Facial Biases in the Evaluation of Baseball Pitchers

DARTMOUTH

Matthew R. Schnell

Honors Thesis

Professor Michael C. Herron, Advisor

Program in Quantitative Social Science

Dartmouth College

June 1, 2022

Acknowledgements

Ever since I read Michael Lewis' *Moneyball* when I was 11, I knew I wanted to work in baseball. The book gave many concrete examples of baseball scouts being biased by factors about players that were completely independent from their playing skills. I found it amazing that it took so many decades for teams to start employing analytics and technology to improve their processes about both playing strategy and, more importantly, finding players that other teams valued less. Even from such a young age, I knew I wanted to help baseball organizations continue enhancing their data-driven capabilities.

The idea for this thesis was sparked the summer before my senior year, when I re-read *Moneyball* for the third time. Despite the surge in analytics that teams employ to scout players, I still consistently noticed scouting reports, particularly from broadcasters and analysts on television and publications, of players referencing parts of players bodies that held no true value in defining and projecting players' abilities. After talking with a few college baseball coaches, I became even more intrigued by this idea that evaluators were choosing players based on factors such as their "baseball intangibles" (whatever that means) or whether they had facial hair or not. I hope my thesis can supplement Lewis' work by providing empirical research on the notion that baseball scouts are incorrectly using appearance-based heuristics to evaluate players.

My thesis has been the culmination of my time at Dartmouth, and I am so grateful for everyone who has helped me during my Dartmouth journey. It has been such a joy to use everything I have learned from Dartmouth to work on a subject that I am so passionate about. I have been pushed to find new ways to solve problems, communicate clearly and concisely to explain complex ideas in simple ways, improve my analytical skills, work hard, and persevere through challenges. Never before have I worked on something over such an extended period that I was truly interested in. I am confident that I will continue to use all of the skills I have developed during these past nine months of working on my thesis for the rest of my life.

This project would not have been made possible without the help and support of my professors, family, and friends. First, I want to show my appreciation to my thesis advisor, Professor Michael Herron. Over winter break during my freshman year, Professor Herron kindly helped me work on a project analyzing the effects of kickoffs in college football. He responded to my emails so quickly,

even though I am certain he had a million other things to do. Since then, I knew I wanted to use Professor Herron as a resource and mentor for the rest of my Dartmouth experience. I am proud to say that I might be one of the few students at Dartmouth to take every class offered by Professor Herron. In all of his classes, Professor Herron challenged me to think critically and creatively to find solutions to problems. Throughout my time at Dartmouth, Professor Herron has readily been available to offer his thoughts on sports analytics, advice on career paths, and ideas for research projects. Professor Herron has continuously supported me, particularly in my endeavors to write this thesis. I am so grateful for Professor Herron's willingness to meet with me whenever I faced any troubles, answer every one of my questions patiently, and encourage me in times of doubt. Without Professor Herron, there is absolutely no way I would have completed this project.

I would like to thank a few additional professors for helping with my thesis. Professor Charles Crabtree, all the way in California, for offering tremendous feedback for my experiment designs. Professor Crabtree was always willing to answer my many emails and meet over Zoom. And Professor Brendan Nyhan and Professor John Carey, for providing helpful insight as additional readers — both were so supportive during my thesis defense, pushing me to go deeper in my research. Their feedback after my defense was so incredibly helpful in improving my thesis.

I owe a huge sense of gratitude to all the baseball coaches who offered to talk with me about my project. Specifically, I want to acknowledge Coach Robert Whalen, Coach Bryan Adamski, and Rob Friedman, for helping me understand how baseball scouts evaluate players and offering their feedback on the videos of my pitchers. A huge thanks goes out to all of the baseball coaches who completed my survey. Without their help, my thesis would be incredibly different, and might not have even existed in the end.

I would like to thank Undergraduate Research for providing me with the Kaminsky Fund Award — writing my thesis would not have been possible without this grant. I would also like to express a warm gratitude to Rachel Bibeault from Undergraduate Research, who, for months, readily answered the dozens of questions I sent her way to help me get my surveys into the field as soon as possible.

A huge acknowledgement to my peers also writing theses. Their help, feedback, and laughter helped immensely over these past tiring months. I could always count on them to provide support

during the challenging times.

Lastly, I want to thank my family. My brothers Zach and Jacob offered tremendous love and support throughout this project, always being there to listen to me and offer to help in any way possible. Mom and Dad, for instilling in me the value of determination, curiosity, and hard work since my early childhood. None of this would have been made possible without the love and encouragement from my family.

Abstract

Individuals consistently judge others based on their faces, potentially biasing their decisions about others. While many studies reveal how people are subject to facial biases, there has been no empirical research on how, if at all, a baseball player's face biases baseball evaluators. I filmed nine different baseball pitchers throwing and photographed head shots of their faces, and then employed three survey experiments to understand how coaches' evaluations are influenced by features of these players' faces. I assessed which pitchers' faces were most associated with a set of four attributes. I added to the results from the first experiment by studying how attractive each pitcher's face was. Employing a principal component analysis, I determined that the first principal component explained roughly half of the variance in how the pitchers' faces looked among four attributes, and players were ranked on this principal component. Lastly, I analyzed how college baseball coaches evaluated four players taken from the first experiment when they saw their faces in videos of the four pitchers throwing compared to when they could not see their faces in the videos. Using an ordinary least squares regression, I found that, for two of the four players evaluated, baseball coaches rated the players significantly differently when they could see their faces in the pitching videos compared to when their faces were blurred out. My findings suggest that pitchers' faces can bias baseball scouts' evaluations of pitchers.

Contents

1 Introduction

People consistently judge their peers simply by looking at their faces. Individuals make consequential decisions based on their perceptions of faces, even though these perceptions are usually wrong (Jaeger et al., 2019a; Olivola, Funk, et al., 2014). In business settings, for example, companies are more likely to hire CEOs whose faces appear competent, dominant, and successful, despite those CEOs sometimes being less qualified than other candidates (Rule & Ambady, 2008). Baseball evaluators, particularly baseball coaches and scouts, have also been shown to suffer from visual biases: scouts have a tendency to over-focus on a prospect's appearance when making their player evaluations (Bellemore, 2001; Lewis, 2003; Smith & Harrison, 1996).

When baseball scouts evaluate a player, they combine objective information, such as a pitcher's statistics, with subjective information that scouts may disagree on, such as whether a pitcher's throwing motion is high quality or not (Streib et al., 2012). Scouts typically supplement a player's metrics, such as a pitcher's velocity and spin rate of his fastball, by watching the player during warm-ups, live games, practice, and on film, ultimately trying to determine how likely it is for that player to be successful (Streib et al., 2012). In addition to having their scouts watch players live, teams use an increasing amount of metrics on players to aid their decision making and track the thousands of players they are monitoring (Eggers, 2012). Organizations are hiring younger, quantitative minds in their scouting departments to analyze millions of data points regarding players' statistics (Apstein, 2019). While baseball teams are using more and more analytical methods to evaluate players, clubs still rely heavily on their scouts to watch and offer their opinions on players (Lindbergh & Arthur, 2019; Streib et al., 2012). Scouts, who historically have suffered from appearance biases, are susceptible to implicit biases because they watch players live, notably judging players based on their appearances (Lewis, 2003; Lindbergh & Arthur, 2019).

When forecasting a player's baseball abilities and potential, statistical validation is usually much more reliable for a baseball scout than using intuition (Lewis, 2003). Despite this, scouts frequently rely on a player's appearance to judge players' talent. Listening to baseball evaluators speak about certain players and player evaluation methods, I learned that scouts explicitly account for a player's appearance when evaluating his talent. One scout reported that he would not sign a

player, "Unless I could look in his face and see what I wanted to see: drive, determination, maturity, whatever" (Gines, 2017). Scouts have also commented on how players dress themselves. One former evaluator wrote about a player, "Even with ear rings and high tops, he can swing the bat" (Lindbergh & Arthur, 2019). A scout from one Major League Baseball organization told me that his team's scouts were instructed to look at a prospect's facial hair to understand how much more room the player had to develop into their bodies and continue growing.[1] Through conversation with another college baseball coach, I learned that this coach prefers recruiting players who appear more tan or sunburned, implying that these players spent more time outside working on their baseball abilities and have a higher work ethic.[2]

That baseball evaluators judge players by their faces and appearances is not surprising. Individuals routinely face uncertainty when making decisions, and people rely on a number of heuristics to reduce the complexity of evaluating others (Tversky & Kahneman, 1974). In his book *Moneyball*, Michael Lewis described several of these sight-based heuristics that scouts use to help them evaluate baseball prospects, such as a player's weight. Lewis writes about one player in particular who, despite his hitting prowess and ability to frequently get on base, scouts wanted to rule out as a viable prospect solely because of his excessive weight, seemingly misshapen body, and "unnatural" appearance (Lewis, 2003). In this example, the baseball scouts use this player's weight as a sight-based heuristic to help them evaluate the player. That is, the scouts used the player's body appearance instead of other key information about the player, such as the number of walks the player had in his prior season.

While baseball evaluators believe that a player's face holds salient information, there is no research suggesting that the face provides meaningful information about a player. One conversation with a baseball scout revealed that many scouts are instructed to rely on players' faces when making evaluations.[3] For example, evaluators sometimes focus on a pitcher's facial expression to understand how much effort they are using when pitching, suggesting that a pitcher whose facial expression looks like he is throwing too hard might be more prone to injury. However, no empirical evidence exists that would suggest that focusing on a pitcher's facial expression while throwing has any significance

[1]The MLB is the top professional baseball league in the United States. Players from all across the globe aim to take their talents to this league. The conversation with this MLB scout occurred on April 20, 2022

[2]This conversation occurred on April 26, 2022

[3]This conversation occurred on April 20, 2022

as to how much effort they are exerting when pitching.

