2021

Would “Robot Umpires” Reduce Discrimination? Measuring Racial Bias in Major League Baseball Umpires

Hank Snowdon

Claremont McKenna College

Would "Robot Umpires" Reduce Discrimination? Measuring Racial Bias in Major League Baseball Umpires

submitted to
Professor Angela Vossmeyer

by
Hank Snowdon

for
Senior Thesis
Spring 2021
April 30, 2021

Abstract

Utilizing thirteen years of Major League Baseball pitch-tracking and play-by-play data, this study investigates racial discrimination by umpires when making pitch calling decisions. Two models are formulated, one that predicts the probability of a strike erroneously being called a ball (batter favoritism) and one that predicts the probability of a ball erroneously being called a strike (pitcher favoritism). The probabilities are modeled as a function of whether or not the pitcher's or batter's race is the same as the umpire's. With over 3 million pitch observations, multiple sub-sample and time trend analyses are conducted to examine with whom the discrimination lies and how it changes throughout the sample. The results suggest that umpires are significantly more likely to make calls that favor players of the same race, and that these effects have not diminished between 2008 and 2020. Furthermore, these biases seem mostly held by White umpires, who account for a wide majority of umpires in MLB.

Acknowledgements

I would first like to thank Professor Angela Vossmeyer, Ph.D. for her invaluable guidance throughout this project. Without her feedback, encouragement, and expertise this paper would not have been possible. I would also like to thank my parents for their never-ending support, and for introducing me to baseball so many years ago. I would never be where I am today without all they have done for me. Lastly, I am deeply grateful for all my incredible friends that have made the last four years so entertaining, interesting, and rewarding.

Contents

1 Introduction

The study of racial discrimination in Major League Baseball is not a new subject — numerous researchers have widely explored the topic, across a variety of angles, in the last half-century. MLB was a segregated league until 1947, when Jackie Robinson famously broke the "color barrier," the rule that barred Black players from playing alongside Whites. The league has become increasingly integrated in the years following Robinson's feat, yet, it is unclear if Black players now face the same opportunities as their White counterparts.

A recent influx of public-facing baseball data now allows for advanced analyses into potential racial discrimination in MLB. One main area of possible bias is umpiring — despite calls for robotic umpires as sports around the world turn digital, MLB continues to employ real people to call pitches in their games. These umpires are tasked with deciding whether or not pitches (which often travel upwards of 100 mph), travel through an imaginary box, known as the strike zone. If the pitch traveled across home plate, and was between the batter's knees and chest, the umpire calls a strike — otherwise the pitch is a ball. This decision is one of the most important officiating activities in sports, and is one that can happen over a hundred times in each game. With each MLB team playing 162 games a season, a staggering number of these calls are made per year.

Data from these pitches provide an ideal empirical setting to tackle the following question — are umpires influenced by racial biases when they call balls and strikes? While discrimination can often be difficult to quantify and identify, the nature of pitch calling decisions lends itself perfectly to identifying race-based biases. Each call by an umpire is, by definition, right or wrong — either the pitch crossed through the zone or it did not. By modeling the probability of these missed calls, using variables that identify the umpire's race and those of the pitcher and batter, umpire discrimination can be identified and measured.

This paper considers pitches thrown between 2008 and 2020 — the entire period in which pitch data has been collected and published by MLB. Two probabilistic models are constructed, one that predicts missed calls which favor the batter (strikes called balls), and one that predicts missed calls which favor the pitcher (balls called strikes). These models are functions of whether or not the umpire's race matches that of the batter and/or the pitcher, along with many other controls.

With over 3 million observations, the extensive sample used considers thousands of pitches per potential race-match combination. This allows for comprehensive restricted sample analyses, which measure how the effects change across different umpire races. The length of the sample also allows for in-depth analyses of changes in discrimination over time, by examining the trends of the race-match effects over thirteen seasons.

