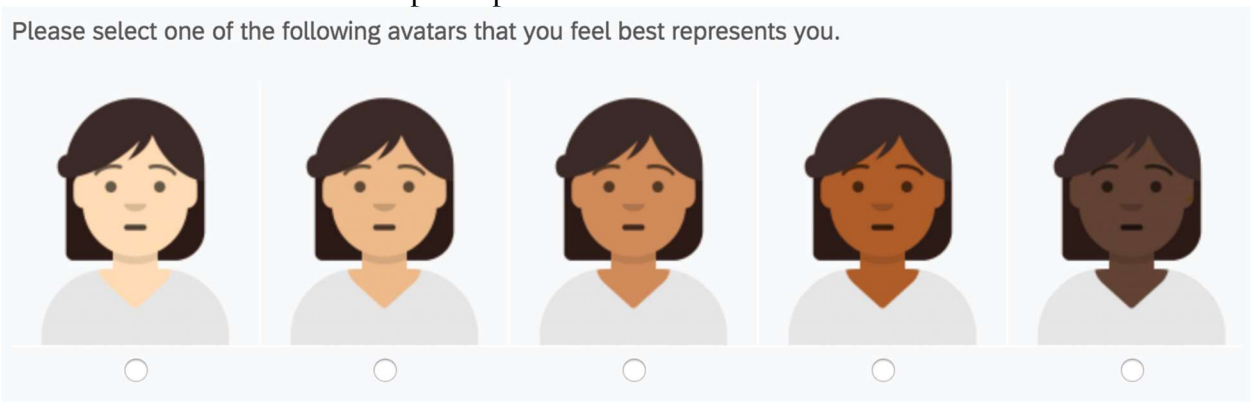
Appendix C

Avatar Options for Participants

For participants who identified as male:



For participants who identified as female:



Appendix D

Likert scale questions for judgments of partner:

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
My partner exhibited warmth	○	○	○	○	○	○	○
My partner was cooperative	○	○	○	○	○	○	○
My partner displayed leadership	○	○	○	○	○	○	○
My partner was likable	○	○	○	○	○	○	○

Appendix E

Likert scale questions for attribution of partner's ability:

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
My partner's performance was the result of luck	○	○	○	○	○	○	○
My partner's performance was the result of skill	○	○	○	○	○	○	○
My partner's performance was the result of applied effort	○	○	○	○	○	○	○
My partner's performance was the result of inherent ability	○	○	○	○	○	○	○

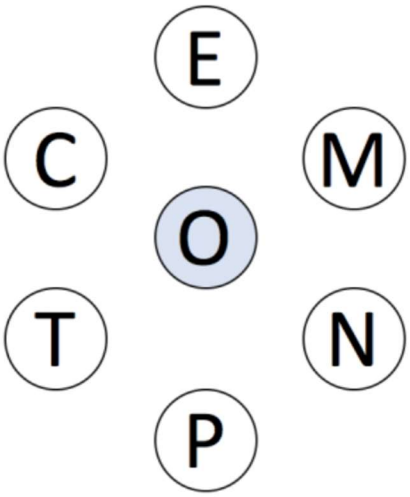
Appendix F

Likert scale questions for attribution of partner's performance:

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I would recommend my partner as a colleague	○	○	○	○	○	○	○
I feel I was similar to my partner	○	○	○	○	○	○	○
I believe my partner was better than me at the games I played	○	○	○	○	○	○	○
My partner would perform similarly well on other games	○	○	○	○	○	○	○

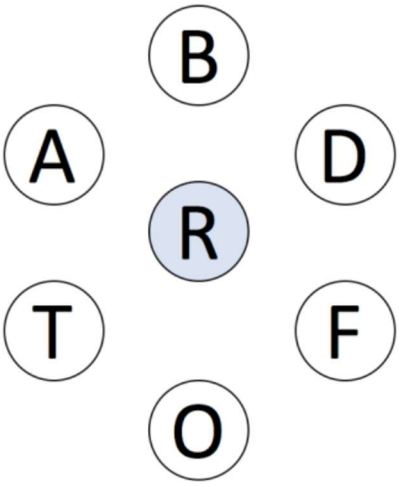
Appendix G

Superadditive Task for Round 1



Appendix H

Superadditive Task for Round 2



Appendix I

Emergent Game Task for Round 1

A delivery van is set to deliver soda to seven grocery stores—A, B, C, D, E, F, and G—over the course of seven days. Exactly one delivery is made each day and soda can't be delivered to any store twice:

-The delivery B comes after the delivery to C.
-The delivery to F comes after the delivery to D.
-D's delivery comes before A's delivery, and there are 4 deliveries between them.
-Delivery to C is either on the first or third day.
-Delivery to E occurs directly before B's delivery
-Delivery to G must occur on the last day

Adapted From "An Intro To Basic Linear (Ordering) Games - Free LSAT Logic Games Lesson" by Evan Jones, 2013 (https://lawschooli.com/intro-basic-linear-ordering-games/). Adapted with permission.