This study addresses whether baseball scouts' evaluations of pitchers are biased by a player's face. While baseball scouts have been recognized as being biased by a player's appearance and race, I have not found any literature explicitly researching the impact a face, in particular, has on a baseball evaluator's judgement of a player. This research aims to understand the extent to which scouts use players' faces as heuristics to evaluate players and how the face affects scouts' evaluations of players.

2 Literature

There are four bodies of literature that have informed my study. In what follows, I first discuss implicit facial biases in different areas. Second, I write about the non-facial biases that exist in sports evaluation. Third, I review specific facial biases in sports. I conclude this section by focusing on how existing literature examines facial biases in baseball.

2.1 Facial biases

Individuals frequently use faces to identify a person's character and intentions (Jaeger et al., 2019a; Olivola, Funk, et al., 2014; Todorov et al., 2015). The first impression individuals make of their peers depends on how a person perceives another person's facial appearance. A person's face is a rich source of information about a person, such as their age, race, and ethnicity (Olivola, Funk, et al., 2014). Individuals often use a person's face to infer that person's character and abilities (Todorov et al., 2015). For instance, people with angry-looking faces are perceived by others as having a dominant and aggressive personality (Montepare & Dobish, 2003; Oosterhof & Todorov, 2008; Todorov & Duchaine, 2008), while those with kinder-looking faces are described as more trustworthy (Krumhuber et al., 2007). Furthermore, adults with faces that appear more baby-faced are perceived as having more childish traits, such as being physically weak, naive, submissive, kind, and warm (Berry & McArthur, 1986; McArthur & Apatow, 1984; Montepare & Zebrowitz, 1998; Zebrowitz & Montepare, 1992). This is an issue because studies show that facial appearances are a very weak indicator of a person's true personality and underlying traits (Olivola, Funk, et al., 2014;

Todorov et al., 2015). Moreover, people are poor at evaluating social characteristics from faces; facial appearances bias human judgements, and there is little value in evaluating a person based on their face (Olivola, Funk, et al., 2014; Olivola & Todorov, 2010; Todorov et al., 2015).

Numerous studies have shown that individuals possessing specific facial features are more likely to receive favorable outcomes and avoid unfavorable outcomes (Olivola, Funk, et al., 2014). Employers are much more likely to hire applicants with attractive faces (Heilman E. & Saruwatari R., 1978). Workers with more attractive faces also earn greater salaries compared to those with less attractive faces (Frieze et al., 1991). Research shows that people select their leaders based on how they perceive the faces of their candidates. In politics, one study concluded that voters elect their candidates based on how "politician-like," sociable, and threatening their faces appear (Olivola, Eubanks, et al., 2014). Within the military setting, officers whose faces are viewed as more dominant-looking are promoted much faster and to higher ranks than those whose faces are perceived as weaker (Olivola, Eubanks, et al., 2014).

Existing literature also underlines people's tendency to spontaneously judge individuals trustworthiness based on their facial appearances (Jaeger et al., 2019a; Olivola, Funk, et al., 2014; Rezlescu et al., 2012; Todorov et al., 2015). Given two individuals who are identical in every way except their facial features, the person with the "right" face is perceived as more trustworthy (Oosterhof & Todorov, 2008). Several studies have examined how having a trustworthy or untrustworthy face affects individuals in real-world situations. Regarding criminal trials, certain people have a higher probability of being selected from a police lineup if their face appears untrustworthy, while those with trustworthy faces are more likely to avoid selection (Flowe & Humphries, 2011). Further, those on trial with faces that either fit the stereotype of the description of the crime or have untrustworthy faces are much more likely to be convicted by the jury (Korva et al., 2013). In the financial realm, research that analyzed peer-to-peer lending websites found that people whose faces seem more trustworthy earn more investments and loans from other people (Duarte et al., 2012). In more controlled experiments, studies used computer software to manipulate a face to make it appear more or less trustworthy, determining that untrustworthy faces attracted less investments and rewards from participants than the trustworthy faces did (Rezlescu et al., 2012; Van't Wout & Sanfey, 2008).

Facial biases still persist even when people have more accurate background information about

individuals (Todorov et al., 2015). When valid cues about a person's character traits are available, people choose to ignore this information and rely on facial appearances to make consequential decisions (Jaeger et al., 2019b). An additional study determined that despite survey participants obtaining prior background information about people, the respondents still chose to invest more money in the individuals with more trustworthy-appearing faces (Rezlescu et al., 2012). People decide to ignore diagnostic facts about people in favor of their facial appearances, in spite of a face having poor predictive power about a person's character. Introducing facial cues actually causes people to discount prior information even more (Jaeger et al., 2019b).

2.2 Non-facial biases in sports

When evaluators judge the performance of athletes, they are potentially biased by many different factors. In the Olympic Games, judges are greatly affected by sequencing biases: an athlete's performance score rose if they performed after an athlete who was scored poorly. Similarly, if an athlete followed another athlete who scored well, then they were more likely to receive a lower score (Kramer, 2017). Additionally, Olympic gymnastic judges give higher scores to participants from their own countries (Ansorge & Scheer, 1988) and to participants who try more challenging routines yet are unsuccessful in their attempts (Morgan & Rotthoff, 2014). Voters for the Fifa Ballon D'or, the award for the best soccer player in the world, suffer from the same biases – jury members disproportionately vote for players from their own country and on teams who play in their own country's league (Coupe et al., 2018). Similarly, judges evaluating female figure skaters voted more favorably for athletes from their own countries (Seltzer & Glass, 1991) who were more well-known than they did for athletes whose names were not recognized by the judges, suggesting a reputation bias (Findlay & Ste-Marie, 2004).

Studies show that teams drafting prospects in professional sports leagues are biased by factors that are not predictive of a player's future abilities. In the National Hockey League, scouts are biased in favor of both European players and taller players, despite the fact that nationality and height have not been shown to be an indicator of future player success (Voyer & Wright, 1998). Within the National Basketball Association, one study determined that given two players with identical statistics, basketball talent evaluators are biased towards selecting the player with a higher

high-school recruiter ranking, despite this ranking having no correlation with future player success (Sailofsky, 2018).

2.3 Facial biases in sports

Researchers have shown that facial biases regarding the quarterback position in the National Football League (NFL) exist. Quarterback is widely viewed as the most important position in football. This position runs and communicates a team's offensive strategy (Williams et al., 2010). Thus, teams typically market their quarterbacks as the face of their organizations. Since teams depend so greatly on their quarterbacks not only for on-field success but also for marketing purposes, NFL executives have been shown as more biased in favor of attractive quarterbacks. Even controlling for player performance, quarterbacks with more attractive faces earn more money than those with less attractive faces (Berri et al., 2011). That said, one study found that a quarterback's facial attractiveness might signal that player's athleticism. Researchers found a positive correlation between a quarterback's on-field success and facial attractiveness (Williams et al., 2010).

The majority of existing research about facial appearances in sports is related to racial biases. One study analyzed language used by NFL draft experts, finding that these experts perpetuate stereotypes of black athletes when evaluating black quarterback prospects (Bigler & Jeffries, 2008). Furthermore, commentators for college football games depict players much differently during games based on the player's ethnicity (Billings, 2004). These stereotypes affect how NFL consumers perceive black players versus white players, as fans also perpetuate these stereotypes in their judgements of players (Ferrucci et al., 2013).

2.4 Biases in baseball

Baseball scouts have a strong tendency to rely too heavily on sight-based evaluations of players (Lewis, 2003). Research shows that, historically, baseball scouts suffer from racial biases when evaluating pitchers. For instance, white players earn promotions to MLB much faster (Bellemore, 2001) and play more high-profile positions, such as infield, than similarly qualified minority players (Curtis & Loy, 1978). Scouts also rely on racial stereotypes of players to determine which position they should play (Smith & Harrison, 1996).

Racial biases spill into all facets of baseball. Baseball umpires calling balls and strikes for players are more likely to favor calls for players of their own race (Parsons et al., 2011). Furthermore, race affects how baseball cards for players are priced at their initial release; given two comparable career performances, prices for a black player will decline much more significantly than the prices for a white player (Burge & Zillante, 2017).

Aside from racial biases in the sport, studies show that scouts favor high school baseball prospects, despite college players having a higher probability of displaying success in MLB than high school players (Lewis, 2003; Spurr, 2000). Teams typically overestimate how frequently high school prospects become star players relative to college players (Burger & Walters, 2009).

3 Hypotheses

As I have reviewed above, the majority of literature on facial biases examines the impact of these biases in areas outside of sports. Not withstanding research on the facial attractiveness of NFL quarterbacks, there has been limited discussion about how athletes' faces bias scouts. And in baseball, aside from examining racial and ethnic biases, research on the impact that a player's face has on baseball evaluators exists only minimally. While baseball evaluators being biased by sight-based heuristics has been discussed, notably in Lewis's *Moneyball*, I have not found any empirical studies specifically focusing on how a baseball player's face might bias baseball evaluators. Given the existing literature I described on facial biases, individuals make less accurate decisions because of such facial biases. If baseball evaluators are also biased by players' faces when making evaluations, then they, too, might make less accurate judgements on players' abilities. With this as background, this thesis adds to both the existing literature around facial biases in general and overall biases in baseball by understanding how a baseball pitcher's face biases baseball scouts.