After undertaking this empirical approach, this paper presents clear evidence of umpire racial discrimination. Overall, umpires exhibit pitcher favoritism when their race matches the pitcher's, and batter favoritism when their race matches the batter's. Further, the time trend analyses suggest that the biases do not meaningfully decrease in size or significance throughout the sample. It also appears that the biases are largely driven by White umpires, while Hispanic umpires exhibit slight evidence of discrimination *against* Hispanic players, and discrimination shown by Black umpires is largely minimal or insignificant depending on the specification. These results, along with the fact that most MLB umpires are White, indicate that White players benefit the most from umpire discrimination in MLB.

These effects could have serious consequences for certain players and teams, as missed calls can have huge impacts on careers and game outcomes. As more teams adopt data-driven, results-oriented approaches to their front offices, the margins of success can be increasingly thin for players on the cusp of making it in the league. If minority players are called out on borderline pitches due to their race, it could potentially have drastic effects for some. The evidence presented in this paper provides greater incentives for MLB to consider electronic strike zones if the league is serious about providing equal opportunities to all players, regardless of race.

2 Literature

The labor discrimination model outlined by Becker et al. (1971) is useful to frame the discussion of different types of economic discrimination. Becker identifies three main sources of taste-based discrimination — prejudice from consumers, co-workers, and employers. Consumer-based discrimination occurs when customers will pay more for goods and services from groups they like, and less from those they do not. In baseball this would occur if fans disliked minority-race players, as they would spend less on tickets, memorabilia, etc. Employer-based discrimination occurs when employers exhibit prejudices against workers of a certain background, and would pay more — while likely limiting their production — to not employ them. Before 1947 baseball was an extreme example of this, as teams could have paid significantly less for talented Black players, but refused to do so as a consequence of their racist tendencies. Finally, co-worker discrimination exists when firms recognize that employing co-workers that discriminate against eachother would lower their production — this could partially explain why some MLB teams refused to integrate for so long.

Some of the earliest research into racial discrimination in MLB was conducted by Pascal and Rapping (1970). While the authors do not find any evidence of racial discrimination on players' salaries in the 1968 season, they do find other relevant results. For one, they show that player bonuses in the 1950s varied significantly by race, but that those differences were eliminated by 1968 — they attribute this decline to either a drop in league-wide prejudice or an increase in the quality of player information. Further research continued through the 1970s. Gwartney and

Haworth (1974) show that in the years immediately following integration, having more Black players on a team was a significant predictor of an increase in wins. They attribute this result to the fact that Black players were proportionally cheaper and more productive than their White counterparts at this time, as they demanded lower wages on average. This wage gap does appear to have diminished over time, however, as Irwin (2004) demonstrates that race was not a significant predictor of salary for players signing free agent contracts between 1997-2003. This is one of a number of additional studies on this topic. The Gwartney and Haworth study also investigates consumer discrimination — recognizing that while hiring more Black players had a positive effect on winning at the time, consumer discrimination (in the form of fans spending less money on tickets) could offset the financial benefits of winning for teams. Instead they find the opposite, showing that in the 1950s, each additional Black player employed was correlated with an increase in yearly attendance between 50,000 and 60,000 fans. Following Gwartney and Haworth, numerous other studies furthered the research into consumer discrimination in MLB. Sommers and Quinton (1982) find no evidence of fan-based discrimination when investigating team revenues from 1976-77. Nonetheless, consumer discrimination has been proven in some instances, mainly in studies of baseball cards and memorabilia. Nardinelli and Simon (1990) investigate the determinants of 1989 prices for 1970 Topps baseball cards, controlling for variables like player career performance. They find race had a significant effect on card price — cards sold for greater than 10% more, all else equal, for White players than their minority counterparts. Conversely, Gabriel, Johnson, and Stanton (1995) conduct a similar study for rookie cards between 1984-1990 and do not find evidence that price differences were driven by race.

Overall, the research discussed above on racial discrimination in MLB presents mixed results. While some research suggests there is racial bias toward non-Whites in MLB, there is no wide consensus. Salary-based discrimination appears to have decreased over time, as player performance and salaries have become increasingly public information. Consumer discrimination could still be prevalent, but it has not been proven consistently on a wide scale.