Appendix J

Emergent Game Task for Round 2

An advertising executive must schedule the advertising during a particular television show. Seven different consecutive time slots are available for advertisements during a commercial break, and are numbered one through seven in the order that they will be aired. Seven different advertisements A, B, C, D, E, F, and G must be aired during the show. Only one advertisement can occupy each time slot. The assignment of the advertisements to the slots is subject to the following restrictions:

-A must be aired directly before C.
-B must be in the first slot and D cannot be directly after B.
-C must be aired during an earlier time slot than G.
-E must be aired during an earlier time slot than A.
-If E does not occupy the fourth time slot, then D must occupy the fourth time slot.
-F must be aired after E and before D.

Adapted From "Sample Logic Game" by Griffon Prep, n.d. (https://www.griffonprep.com/logicgame.html). Adapted with permission.

Appendix K

White Avatar

Avataaars Generator. (n.d.). Retrieved from https://getavataaars.com/

Appendix L

Black Avatar

Avataaars Generator. (n.d.). Retrieved from https://getavataaars.com/

Appendix M

5 White names selected from the Top 20: Molly, Emily, Amy, Heather, Katherine

ABC News. (2015, March 1). Retrieved from https://abcnews.go.com/2020/top-20-whitest-blackest-names/story?id=2470131

Appendix N

5 Black names selected from the Top 20: Nia, Jada, Jasmine, Aliyah, Tiara

ABC News. (2015, March 1). Retrieved from https://abcnews.go.com/2020/top-20-whitest-blackest-names/story?id=2470131

Appendix O

3 clues for Superadditive task:

"Keep track of wrong orders you have already tried to avoid repeating work"
"If one slot has a choice of two elements, build out possible orders for each different condition"
"Combine rules to get information about the order of multiple elements"

Appendix P

3 clues for Emergent task:

"Think of a word then add a letter to it (even for three letter words)"
"Remember to repeat letters, including the center letter"
"Find sets of letters that go together (i.e. 'sh' 'ch') and double letters"

References

Adler, M. (1985). Stardom and talent. The American economic review, 75(1), 208-212.

Agocs, C., & Jain, H. C. (2001). Systemic racism in employment in Canada: Diagnosing systemic racism in organizational culture. Toronto: Canadian Race Relations Foundation.

Agrawal, A. K., McHale, J., & Oettl, A. (2014). Why stars matter (No. w20012). National Bureau of Economic Research.

Aguinis, H., O'Boyle Jr, E., Gonzalez-Mulé, E., & Joo, H. (2016). Cumulative advantage: Conductors and insulators of heavy-tailed productivity distributions and productivity stars. Personnel Psychology, 69(1), 3-66.

Aguinis, H., & O'Boyle Jr, E. (2014). Star performers in twenty-first century organizations. Personnel Psychology, 67(2), 313-350.

Azoulay, P., Fons-Rosen, C., & Graff Zivin, J. S. (2019). Does science advance one funeral at a time? American Economic Review, 109(8), 2889-2920.

Azoulay, P., Graff Zivin, J. S., & Wang, J. (2010). Superstar extinction. The Quarterly Journal of Economics, 125(2), 549-589.

Avataaars Generator. (n.d.). Retrieved from https://getavataaars.com/

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Brown, J. (2011). Quitters never win: The (adverse) incentive effects of competing with superstars. Journal of Political Economy, 119(5), 982-1013.

Burke, M. A., Fournier, G. M., & Prasad, K. (2007). The diffusion of a medical innovation: is success in the stars? Southern Economic Journal, 588-603.

Call, M. L., Nyberg, A. J., & Thatcher, S. (2015). Stargazing: An integrative conceptual review, theoretical reconciliation, and extension for star employee research. Journal of Applied Psychology, 100(3), 623.

Chen, J. S., & Garg, P. (2018). Dancing with the stars: Benefits of a star employee's temporary absence for organizational performance. Strategic Management Journal, 39(5), 1239-1267.

Chmait, N., Robertson, S., Westerbeek, H., Eime, R., Sellitto, C., & Reid, M. (2019). Tennis superstars: The relationship between star status and demand for tickets. Sport Management Review.