To identify if facial biases exist in the evaluation of baseball players, I compared how baseball evaluators rate pitchers on their throwing abilities when they see their faces in videos of them throwing compared to when they cannot see the pitchers' faces in the videos. I pose and examine two possible hypothesis in my research:

Hypothesis 1: There is no significant difference in the ratings assigned to a pitcher when a baseball coach sees the pitcher's face compared to when the coach does not see his face. This suggests that seeing a pitcher's face has no impact on how a baseball coach evaluates a pitcher.

Hypothesis 2: There is a significant difference in the ratings assigned to a pitcher when a baseball coach sees the pitcher's face compared to when the coach does not see his face in the evaluation. A positive difference between a player's average rating when his face is visible and when it is not visible in the evaluation means that seeing the pitcher's face makes a baseball evaluator more likely to assign the pitcher a higher rating than they would have if they did not see that pitcher's face when evaluating him. However, if this difference is negative, then seeing a pitcher's face makes a baseball evaluator more likely to assign the pitcher a lower rating than they would have if they did not see that pitcher's face when evaluating him.

If Hypothesis 1 is true for a player, then baseball evaluators do not rely on that player's face when judging his talent. In this situation, scouts are not biased by how the player's face looks, and evaluate the player based on other factors that are more relevant to his throwing ability. If scouts really are judging players by their abilities and not subject to facial biases, then Hypothesis 1 would be true for every player examined.

Meanwhile, if Hypothesis 2 is true for a player, then baseball evaluators are affected by that player's face. Thus, a scout would diminish the facial bias by removing the player's face from the video of him throwing. Since there is no research suggesting the face holds important information about a baseball player's ability, and existing literature says facial biases hurt individuals' abilities to make accurate judgements of others, baseball evaluators could improve their scouting by ignoring the face altogether from their evaluations.

4 Research design

I used three survey experiments to identify facial biases in baseball player evaluation. In this section, I describe the design for the three survey experiments as well as the statistical methods used

to analyze my results.

4.1 Experiment 1A: Perceptions of pitchers' faces

I conducted a survey experiment to identify the characteristics associated with different baseball players' faces. Specifically, the survey captured four attributes regarding a baseball pitcher's face: (1) how much they look like a baseball player, (2) smartness, (3) athleticism, and (4) intimidation.

4.1.1 Participants

I recruited 1,377 participants for this survey experiment from Amazon's Mechanical Turk (MTurk). MTurk is widely used in social science research to gather survey participants and, specifically, to ask respondents to evaluate individuals' faces across different attributes (Olivola, Funk, et al., 2014; Paolacci & Chandler, 2014).

4.1.2 Survey design

Nine different baseball pitchers were used as subjects in the survey. I photographed these players in a single location wearing a blue shirt with black sweatpants. The pitchers' clothes were the same for each subject to limit any potential confounding biases when survey participants, described later, rated pitchers' faces across the four attributes. The nine subjects were students at Dartmouth College, who, at a minimum, pitched in high school. Race was held constant at White for all of the nine pitchers.

Survey participants first consented to take the survey and to answer each question honestly and to the best of their ability. Once consenting to the survey, respondents answered a series of demographic questions regarding age, race, and gender.[4] Respondents then faced a series of three attention checks to ensure that they were paying attention to the survey questions. Respondents who failed to pass attention checks provide low-quality data (Aranow et al., 2020). Thus, responses from individuals who did not correctly answer all three checks were not included in my analysis. The final attention check ensured participants were carefully reading the questions posed to them:

[4]The demographics of the survey participants are summarized in Table A.1 in the Appendix.

People are very busy these days and many do not have time to follow what goes on in the news. We are testing whether people read questions. To show that you've read this much, answer both **"extremely interested"** and **"very interested"**.

Participants were given *Extremely interested*, *Very interested*, *Moderately interested*, *Slightly interested*, *Not interested at all* as response choices to the question.

After answering the demographic and attention check questions, respondents saw four randomly selected photographs from the nine different pictures of the players' faces. Figure 1 is a sample head shot of a player.

Figure 1: Sample player head shot

Respondents read four different statements after seeing each pitcher's face:

1. *This individual plays baseball.*
2. *This individual is intimidating.*
3. *This individual is smart.*
4. *This individual is athletic.*

Each individual statement represents a specific attribute: (1) how much they look like a

baseball player, (2) smartness, (3) athleticism, and (4) intimidation. These four attributes were selected due to conversations I had with several college baseball coaches during the course of my thesis.[5] The coaches suggested that they wanted smart, athletic players, who appeared confident and determinant. These coaches also told me that they tried to look for players who had the "baseball intangibles," meaning they play intelligently and know the game of baseball very well. I attempted to capture this idea by asking whether a player looked like a baseball player.

For each statement about an attribute, respondents were asked:

How much do you agree with the following statement?

Respondents selected their answers to this question using a 7-point Likert scale, with values ranging from one *(strongly disagree)* to seven *(strongly agree)*. The seven-point rating scale is used widely in social-science experiments (Menold, 2020).

The order that participants answered each question about how much they agreed or disagreed with the four statements about a player was kept consistent for each player throughout the survey, in the same order as the statements are written above.

Each statement about a specific attribute was shown on its own page in the survey below a head shot of a player. For example, for the first pitcher shown to a respondent, a respondent answered only the question about how much they agreed or disagreed with the statement that the individual plays baseball. Then, they clicked a *Next* button that took them to the question about how much they agreed or disagreed with the second statement. Respondents answered questions in this manner until they answered the question about how much they agreed or disagreed with each of the four statements for the four pitchers they were assigned. This method was done to make it more challenging for respondents to straight-line their answers, which occurs when respondents answer the same response for each question to finish the survey as quickly as possible. If the four statements about a single pitcher were shown on the same page, then respondents would be able to easily respond the same way to the question about how much they agree or disagree with each of the four statements.

[5]Conversations occurred on October 14, 2021 and October 21, 2021.

4.1.3 Methodology

I performed a principal component analysis (PCA) on the survey results. I used this analysis to see if I could reduce the four questions regarding facial attributes into just one or two dimensions for each pitcher. I examined whether there was a single principal component that explained most of the variance in the pitchers' faces. If this were the case, I planned to use this specific principal component to evaluate each of the four pitchers.

As part of my PCA, I converted each survey response to the questions about facial attributes from the Likert scale to a numerical scale, so that a *Strongly disagree* answer received a value of one while a *Strongly agree* answer received a value of seven. The loadings for each principal component signal how much of that principal component reflects a given attribute. A positive loading for a given component means that the attribute is positively correlated on the component. A negative loading, meanwhile, suggests that the attribute is negatively correlated on the component.

I selected the principal component that explains most of the variance among the players, and calculated the average principal component value for each of the nine players. I ranked the players in order of the average value of the principal component that explains most of the variance among the players. The four players who ranked highest by average principal component were selected to be evaluated by the college baseball coaches in Experiment 2.

4.2 Experiment 1B: Measuring attractiveness

In the event that coaches evaluated a player worse when they saw his face in the video of him throwing compared to when they did not see his face, the attractiveness of a player might explain why the coaches grade the player lower when the player's face is visible. Less attractive individuals have been shown to be less likely to be hired and, once employed, earn less money than their more attractive peers (Frieze et al., 1991; Heilman E. & Saruwatari R., 1978). If relatively unattractive faces are rated lower than others, then this could suggest the coach's are negatively biased against pitchers whose faces are not attractive. Because I did not include a question about attractiveness in Experiment 1A, I developed an additional survey that incorporated the attractiveness of the players' faces. I used these survey results to see how including a question about facial attractiveness affected

the principal components in my PCA and the average principal component value for the players that I found in Experiment 1A.

4.2.1 Participants

I recruited 636 additional Amazon MTurk survey respondents for Experiment 1B.[6]

4.2.2 Survey design

I used the same survey to the one described in Experiment 1A but included a fifth statement regarding the attractiveness of each pitchers' face:

This individual is attractive.

Respondents were given the question, *How much do you agree with the following statement?* to this statement, and answered on the same 7-point Likert scale used in Experiment 1A.

4.2.3 Methodology

I performed an additional PCA on the results from this second survey that included a question about facial attractiveness. I followed the exact same methodology for the PCA described in Experiment 1A.

I compared the results from the two separate analyses in Experiment 1A and Experiment 1B to see how adding attractiveness affected the proportion of variance in the players' faces could be explained by each principal component. Moreover, I used this second PCA as a robustness check on the set of results from Experiment 1A, determining how adding a question about attractiveness affected the different principal component and which players had the highest average principal component value.

4.3 Experiment 2: Evaluating baseball pitchers

To estimate the causal effect that observing a pitcher's face has on their evaluation, I conducted a survey in Experiment 2 that asked baseball evaluators to rate players based on their pitching

[6]The demographics of the survey participants are summarized in Table A.2 in the Appendix.

motions.

4.3.1 Participants

I used college baseball coaches as my survey participants in this experiment. College coaches must judge high school pitchers based on their throwing abilities, and they choose the players they desire to recruit for their teams. Thus, college baseball coaches evaluate baseball players on a daily basis and are an ideal choice as the baseball evaluators. I would have sent the survey to professional baseball scouts but did not have access to their email addresses. Moreover, the sample size of college baseball coaches is much greater than the number of professional baseball scouts.[7]

I purchased a set of college baseball coaches' email addresses from College Directories, who offers information regarding all college athletic departments and their coaches contact information. I sent the survey to a set of 4,185 National Collegiate Athletic Association baseball coaches.