Rather than focusing on consumer discrimination or employer discrimination, this paper builds upon previous research on potential umpire racial bias toward MLB players. Research on officiating bias is a deeply explored topic in and of itself. Looking at soccer, Nevill, Newell, and Gale (1996) prove that home teams are given significantly more penalty kicks by officials than visiting teams, when examining evidence from English and Scottish leagues. Sutter and Kocher (2004) furthered this research, showing that home teams are given more extra time when trailing in games, and that visiting teams were awarded significantly fewer penalties. Garicano, Palacios-Huerta, and Prendergast (2005) find the same results regarding extra time, and show that the effects increase when crowd sizes are larger. These studies demonstrate that referee bias can exist on a significant level — in this case, to conform to social pressure and satisfy home fans. This social pressure theory was extended to MLB umpires by Mills (2014), whose results suggest that umpires make more favorable pitch calls for both players that are more experienced and have higher status, as well as for catchers, as they are physically closer to umpires throughout the game.

In perhaps the most well-known study on racial discrimination and officiating, Price and Wolfers (2010) investigate calls in the National Basketball Association from 1991-2004. The paper examines whether or not referees call more fouls on players that do not have the same race as them. Their findings are strong and significant — up to 4% fewer fouls were called on players whose races matched the officials'. They also find that these players scored up to 2.5% more points when the officials were same-race. They conclude that these racial biases were large enough to significantly affect the probabilities of teams winning games.

To extend this research to baseball, Parsons et al. (2011) examine MLB umpires' pitch calling decisions from 2004-2008. They mirror Price and Wolfers' methodology by studying umpire-pitcher race matches and non-matches to see if pitch calls were significantly different for the two. The study shows that umpires were more likely to make favorable pitch calls when their race matched that of the pitcher. However, they find this result to only be significant in parks where the umpires were not monitored by MLB's QuesTec evaluation system, technology that was only implemented in one-third of the stadiums during the selected time period. As PITCHf/x pitch tracking technology was not introduced across all MLB stadiums until 2008, this study does not use pitch tracking data to evaluate the umpires. Instead, they use total strike percentage as their main dependent variable.

While Parsons et al. did find evidence of discrimination, some following studies using similar methodologies have not robustly reproduced the results. Tainsky, Mills, and Winfree (2015) consider MLB games from 2007-2008, and using a wide variety of specifications, show that evidence of race-match discrimination is sensitive to the both the methodologies and specifications used. Birnbaum (2010) also shows that evidence of this discrimination is limited.

PITCHf/x technology's league-wide implementation, as mentioned above, has given a reason to revisit this research. These cameras track detailed information about every pitch thrown in MLB games, including pitch location, velocity, break, acceleration, release point, and more. This has allowed for researchers to build predictive models on the pitch tracking data and test the significance of other covariates on umpires' calls. Kim and King (2014) take this approach to show that player status leads to umpires making more pitch calls in their favor. Their methodological strategy is to split bad calls into ones whose quality are under-recognized (the umpires call strikes balls) and pitches whose quality are over-recognized (the umpires call balls strikes). In more common baseball terminology, this can be seen as an umpire "widening" or "tightening" the strike zone. Verducci (2017) follows the same methodological approach, along with the idea of umpire-player race matches outlined by Parsons et al. and Price and Wolfers, to determine the effect of race on umpire decisions. After running logistic regression models that predicted the chance of over or under-recognition, Verducci finds somewhat weak evidence of pitch-calling racial discrimination by umpires, with the highest levels of bias coming from White umpires against Hispanic players. This study examines pitch data from 2015 and 2016.

This paper contributes to the literature on racial discrimination in baseball by extending Verducci's analysis over a significantly longer period — 2008-2020, the entire period in which pitch tracking data is available. The panel structure of the data provides a unique look at potential changes in umpire discrimination over time, and allows for a much more complete analysis. The past studies discussed above all look at individual snapshots of time in MLB — it appears no work has been done to study the trend of umpire discrimination over time while using pitch-tracking data. The trend is likely of greater importance to league officials than the snapshots, as policy measures could more substantially be made on the basis of rising or falling discrimination. This analysis has multiple objectives — to more fully determine if this type of racial discrimination exists at all, while also determining if it is increasing or decreasing. Further, the behavior of White umpires is compared to that of Hispanic and Black umpires, a breakdown that would be difficult without using thirteen seasons of pitch data.