Crocker, J., & Luhtanen, R. (1990). Collective self-esteem and ingroup bias. Journal of personality and social psychology, 58(1), 60.

Davies, J. (n.d.). Word Cloud Generator. Retrieved from https://www.jasondavies.com/wordcloud/

Deitch, E. A., Barsky, A., Butz, R. M., Chan, S., Brief, A. P., & Bradley, J. C. (2003). Subtle yet significant: The existence and impact of everyday racial discrimination in the workplace. Human Relations, 56(11), 1299-1324.

Denrell, J. (2005). Should we be impressed with high performance?. Journal of Management Inquiry, 14(3), 292-298.

Denrell, J., & Liu, C. (2012). Top performers are not the most impressive when extreme performance indicates unreliability. Proceedings of the National Academy of Sciences, 109(24), 9331-9336.

Epstein, C. F. (1973). BLACK AND FEMALE-DOUBLE WHAMMY. *Psychology Today*, *7*(3), 57.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. Annual review of psychology, 47(1), 273-305.

Feather, N. T. (1989). Attitudes towards the high achiever: The fall of the tall poppy. Australian Journal of Psychology, 41(3), 239-267.

Franck, E., & Nüesch, S. (2012). Talent and/or popularity: what does it take to be a superstar? Economic Inquiry, 50(1), 202-216.

Gallardo-Gallardo, E., Dries, N., & González-Cruz, T. F. (2013). What is the meaning of 'talent' in the world of work? Human Resource Management Review, 23(4), 290-300.

Galperin, R. V., Hahl, O., Sterling, A. D., & Guo, J. (2019). Too good to hire? Capability and inferences about commitment in labor markets. Administrative Science Quarterly, 0001839219840022.

Gergaud, O. Verard V. (2006). Untalented but Successful [PowerPoint slides]. Retrieved from https://www.slideserve.com/kishi/untalented-but-successful

Greer, L. L., & Chu, C. (2019). Power Struggles: When and Why the Benefits of Power for Individuals Paradoxically Harm Groups. Current opinion in psychology.

Grissom, J. A., & Redding, C. (2015). Discretion and disproportionality: Explaining the underrepresentation of high-achieving students of color in gifted programs. Aera Open, 2(1), 2332858415622175.

Grossman, S. J., & Hart, O. D. (1992). An analysis of the principal-agent problem. In Foundations of Insurance Economics (pp. 302-340). Springer, Dordrecht.

Groysberg, B. (2010). Chasing stars: The myth of talent and the portability of performance. Princeton University Press.

Groysberg, B. (2008). How star women build portable skills. Harvard Business Review, 86(2), 74.

Groysberg, B., & Lee, L. E. (2009). Hiring stars and their colleagues: Exploration and exploitation in professional service firms. Organization science, 20(4), 740-758.

Groysberg, B., Nanda, A., & Nohria, N. (2004). The risky business of hiring stars. Harvard business review, 82(5), 92-101.

Groysberg, B., Polzer, J. T., & Elfenbein, H. A. (2011). Too many cooks spoil the broth: How high-status individuals decrease group effectiveness. Organization Science, 22(3), 722-737.

Hahl, O., & Zuckerman, E. W. (2014). The denigration of heroes? How the status attainment process shapes attributions of considerateness and authenticity. American Journal of Sociology, 120(2), 504-554.

Hausman, J. A., & Leonard, G. K. (1997). Superstars in the National Basketball Association: Economic value and policy. Journal of Labor Economics, 15(4), 586-624.

Hill, B. (2014). The superstar effect in 100-meter tournaments. *International Journal of Sport Finance*, *9*(2), 111.

Hong, L., Page, S. E., & Riolo, M. (2012). Incentives, information, and emergent collective accuracy. Managerial and Decision Economics, 33(5-6), 323-334.

Jane, W. J., Yao, J. L., & Wang, J. S. (2018). Having Good Friends is a Good Thing: The Effects of Peers and Superstars on Performance in Swimming Competitions. International Journal of Economic Sciences, 7(1), 39-64.

Jaxon, J., Lei, R. F., Shachnai, R., Chestnut, E. K., & Cimpian, A. (2019). The acquisition of gender stereotypes about intellectual ability: Intersections with race. Journal of Social Issues.

Jensen, J. M., Patel, P. C., & Raver, J. L. (2014). Is it better to be average? High and low performance as predictors of employee victimization. Journal of Applied Psychology, 99(2), 296.

Kehoe, R. R., Lepak, D. P., & Bentley, F. S. (2018). Let's call a star a star: Task performance, external status, and exceptional contributors in organizations. Journal of Management, 44(5), 1848-1872.