I emailed the coaches on three separate days over a two-week period, collecting 440 responses overall. While I received some responses saying that the coach was no longer working at the school to which I sent the emails, 90 percent of the emails I sent were successfully delivered to the email addresses on the purchased set. Each email included a link to my survey and stated that coaches should not take the survey twice.

4.3.2 Survey design

My survey used recorded videos of each of the four pitchers selected from Experiment 1A throwing at maximum effort. To ensure consistency across videos, each pitcher was right-handed, threw a fastball, and stood in an identical location. I only asked the coaches to evaluate four pitchers so as to limit the length of the survey, hoping coaches would be more likely to take a shorter survey. The video clips record the pitchers throwing from the side angle. While this side-angle view is not the only angle coaches look at when scouting pitchers, coaches need to see this angle to observe a pitcher's full torso in motion. A pitching instructor told me that he believes a pitcher's side view is one of the ideal angles when watching a pitcher throw because it reveals a lot about a pitcher's

[7]There are 30 teams in the MLB, with a typical organization employing between 8-12 full-time scouts (Apstein, 2019)

mechanics, torso and full-body rotation, and athleticism.[8]

I filmed each pitcher throwing multiple pitches. Since I wanted to reduce the length of the survey for the coaches to ensure that each coach completed the entire survey, I selected just one throw for each player, choosing what I believed was the highest-quality pitch that included the pitcher's entire body and range of motion. I saved one version of each video in normal speed and converted a second version of each video to slow-motion at a speed of 50 percent of the original video.

After speaking with several college baseball coaches and evaluators, I learned that coaches typically watch pitchers throw at live speed and at a reduced speed.[9] Full speed videos are best for coaches to see how athletic a player moves during their throw, enabling coaches to focus on the pitcher's tempo, rotation, and explosiveness. In contrast, slow-motion videos allow coaches to examine various parts of the body during the throw that they can not catch at live speed. At this stage, I had four separate videos, one for each pitcher, where each video showed the pitcher throwing a pitch at live speed followed by the exact same pitch in slow-motion.

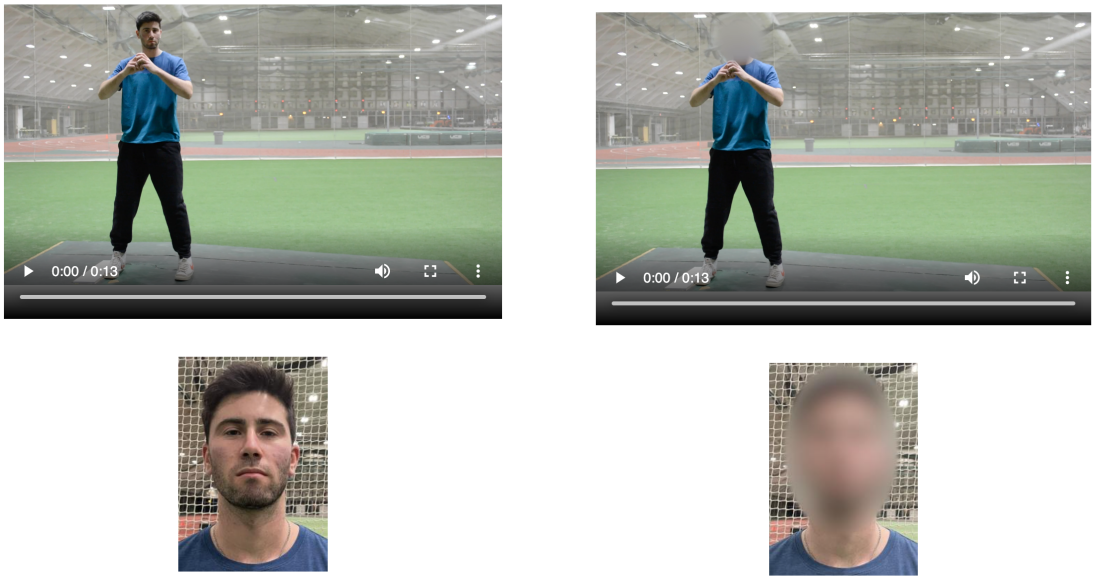
My survey only asked coaches to evaluate the players based on these videos. Coaches typically evaluate players across many different metrics, including their pitching statistics (Lin et al., 2011). However, the coaches I talked to emphasized that they would never choose a player without having watched them throw, either in person or in a video.[10] While there are many different factors that coaches use to evaluate players, seeing a player's pitching motion is the most important dimension when discussing the pitcher's ability. While I could have included statistics for each pitcher, I decided to only show videos of the pitchers throwing to the coaches to reduce the complexity of the evaluation process and make it exclusively about their throwing motions.

I created a separate set of videos, called the blurred version of the videos, by editing the original versions of each video. In the blurred versions of the videos, I edited the footage so that the pitchers' faces were blurred out. When watching the pitchers throw in these particular videos, coaches were not able to see the pitcher's face at all in the videos. After editing, there were eight total videos, four in the unedited set of videos and four in the blurred version of the videos. Moreover,

[8]Conversation with the pitching instructor occurred on January 12th, 2022.
[9]Conversations occurred on October 14, 2021, October 21, 2021, and April 20, 2022
[10]Conversations occurred on October 14, 2021 and October 21, 2021

each pitcher had two videos: one where their face was visible, in the unedited version, and one where their face was blurred, in the blurred version.

To ensure coaches could clearly see a pitcher's face while the pitcher is throwing, I included the pitcher's head shot below the video of them throwing. I added this photo but with the player's face blurred out for the blurred version of the videos to maintain consistency between the two different versions of the survey where the videos are unedited and the version where the pitchers' faces are blurred. Figure 2 shows two versions of the same pitcher. The image on the left is drawn from the unedited version of the video of the pitcher throwing, while the image on the right is drawn from the blurred version of the same pitcher throwing where his face is blurred out in such a way that survey respondents can not see his face.



(a) Unedited version

(b) Blurred version

Figure 2: Unedited and blurred version of a video

The questions in my survey experiment followed a similar path to the surveys in Experiments 1A and 1B. Once coaches consented to take my survey and answer truthfully, they answered questions regarding their demographics and background.[11] Coaches were required to pass three separate attention checks to ensure they read the survey thoroughly and gave high-quality answers. The three attention checks were the same as the ones used in the surveys from Experiments 1A and 1B.

Coaches were randomly shown either the set of the normal videos of the pitchers throwing or the set of blurred videos in which pitchers' faces were not visible. I varied which set of videos each coach saw because I did not want a coach to see both versions of the videos. By seeing a blurred and unedited version of pitchers throwing, a coach might have been able to guess that the survey was trying to understand how their evaluations would be affected by a player's face and potentially rate the player differently than they otherwise would have. The ordering of the four videos in each set was also randomized. Coaches evaluated each of the four pitchers one at a time. Once a coach finished rating one pitcher, the coach would click on a *Next* button to move to the following page and evaluate the next pitcher, until they rated all four pitchers.

For all four pitchers, coaches were asked, *How would you grade this pitcher* on a rating scale from 20 to 80. The 20-80 grading scale is used widely by baseball scouts and endorsed by Major League Baseball to provide a standardized and quantifiable rating system (Gines, 2017; McDaniel, 2014). A player receiving a rating of 20 is considered poor, while a pitcher earning a grade of 80 is viewed as the best. A pitcher with a rating of a 50 is viewed as average (McDaniel, 2014).

4.3.3 Methodology

Data from respondents who failed to pass all three attention checks were not included in the analysis. I employed an ordinary least squares regression to understand the effect that blurring a player's face has on his evaluation rating:

$$Rating = \beta_0 + \beta_1\, Blurred + \beta_2\, Player1 + \beta_3\, Player2 + \beta_4\, Player3 + \beta_5\, Blurred\, Player1 + \beta_6\, Blurred\, Player2 + \beta_7\, Blurred\, Player3 + \epsilon$$

[11] The demographics of the coaches who completed the survey are summarized in Table A.3 in the Appendix.

Standard errors are clustered by individual college coaches, who serve as the raters. The dependent variable *Rating* in the linear regression is the evaluation rating from the college baseball coaches given to the players. The binary variable *Blurred* is a dummy variable for whether the player's face is blurred in the video or not. There are dummy variables for the four players, represented by the binary variables *Player1*, *Player2*, and *Player3*. The intercept, β_0, represents the fourth player's average rating and the other three players are all relative to this baseline.

I included interaction variables between *Blurred* and the dummy variables for each player. The coefficients on these interaction variables, β_5, β_6, and β_7, represent the difference in a player's average rating when the coach could see the player's face compared to when the coach could not see his face. If the coefficient on any of these interaction variables is statistically significant and negative, then this implies that the coaches rate that player higher when they can see his face compared to when their face is not visible in the video. Meanwhile, if the coefficient is statistically significant and positive, then that player is rated higher by the coaches when the coaches cannot see his face compared to when his face is visible. A non-statistically significant coefficient means that the coaches do not evaluate the player any differently when they see his face in the video compared to when they can see his face.

5 Results

In this section, I describe the results from my three survey experiments.