3 Data

3.1 Overview

The sample used in the empirical analysis contains observations at the individual pitch level. Along with information about the location and metrics of each pitch, the dataset contains variables which capture the situation surrounding each pitch, including runners on base, outs, count, etc. The sample also contains race information for the pitcher, batter, and umpire involved, along with terms that indicate when those races match.

3.2 Player and umpire race classification

One of the major obstacles to conducting this research is the racial classification of the players and umpires in the sample. Multiple strategies have been taken in other studies to make these types of classifications. For instance, Parsons et al. had research assistants manually assign players to races, while Verducci created an Amazon Mechanical Turk task in which random Internet users code players and umpires by race according to their pictures.

I employ a five-part strategy. First, I take advantage of publicly available data in the form of Wikipedia lists. The first list used is "African-American baseball players," a list which contains all baseball players identified on their Wikipedia pages as African-American. The contents of this list are scraped from Wikipedia and used to construct a "black" variable in my dataset. The second Wikipedia list used is the "List of current Major League Baseball players by nationality." Through this list the nationality of almost all current players is identified, and with the help of a list of

Hispanic and Asian countries, I begin to construct "hispanic" and "asian" variables. Next, World Baseball Classic rosters are used to further identify players' nationality, which allows me to categorize more players as Hispanic and Asian. The WBC is a tournament in which national baseball teams compete against each other every four years (with the exception of the first two tournaments, which happened three years apart). The tournament has been held in 2006, 2009, 2013, and 2017. Through the use of these tournament rosters, more players in my sample that played for Hispanic and Asian countries are classified. The Lahman Database, a vast baseball database constructed by Sean Lahman, is subsequently utilized to then identify the birthplace of every player in my dataset. This data is supplemented by player information collected from Retrosheet, as the Lahman data had not yet been updated during my research to include players who debuted in the COVID-shortened 2020 season. Players born in Hispanic and Asian countries are also classified using this data. As the final player-race identification strategy, the r-package "predictrace" is used. After inputting a player's last name, this package returns the likelihood (based on US Census data) that the name belongs to someone that is Native American, Asian, Black, Hispanic, White or multiple races. Using these probabilities, I manually check players with a high probability of being Hispanic (that are not yet labeled as Hispanic) by looking up their nationality information on Baseball Reference. The same strategy is taken for Black and Asian players. This last step accounts for any players not yet categorized as any race other than White.

A few checks are also in place to verify the accuracy of these classifications. It is important to recognize that the Wikipedia data is less reliable than the birthplace data for two main reasons. The first is that Wikipedia is open source, and can be edited by anyone, which creates opportunities for error. Second, the Wikipedia lists do not have unique identifiers for each player as the Lahman database does, which opens the possibility to duplicate names and misclassification. Due to these potential problems, I manually checked each player identified as Black on this list on Baseball Reference to ensure they were Black, and removed a few inaccurate categorizations. I also manually checked every name that appeared more than once in the sample, in order to ensure players' races were categorized correctly.

Since only 161 umpires worked an MLB game behind home plate between 2008-2020, I was able to manually determine each umpire's race by looking them up on the Internet, using a combination of Wikipedia, CloseCallSports, Google Images, and Baseball Reference to determine their races. Of the 161 home-plate umpires, 135 are White, eleven are Black, and eleven are Hispanic. While the umpire races heavily skew White, this should not present problems with the analysis, as the number of pitches thrown over thirteen seasons is overwhelmingly large.

A small number of games in the sample have multiple home plate umpires listed — this was often due to rainouts that caused games to be rescheduled with different umpiring crews at later dates. Pitches from these games were removed from the sample, as it was not possible to determine which umpires were calling which specific pitches with the data at hand. Pitches involving Asian pitchers or batters were also removed from the data, as they only make up a small minority of total players, and no home plate umpires were identified as Asian. Further analysis would be necessary to study potential pitch-calling discrimination against Asian players in MLB.