Kehoe, R. R., & Tzabbar, D. (2015). Lighting the way or stealing the shine? An examination of the duality in star scientists' effects on firm innovative performance. Strategic Management Journal, 36(5), 709-727.

Kim, E., & Glomb, T. M. (2014). Victimization of high performers: The roles of envy and work group identification. Journal of Applied Psychology, 99(4), 619.

Kraft-Todd, G. T., Reinero, D. A., Kelley, J. M., Heberlein, A. S., Baer, L., & Riess, H. (2017). Empathic nonverbal behavior increases ratings of both warmth and competence in a medical context. PloS one, 12(5), e0177758.

Kreuger, A. (2019, June 1). The Economics of Rihanna's Superstardom. Retrieved from https://www.nytimes.com/2019/06/01/opinion/sunday/music-economics-alan-krueger.html.

Lockwood, P., & Kunda, Z. (1997). Superstars and me: Predicting the impact of role models on the self. Journal of personality and social psychology, 73(1), 91.

Marr, J. C., & Thau, S. (2014). Falling from great (and not-so-great) heights: How initial status position influences performance after status loss. Academy of Management Journal, 57(1), 223-248.

Marshall, P. D. (2014). Celebrity and power: Fame in contemporary culture. U of Minnesota Press.

Marshall, A. (2009). Principles of economics: unabridged eighth edition. Cosimo, Inc.

Mills, L. S., Soulé, M. E., & Doak, D. F. (1993). The keystone-species concept in ecology and conservation. BioScience, 43(4), 219-224.

Nihalani, P. K., Wilson, H. E., Thomas, G., & Robinson, D. H. (2010). What determines high- and low-performing groups? The superstar effect. Journal of Advanced Academics, 21(3), 500-529.

Nyberg, A. (2010). Retaining your high performers: Moderators of the performance–job satisfaction–voluntary turnover relationship. Journal of applied psychology, 95(3), 440.

O'Boyle Jr, E., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. Personnel Psychology, 65(1), 79-119.

Oettl, A. (2012). Reconceptualizing stars: Scientist helpfulness and peer performance. Management Science, 58(6), 1122-1140.

Oldroyd, J. B., & Morris, S. S. (2012). Catching falling stars: A human resource response to social capital's detrimental effect of information overload on star employees. Academy of Management Review, 37(3), 396-418.

Page, S. E. (2007). Making the difference: Applying a logic of diversity. Academy of Management Perspectives, 21(4), 6-20.

Pettit, N. C., Sivanathan, N., Gladstone, E., & Marr, J. C. (2013). Rising stars and sinking ships: Consequences of status momentum. Psychological science, 24(8), 1579-1584.

Pierce, S., Hodge, K., Taylor, M., & Button, A. (2017). Tall poppy syndrome: Perceptions and experiences of elite New Zealand athletes. International Journal of Sport and Exercise Psychology, 15(4), 351-369.

Pollard, D. A. (1999). Unconscious Bias and Self-Critical Analysis: The Case for a Qualified Evidentiary Equal Employment Opportunity Privilege. Wash. L. Rev., 74, 913.

Rivera, L. A., & Tilcsik, A. (2019). Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. American Sociological Review, 84(2), 248-274.

Rosen, S. (1981). The economics of superstars. The American economic review, 71(5), 845-858.

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.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. science, 311(5762), 854-856.

Scott, K. A., & Brown, D. J. (2006). Female first, leader second? Gender bias in the encoding of leadership behavior. *Organizational behavior and human decision processes*, *101*(2), 230-242.

Storage, D., Horne, Z., Cimpian, A., & Leslie, S. J. (2016). The frequency of "brilliant" and "genius" in teaching evaluations predicts the representation of women and African Americans across fields. PloS one, 11(3), e0150194.

Swaab, R. I., Schaerer, M., Anicich, E. M., Ronay, R., & Galinsky, A. D. (2014). The too-much-talent effect: Team interdependence determines when more talent is too much or not enough. Psychological Science, 25(8), 1581-1591.

Tanaka, R., & Ishino, K. (2012). Testing the incentive effects in tournaments with a superstar. Journal of the Japanese and International Economies, 26(3), 393-404.

Ezersky, S., (2019). Retrieved from https://www.nytimes.com/puzzles/spelling-bee

Van Laer, K., & Janssens, M. (2011). Ethnic minority professionals' experiences with subtle discrimination in the workplace. Human Relations, 64(9), 1203-1227.

Williams, J. C., & Multhaup, M. (2018). For women and minorities to get ahead, managers must assign work fairly. Harvard Business Review, 2-9.