5.1 Experiment 1A

For the first survey, 1,132 of 1,377 respondents passed all three attention checks. Only responses from this subset of survey participants were used in the analysis.[12] To maintain privacy for the nine players used in the survey, their names are not used in the discussion that follows.

I converted the 7-point Likert scale responses to numerical values. Table 1 shows the average response for the pitchers' faces across the four different attributes that I described above, along with how each player ranks in each attribute in terms of average response. A player with a high average

[12] Using the responses that were dropped due to a failed attention check yielded qualitatively similar results.

response for how athletic they look, for instance, implies that the player has an athletic-looking face. If a player has a low average response for athleticism, meanwhile, then the player does not have an athletic-looking face.

Table 1: Average responses for players' facial attributes

	Smart		Athletic		Intimidating		Baseball	
Player	Average	Rank	Average	Rank	Average	Rank	Average	Rank
A	5.01	7	5.30	8	3.98	6	5.11	5
B	4.96	8	5.60	1	4.92	1	5.21	1
C	5.05	6	5.39	6	4.06	4	5.06	6
D	5.20	2	5.43	5	4.09	3	5.13	4
E	5.27	1	5.44	3	4.33	2	5.13	2
F	5.20	3	5.45	2	3.97	7	5.13	3
G	5.08	5	5.44	4	4.03	5	4.96	7
H	5.09	4	5.17	9	3.80	8	4.93	8
I	4.84	9	5.33	7	3.74	9	4.91	9

As evidenced by Table 1, in terms of the average response, Player E has the smartest-looking face while Player B looks the most athletic, intimidating, and most like a baseball player.

The loadings from a PCA show how much each attribute contributes to a specific principal component. Table 2 summarizes the loadings in each of the four principal components, named $PC1_A$ through $PC4_A$.

Table 2: Principal component loadings

Attribute	$PC1_A$	$PC2_A$	$PC3_A$	$PC4_A$
Baseball	0.52	-0.33	0.38	0.69
Intimidating	0.52	0.33	0.57	-0.55
Smart	0.48	0.63	-0.56	0.25
Athletic	0.48	-0.62	-0.46	-0.41
Proportion of variance	0.48	0.20	0.16	0.16

The total proportion of variance for each principal component, also contained in Table 2, reveals how much of the variance in facial attributes each component explains. The first principal component, labeled $PC1_A$, explains 48 percent of the variance among the four attributes. $PC1_A$ has a strong, positive loading for all four attributes, meaning that the four attributes are correlated with $PC1_A$.

Meanwhile, the second principal component, labeled $PC2_A$, loads positively on how intimidating and smart the player looks. How much the player looks like a baseball player and how athletic the player looks negatively loads on $PC2_A$. Therefore, $PC2_A$ describes how intimidating and smart the player is, yet is negatively correlated with how much they look like a baseball player and how athletic the player's face looks. This means that, in $PC2_A$, a player with a more intimidating and smarter-looking face would rank higher on $PC2_A$ while a player whose face looks athletic and like a baseball player's would be lower on $PC2_A$.

Table 3 shows the average principal component for each of the four pitchers who were used in the second survey. Because $PC1_A$ explains nearly half of the variation across the four attributes in my study, players are ranked by their average $PC1_A$ value.

Table 3: Average principal components by player

Player	$PC1_A$	$PC2_A$	$PC3_A$	$PC4_A$
B	0.30	-0.06	0.22	-0.24
E	0.15	0.08	-0.02	-0.01
D	0.07	0.03	-0.05	0.06
F	0.05	0.005	-0.10	0.07
C	-0.03	-0.01	0.004	-0.23
G	-0.04	-0.01	-0.06	-0.04
A	-0.06	-0.01	0.04	0.07
H	-0.20	0.10	-0.03	0.10
I	-0.26	-0.12	-0.02	0.01

I selected the four players who have the four highest average first principal component, evidenced by the dashed line in Table 3, to be evaluated by the college baseball coaches in Experiment 2. Player B ranks first on $PC1_A$, followed by Players E, D, and F.

5.2 Experiment 1B

The survey from Experiment 1B received 636 responses, of which 539 of participants passed all three attention checks.[13] Table 4 summarizes the loadings in each of the five principal components, named $PC1_B$ through $PC5_B$, from the PCA on the results from the second survey.

[13]Using the responses that were dropped due to a failed attention check yields qualitatively similar results.

Table 4: Principal component loadings

Attribute	$PC1_B$	$PC2_B$	$PC3_B$	$PC4_B$	$PC5_B$
Baseball	0.43	0.44	-0.49	0.61	0.12
Intimidating	0.44	-0.26	-0.61	-0.57	-0.22
Smart	0.47	-0.38	0.38	0.37	-0.60
Athletic	0.41	0.68	0.43	-0.42	-0.10
Attractive	0.49	-0.36	0.26	-0.02	0.75
Proportion of variance	0.54	0.15	0.12	0.11	0.08

A PCA on the results from the second survey, which included a question about facial attractiveness, yielded results similar to the PCA in Experiment 1A. $PC1_B$ in the second PCA, shown in Table 4, explains 54 percent of the variance in facial attributes. Similar to the PCA in Experiment 1A, $PC1_B$ in Experiment 1B is the most informative component, explaining just over half of the variance in the players' faces. Each of the five attributes positively loads on $PC1_B$ in Experiment 1B, which also occurred for the four attributes on $PC1_A$ in Experiment 1A.

Because an additional question was included in this survey, the second principal component in Experiment 1B, labeled $PC2_B$, is different from the results found in the PCA in Experiment 1A. $PC2_B$ loads positively on how much a pitcher looks like a baseball player and how athletic he looks, which the opposite from the $PC2_A$ in Experiment 1A. How intimidating, smart, and attractive a player looks negatively loads on $PC2_B$. Therefore, $PC2_B$ describes how much a player looks like a baseball player and how athletic the player is, yet is negatively correlated with how intimidating, smart, and attractive the player's face looks. This means that in $PC2_B$, a player whose face looks athletic and like a baseball player's would be higher on $PC2_B$, while a player with an intimidating, smart, or attractive face would rank lower on $PC2_B$.

Table 5 shows the average principal component value for $PC1_B$ through $PC5_B$ for the nine pitchers.

Table 5: Average principal component values by player

Player	$PC1_B$	$PC2_B$	$PC3_B$	$PC4_B$	$PC5_B$
B	0.28	0.07	-0.16	-0.23	-0.01
F	0.20	-0.01	0.19	0.12	0.04
E	0.18	-0.11	0.04	-0.02	0.06
D	0.12	0.00	-0.01	0.05	0.05
H	-0.03	-0.18	0.07	0.06	-0.03
G	-0.07	0.03	0.00	-0.10	-0.04
A	-0.11	0.00	0.01	0.09	0.01
C	-0.17	0.07	-0.03	0.00	0.04
I	-0.38	0.13	-0.07	0.04	-0.10

As described above, Players B, F, E, and D rank highest by their average first principal component value in Experiment 1A. These same four players have the four highest average $PC1_B$ in Experiment 1B, evidenced by the dashed line in Table 5. My results from Experiment 1B serve as a robustness check on Experiment 1A, in that I can see that including a question about facial attractiveness yields similar qualitative results about how informative the first principal component is in explaining the variance in the pitchers' faces and which four players rank highest on this component.

5.3 Experiment 2

440 collegiate baseball coaches completed the third survey, of whom 369 passed all three attention checks.[14]

[14]Using the responses that were dropped due to a failed attention check yields qualitatively similar results.