The final step in gathering the race data is constructing the variables of interest. An umpire-pitcher race match variable (UPM) and an umpire-batter race match variable (UBM) are created — these terms equal one when the race of the umpire and the batter or pitcher is the same.

3.3 Pitch-tracking and play-by-play data

The pitch-by-pitch data used in this study is collected via the MLB Stats API, which houses data from MLB Advanced Media's (MLBAM) Gameday application. This dataset contains detailed observations on every pitch thrown in MLB since 2008. PITCHf/x tracking technology installed in every stadium collected the pitch-tracking data until 2017, when Trackman pitch-tracking systems replaced PITCHf/x. Both systems use high-resolution cameras to collect detailed measurements of the ball's flight toward home plate. The other game-related data present was manually collected by MLB operators. This data is scraped through the "baseballr" r package, created by Bill Petti. The package is also used to download a data frame which matches umpires to each game in the sample.

The resulting data scraped from MLBAM contains an observation for every pitch thrown in an MLB game between 2008-2020, with variables for game situation and pitch information, along with identifiers for the batters, pitchers, and umpires involved with each pitch. To select the observations of interest, the pitches are filtered only for called balls and called strikes. This ensures the model only examines pitches in which the umpire made a pitch-calling decision. The pitches are further filtered to remove outliers, following a similar methodology to Mills (2013). This excludes pitches more than eight inches outside either side of the plate, those more than 7.2 inches above or below the zone, and those with an unknown pitch type, along with intentional balls and pitch-outs. These filters remove pitches that are clear balls, leaving only situations in which the umpire had to make an actual decision to call the pitch a ball or a strike.

According to MLB rules, the strike zone is defined as "the area over home plate from the midpoint between a batter's shoulders and the top of the uniform pants — when the batter is in his stance and prepared to swing at a pitched ball — and a point just below the kneecap." [1] To determine whether or not a pitch should have been called a strike or not, a strike zone is constructed for each at-bat according to this rule. MLB-employed "stringers" manually mark numeric values for the top and bottom of the zone for each plate appearance, which serve as the top and bottom of the strike zones in my analysis. A pitch is considered "over the plate" if its x-position is less than 10 inches from the center of the plate. Although the plate is only 8.5 inches wide from both sides of the center, a standard MLB ball is 3 inches in diameter — the 10 inch window allows for all pitches in which more than half of the ball crossed over the plate. Consequently, a pitch is considered as "inside the zone" if its vertical height is between the top and bottom borders of the zone, and is also "over the plate."

1. The strike zone has changed multiple times in history, most recently in 1996.

Since the zone_top and zone_bottom values are manually determined by MLB operators, a few erroneous values exist in the dataset. To ensure only realistic strike zones are modeled, all pitches are removed in which zone_top is less than zone_bottom. Furthermore, pitches are removed if the zone_bottom values are not between 1 and 2.75 ft off the ground or the zone_top values are not between 2.75 and 4.25 ft.

Other MLBAM glitches allow for a small number of pitches to have three strikes or four balls at the beginning of the pitch — since these are impossible occurrences, these observations are also removed. I also remove pitches that occurred past the 10th inning, to remove outlier innings without enough observations to produce a significant effect.

After filtering for only called balls and strikes, removing pitches with Asian batters or pitchers, and removing observations with unrealistic strike zones or counts, the final sample for analysis includes 3,379,235 pitches thrown between 2008 and 2020. For all of these pitches, covariates are included for the count, outs, inning, away team batting, pitcher and batter wins above replacement, pitcher and batter All-Star appearances, season, distance from strike zone, run margin, pitcher and batter handedness, runners on base, and runner(s) stealing, along with the race match variables described in Section 3.2.

4 Descriptive Statistics

The econometric modeling approach later outlined in Section 5 considers two models, one for both pitches inside and outside the strike zone. This approach allows me to examine two types of missed calls — those which favor the batter and those that favor the batter. For a pitch outside the zone called a strike, the umpire "over-recognizes" the pitch's quality, which benefits the pitcher. When a strike is called a ball, the umpire "under-recognizes" the pitch's quality, benefiting the batter. Table 1 depicts the overall distribution of these missed calls by umpires in the sample.