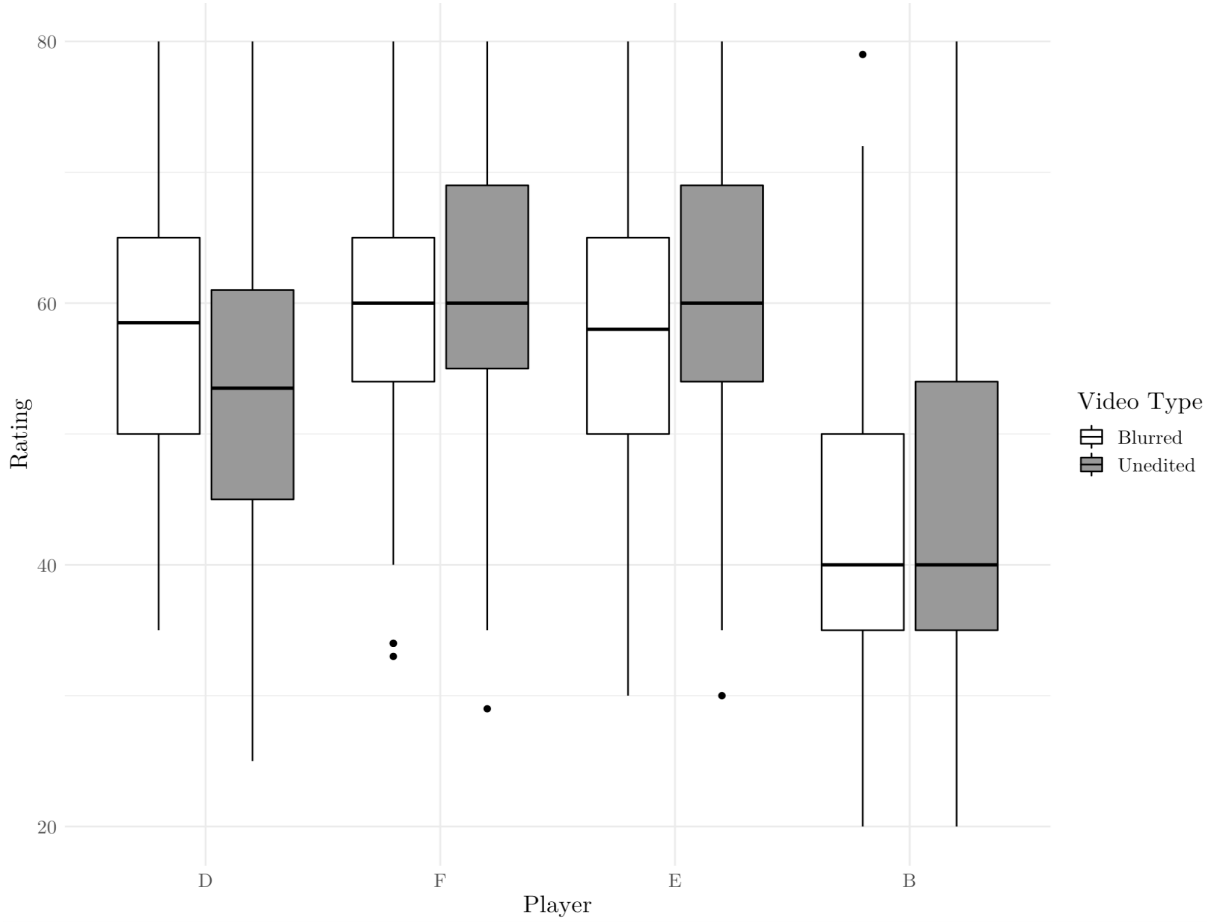


Figure 3: Box plot of player evaluations

Figure 3 is a box plot of the results from the coaches evaluations of each of the four players. Players F, E, and B have slightly higher average ratings when their faces are visible in the unedited version, where the coaches could see the player's face, than they do in the blurred version, where the coaches could not see his face when evaluating him.[15] Player D is the only player who earned a lower average rating when his face was visible to the coaches compared to when the coaches could not see his face. The players received relatively similar ratings except for Player B, who received a much lower rating compared to the other three players.

Table 6 shows the results from the linear regression for the four players evaluated by the college baseball coaches. Player B's average rating is the constant in the regression. The other three

[15]Exact player ratings and the difference between versions of each video can be found in Table A.4 in the Appendix.

players' ratings are all relative to Player B's.

Table 6: Analysis of player evaluations

	Dependent Variable: Rating
Player F	17.178***
	(1.116)
Player E	16.433***
	(0.984)
Player D	9.229***
	(0.897)
Blurred	−0.575
	(1.418)
Blurred * Player F	−1.571
	(1.580)
Blurred * Player E	−2.597*
	(1.432)
Blurred * Player D	5.333***
	(1.236)
Constant	43.720***
	(1.026)

*Note: *p<0.1; **p<0.05; ***p<0.01*

The most important results in Table 6 are the interaction estimates, $Blurred * Player\ F$, $Blurred * Player\ E$, $Blurred * Player\ D$, and *Blurred* variable. For Players F, E, and D, if any of these interactive variables are statistically significant, this means that there is a significant difference between that player's rating when the coaches could see their faces in the videos compared to when they could not see their faces. If the *Blurred* variable is statistically significant, this means that Player B is rated significantly differently by the coaches when they could see his face in the videos.

For Player B, who has the highest average first principal component value in Experiments 1A and 1B, and for Player F, there was no difference in the rating assigned to these pitchers when their faces were visible in the pitching videos compared to when the faces were blurred out. This implies that Hypothesis 1 was true for Players B and F, meaning there is no significant statistical evidence that Player B and F's faces affected how the coaches evaluated them.

Meanwhile, the coaches evaluated Player E higher when they saw his face in the video of him pitching compared to when they could not their faces in the videos. This positive difference is statistically significant ($p < 0.1$), suggesting that Hypothesis 2 was true for Player E. By including his face in the videos, Player E's pitching rating increased when the coaches could see his face compared to when his face was blurred in the videos.

Player D received a statistically significantly ($p < 0.01$) lower rating from the coaches when they could see his face compared to when his face was blurred out. The difference in grades assigned for this pitcher is the largest among the four pitchers. Hypothesis 2 was true for this pitcher, since he earned a statistically significant worse evaluation from the coaching respondents when his face was visible relative to when the coaches rated him without seeing his face.

6 Discussion

The results of my three experiments find evidence that a pitcher's face can bias baseball scouts' player evaluations. If a pitcher's face did not bias baseball evaluators, there would be no significant difference between how they evaluated the pitcher when they saw his face when watching him pitch compared to when they could not see his face. This would mean that Hypothesis 1 would be true for all four pitchers. However, for two of the four pitchers, I find statistical evidence in favor of Hypotheses 2, meaning that the college baseball coaches evaluated the pitchers significantly differently when they could see their faces compared to when they could not see their face. One player's face positively biased the coaches, since the coaches rated this pitcher higher when they could see his face compared to when they could not see his face. Meanwhile, the coaches rated the other pitcher worse when they saw his face in the videos of him pitching compared to when they did not see his face, suggesting this pitcher's face negatively biased the coaches in their evaluation of the pitcher. Of the four pitchers who were evaluated, this pitcher has the worst face in the first principal component, as shown in Table 5 from Experiment 1B. This is consistent with the results that the coaches evaluated him worse when they could see his face in the videos of him throwing. This finding suggests that players with faces that do not look smart, intimidating, athletic, attractive, and like a baseball player's face might receive worse treatment from baseball evaluators than those with faces

who have faces mores consistent with these attributes.

Player B had the highest average first principal component, yet the coaches did not evaluate him any differently when they saw his face compared to when they could not see his face in the videos. Since this pitcher had the highest face on PC1, it could be expected that his face would lead to coaches being positively biased by his face, rating him higher when they see his face in the video compared to when they do not. Looking at Figure 3, Player B received a much worse rating from the college coaches relative to the other three players. Player B's inferior pitching ability could explain why the coaches were not biased by his face. Even though his face was the highest on PC1 in both Experiments 1A and 1B, the coaches were immediately put off by his worse pitching skill that they did not focus as much on his face. Player B's poor pitching talent out-weighs any effect that his face might have on the coaches' evaluations of his abilities.

I find that a pitcher's face can bias how a coach evaluates a baseball pitcher. As Michael Lewis discusses in *Moneyball*, sight-based heuristics cloud baseball scouts' abilities to make accurate judgements on players. If the face is being used to aid baseball evaluators in the selection of their players, is this, too, hurting baseball scouts' ability to make accurate player evaluations? Since I find that baseball evaluators specifically look at a player's face to help make their decisions, how are such facial biases affecting scouts' judgements? Could these facial biases lower the performance of baseball evaluators and reduce the accuracy of scouts' evaluations of pitchers?

There is no empirical research suggesting that a player's face holds any valuable information about the player's ability. Since there is no evidence that using a player's face helps explain their baseball abilities, the evaluation of a player should be determined by their performance and mechanics, not by their face (Lin et al., 2011). Therefore, baseball evaluators selecting, or choosing not to select, specific players based on their facial appearances creates an issue in the evaluation of baseball players. This would entail that many baseball players have been discriminated against based on their face. Scouts could improve their talent identification by ignoring the face altogether when evaluating players. To reduce potential gender, ethnic, or racial biases when evaluating job applicants, some job employers have removed the names of the applicants from their resumes. As a result, these companies have greatly increased the diversity of their workforce (by ContentEngine LLC, 2020). Baseball teams could follow a similar path by blurring out the face in videos of players

to remove any facial biases from the evaluation process.

7 Conclusion

I conclude my thesis by discussing future research about facial biases in baseball evaluation, in addition to my study's limitations.

7.1 Future research

Because this study focused on pitchers, it could be worth exploring how position players' faces affect scouts. I anticipate that scouts are still biased by a player's face no matter their position, yet the actual impact that a player's face has on scouts could differ depending on the player's position. For instance, it would be interesting to examine any possible differences between a pitcher and a catcher, since scouts might be biased to a lesser extent by the face of a catcher, who wears a mask for the majority of a game when he is behind home plate, than they are biased by a pitcher's face.

My findings should not come as a surprise. As discussed earlier in this paper, individuals in all industries judge others based on their facial appearances. While my study does not examine whether these facial biases actually make scouts less accurate in their player evaluations, future research could try to understand the overall impact that these facial biases have on a baseball evaluator's ability to correctly evaluate players. This question is worth exploring to gauge whether or not baseball scouts could improve their talent-identification processes to make more accurate decisions about players, ultimately selecting better players for their teams. One way future research could address this question would be by replicating my methodology but asking evaluators to estimate a real player's statistics and comparing the scout's estimation of the player's statistics to that player's actual statistics. If a scout becomes more accurate when a player's face is removed from the evaluation process, then this could suggest that seeing a player's face makes a scout worse at evaluating players.

7.2 Limitations

One limitation in my study is that I should have carried out Experiments 1A and 1B as one experiment, instead of employing them as separate surveys. I chose the four players from

Experiment 1A to use in Experiment 2, where the baseball coaches evaluated these four players, before I asked about pitcher attractiveness. Since I had already chosen the four players for the coaches to evaluate, I really could only use this extra question about facial attractiveness in Experiment 1B as a robustness check on the results from Experiment 1A. Ultimately, I selected the same four players from Experiment 1B, but should have asked about the attractiveness of the pitchers' faces in the survey in the first experiment instead of creating an additional survey.

Additionally, my study only asks coaches to evaluate four pitchers. To provide stronger evidence that facial biases exist in baseball evaluation, my study should be replicated using many more pitchers to understand whether these facial biases exist over a larger sample of baseball pitchers.

Finally, I gave baseball coaches only one piece of information when making their evaluations of the four pitchers. Scouts typically have more quantitative metrics about the players and additional videos of the pitchers throwing from other angles. A more realistic evaluation of a player would have included several angles of the pitcher throwing, including from behind the pitcher's head to see where the ball is caught in the catcher's glove relative to where the catcher set up his glove at the start of the pitch. This would be included along with more objective information about the pitcher, such as the velocity and spin rate of his fastball. While my single video is useful in making an initial observation about a player, coaches do need much more information on a player to make a more accurate evaluation of a pitcher.

My results suggest that coaches are biased by a pitcher's face. Sight-based biases in baseball have been well documented. This study hopefully contributes to the understanding of facial biases in general and sight-based heuristics in baseball evaluation by specifically examining the how the face biases baseball scouts.

References

Ansorge, C. J., & Scheer, J. K. (1988). International bias detected in judging gymnastic competition at the 1984 olympic games. *Research Quarterly for Exercise and Sport*, *59*(2), 103–107.

Apstein, S. (2019). Scouts are looking for life after baseball. *Sports Illustrated*. `https://www.si.com/mlb/2019/05/08/mlb-scouts-analytics-decline`

Aranow, P. M., Kalla, J., Orr, L., & Ternovski, J. (2020). Evidence of rising rates of inattentiveness on lucid in 2020.

Bellemore, F. A. (2001). Racial and ethnic employment discrimination: Promotion in major league baseball. *Journal of Sports Economics*, *2*(4), 356–368.

Berri, D. J., Simmons, R., Van Gilder, J., & O'Neill, L. (2011). What does it mean to find the face of the franchise? physical attractiveness and the evaluation of athletic performance. *Economics letters*, *111*(3), 200–202.

Berry, D. S., & McArthur, L. Z. (1986). Perceiving character in faces: The impact of age-related craniofacial changes on social perception. *Psychological bulletin*, *100*(1), 3–18.

Bigler, M., & Jeffries, J. L. (2008). "an amazing specimen": Nfl draft experts' evaluations of black quarterbacks. *Journal of African American studies (New Brunswick, N.J.)*, *12*(2), 120–141.

Billings, A. C. (2004). Depicting the quarterback in black and white: A content analysis of college and professional football broadcast commentary. *The Howard journal of communications*, *15*(4), 201–210.

Burge, G. S., & Zillante, A. (2017). Racial discrimination and statistical discrimination: Mlb rookie card values and performance uncertainty. *Social Science Quarterly*, *98*(5), 1435–1455.

Burger, J. D., & Walters, S. J. K. (2009). Uncertain prospects: Rates of return in the baseball draft. *Journal of Sports Economics*, *10*(5), 485–501.

by ContentEngine LLC, T. (2020). What is a neutral or blind resume? (English ed.). *CE Noticias Financieras*.

Coupe, T., Gergaud, O., & Noury, A. (2018). Biases and strategic behaviour in performance evaluation: The case of the fifa's best soccer player award. *Oxford Bulletin of Economics and Statistics*, *80*(2), 358–379.

Curtis, J. E., & Loy, J. W. (1978). Positional segregation in professional baseball: Replications, trend data and critical observation. *International Review of Sport Sociology*, *13*(4), 5–23.

Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *The Review of financial studies*, *25*(8), 2455–2483.

Eggers, B. (2012). The most important offseason acquisition for the san francisco giants could be hadoop. *TechCrunch*. `https://techcrunch.com/2012/11/16/baseball-and-big-data-how-hadoop-and-the-san-francisco-giants-are-taking-on-stats/`

Ferrucci, P., Tandoc, E., Hong, S., Almond, A., & Leshner, G. (2013). Game of colors: Implicit and explicit stereotypes in coverage of america's pastime prepared for association for education in journalism and mass communication annual conference minorities and communication division.

Findlay, L. C., & Ste-Marie, D. M. (2004). A reputation bias in figure skating judging. *Journal of Sport Exercise Psychology*, *26*(1), 154–166.

Flowe, H. D., & Humphries, J. E. (2011). An examination of criminal face bias in a random sample of police lineups. *Applied cognitive psychology*, *25*(2), 265–273.

Frieze, I. H., Olson, J. E., & Russell, J. (1991). Attractiveness and income for men and women in management. *Journal of Applied Social Psychology*, *21*(13), 1039–1057.

Gines, D. S. (2017). Tastes for true talent: How professional baseball scouts define talent and decide who gets to play.

Heilman E., M., & Saruwatari R., L. (1978). When beauty is beastly: The effects of appearance and sex on evaluations of job applicants for managerial and nonmanagerial jobs. *Organizational Behavior and Human Performance*, *23*.

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019a). Lay beliefs in physiognomy explain overreliance on facial impressions.

Jaeger, B., Evans, A. M., Stel, M., & van Beest, I. (2019b). Explaining the persistent influence of facial cues in social decision-making. *Journal of experimental psychology. General*, *148*(6), 1008–1021.

Korva, N., Porter, S., O'Connor, B. P., Shaw, J., & Brinke, L. t. (2013). Dangerous decisions: Influence of juror attitudes and defendant appearance on legal decision-making. *Psychiatry, psychology, and law*, *20*(3), 384–398.

Kramer, R. S. S. (2017). Sequential effects in olympic synchronized diving scores. *Royal Society open science*, *4*(1), 160812–160812.

Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion (Washington, D.C.)*, *7*(4), 730–735.

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. W. W. Norton Company.

Lin, W.-B., Chen, M.-J., Tung, I.-W., & Chen, M.-Y. (2011). An analysis of an optimal selection process for characteristics and technical performance of baseball pitchers. *Perceptual and Motor Skills*, *113*(1), 300–310.

Lindbergh, B., & Arthur, R. (2019). What we learned from our unprecedented access to mlb scouting reports. *The Ringer*. `https://www.theringer.com/mlb/2019/3/4/18249155/cincinnati-reds-scouting-report-series-part-1-data-findings`

McArthur, L. Z., & Apatow, K. (1984). Impressions of baby-faced adults. *Social cognition*, *2*(4), 315–342.

McDaniel, K. (2014). *Scouting explained: The 20-80 scouting scale*. `https://blogs.fangraphs.com/scouting-explained-the-20-80-scouting-scale/`

Menold, N. (2020). Rating-scale labeling in online surveys: An experimental comparison of verbal and numeric rating scales with respect to measurement quality and respondents' cognitive processes. *Sociological Methods & Research*, *49*(1), 79–107.

Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of nonverbal behavior*, *27*(4), 237–254.

Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: The salience and significance of age in social judgments. *Advances in Experimental Social Psychology*, *30*(100), 93–161.

Morgan, H. N., & Rotthoff, K. W. (2014). The harder the task, the higher the score: Findings of a difficulty bias. *Economic Inquiry*, *52*(3), 1014–1026.

Olivola, C. Y., Eubanks, D. L., & Lovelace, J. B. (2014). The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance. *The Leadership quarterly*, *25*(5), 817–834.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in cognitive sciences*, *18*(11), 566–570.

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? reexamining the diagnostic value of appearance-based inferences. *Journal of experimental social psychology*, *46*(2), 315–324.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

Paolacci, G., & Chandler, J. (2014). Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

Parsons, C. A., Sulaeman, J., Yates, M. C., & Hamermesh, D. S. (2011). Strike three: Discrimination, incentives, and evaluation. *American Economic Review*, *101*(4), 1410–35.

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PloS one*, *7*(3).

Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological science*, *19*(2), 109–111.

Sailofsky, D. (2018). Drafting errors and decision making theory in the nba draft.

Seltzer, R., & Glass, W. (1991). International politics and judging in olympic skating events. *Journal of Sport Behavior*, *14*(3), 189–200.

Smith, E., & Harrison, C. K. (1996). Stacking in major league baseball. *Journal of African American men*, *2*(2/3), 113–129.

Spurr, S. J. (2000). The baseball draft: A study of the ability to find talent. *Journal of Sports Economics*, *1*(1), 66–85.

Streib, N., Young, S. J., & Sokol, J. (2012). A major league baseball team uses operations research to improve draft preparation. *Interfaces (Providence)*, *42*(2), 119–130.

Todorov, A., & Duchaine, B. (2008). Reading trustworthiness in faces without recognizing faces. *Cognitive neuropsychology*, *25*(3), 395–410.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual review of psychology*, *66*(1), 519–545.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Van't Wout, M., & Sanfey, A. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*(3), 796–803.

Voyer, D., & Wright, E. F. (1998). Predictors of performance in the national hockey league. *Journal of Sport Behavior*, *21*(4), 456–473.

Williams, K. M., Park, J. H., & Wieling, M. B. (2010). The face reveals athletic flair: Better national football league quarterbacks are better looking. *Personality and individual differences*, *48*(2), 112–116.

Zebrowitz, L. A., & Montepare, J. M. (1992). Impressions of babyfaced individuals across the life span. *Developmental psychology*, *28*(6), 1143–1152.

Appendix

Table A.1: Experiment 1A survey participant demographics

Demographic	Proportion of Sample (%)
Gender	
Man	61.3
Woman	31.7
Other	3.9
Race	
White	74.7
Asian	11.1
Black or African American	8.2
Other	6.0
Age	
18-24	6.0
25-34	51.2
35-44	25.0
45-54	10.7
55-64	5.1
65+	2.0

Table A.2: Experiment 1B survey participant demographics

Demographic	Proportion of Sample (%)
Gender	
Man	61.9
Woman	35.7
Other	2.4
Race	
White	74.7
Asian	11.1
Black or African American	8.2
Other	6.0
Age	
18-24	5.2
25-34	53.8
35-44	25.1
45-54	10.7
55-64	4.2
65+	0.8

Table A.3: Experiment 2 survey participant demographics

Demographic	Proportion of Sample (%)
Gender	
Man	93.0
Woman	0.2
Other	6.8
Race	
White	86.8
Asian	0.0
Black or African American	1.1
Other	12.1
Age	
18-24	1.7
25-34	32.6
35-44	30.0
45-54	21.3
55-64	12.3
65+	2.2

Table A.4: Pitcher evaluations

Player	Rank in $PC1_B$	Video	Mean Rating	n	t	p-value
B	1	Unedited	43.71	157		
		Blurred	43.14	159		
		Difference	0.57		0.41	0.68
F	2	Unedited	60.76	157		
		Blurred	58.75	157		
		Difference	2.01		1.76	0.079 *
E	3	Unedited	60.08	158		
		Blurred	56.80	161		
		Difference	2.85		2.85	0.0046***
D	4	Unedited	52.94	158		
		Blurred	57.59	158		
		Difference	-4.65		-3.71	0.0002***

Note: *p <0.1; **p <.05; ***p <.01

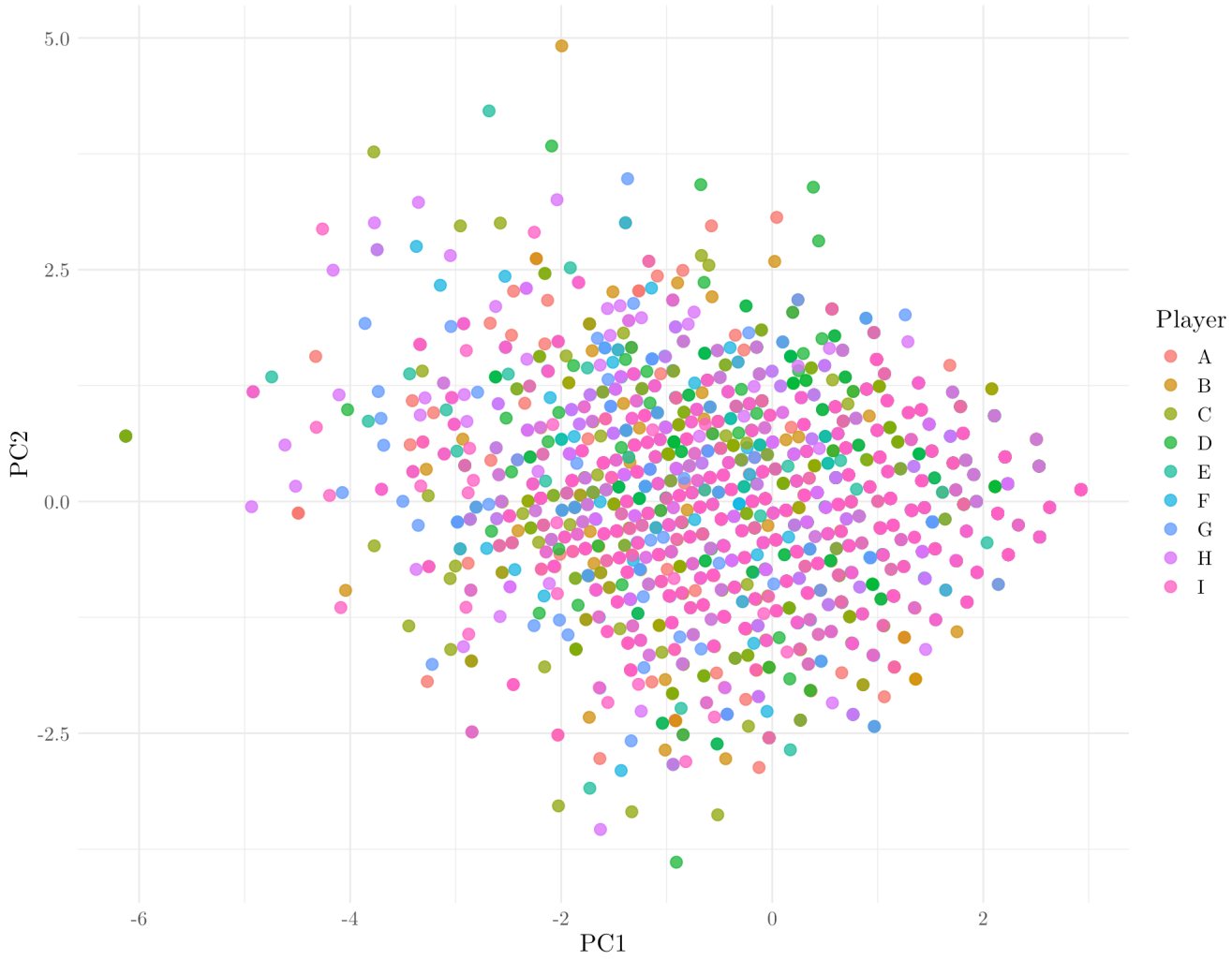


Figure A.2: $PC1_A$ plotted against $PC2_A$ from PCA in Experiment 1A

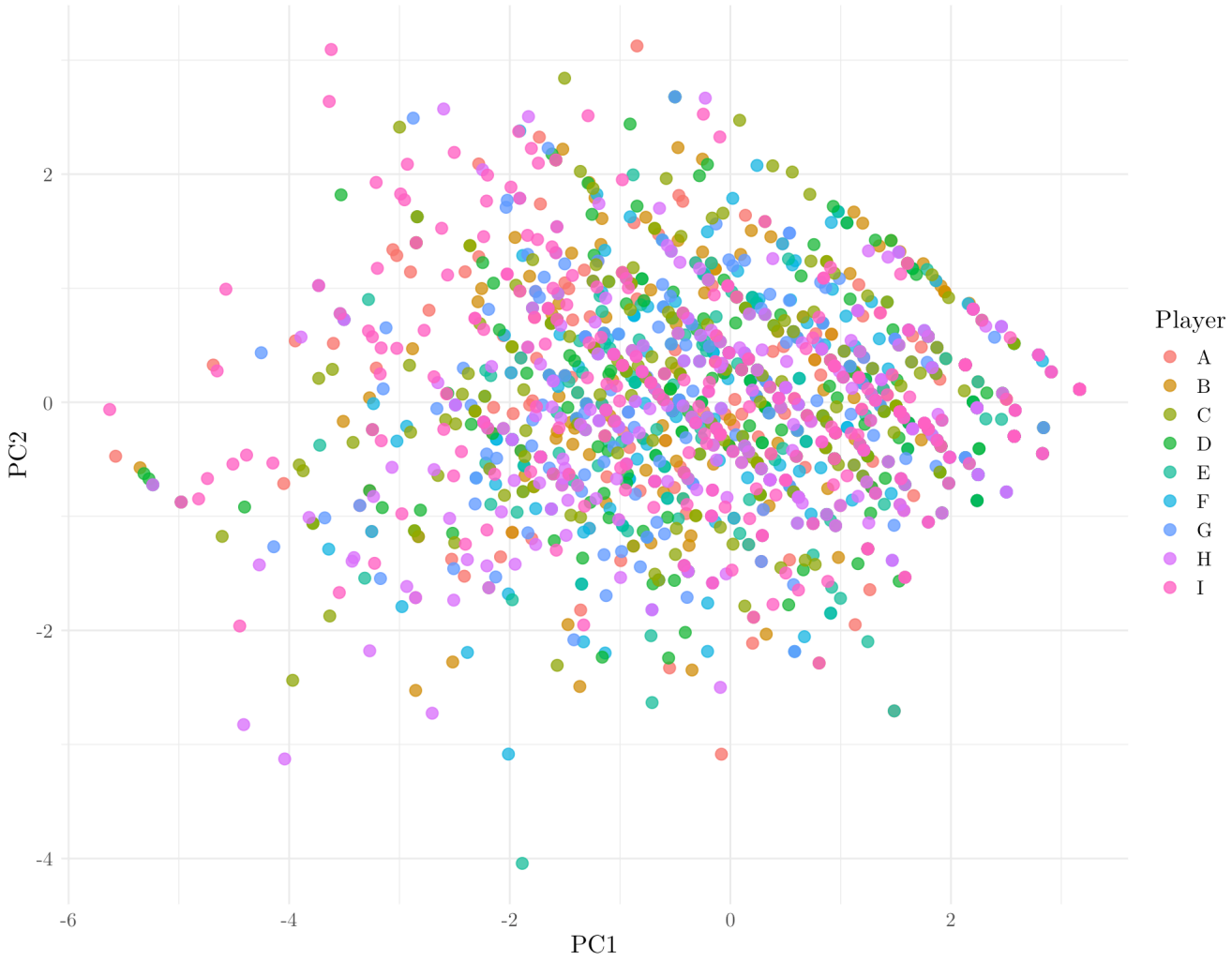


Figure A.2: $PC1_B$ plotted against $PC2_B$ from PCA in Experiment 1B