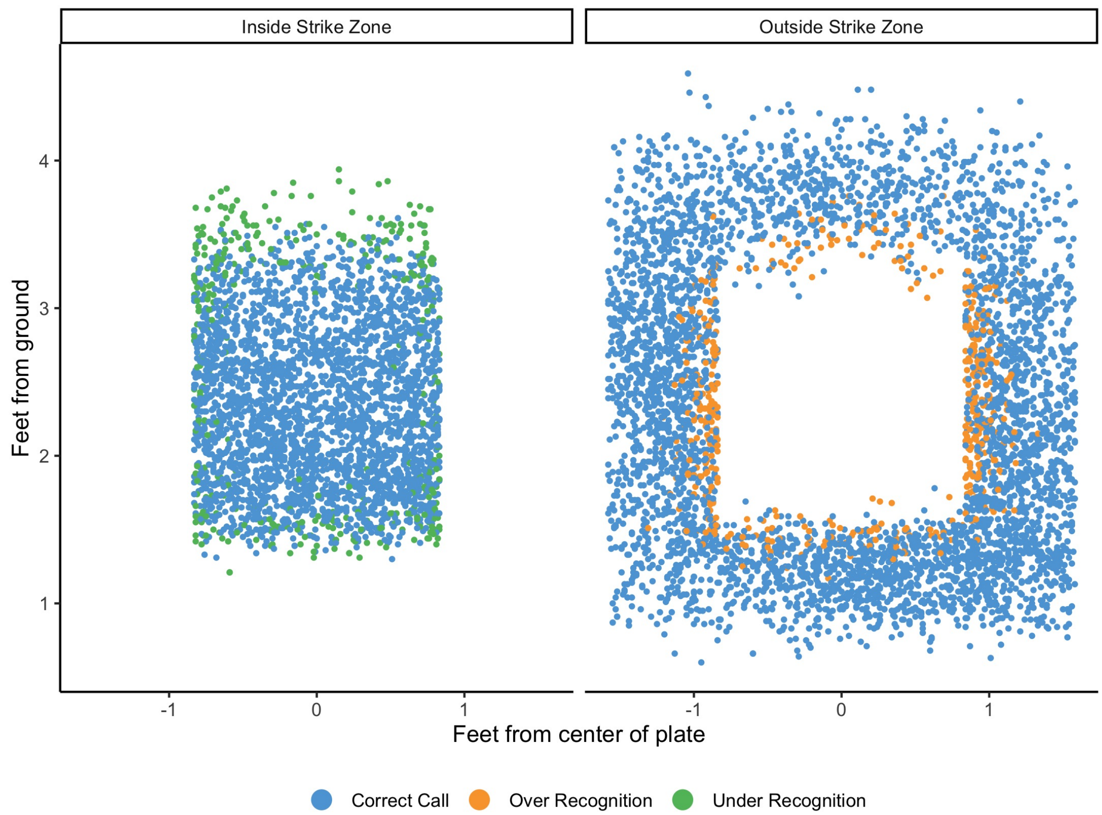
Table 1. Distribution of missed calls

	Called Strike	**Called Ball**	**Total**
Inside Zone	1,163,685 (77.98%)	256,287 (22.02%)	1,419,972 (100%)
Outside Zone	264,973 (15.64%)	1,694,290 (84.36%)	1,959,263 (100%)

Of the 3,379,235 pitches included in the full sample, 1,959,263 are outside the strike zone, while 1,419,972 are inside the zone. Over-recognition occurs on 15.64% of the pitches outside the zone, while under-recognition occurs 22.02% of the time on pitches in the zone. These rates are lower when considering all pitches thrown in a game, but the rates are higher here as the sample only considers pitches in the close vicinity of the strike zone.

Meanwhile, a visualization of the calls is presented in Figure 1. Pitch locations in blue represent correct calls, while green pitches are strikes called balls and orange pitches are balls called strikes.

Figure 1. Pitch locations per correct or incorrect call



Last 8000 pitches from 2020 season plotted.

Table 2 presents the fraction of pitches called by each race of umpire by year. As mentioned before, there are 161 home-plate umpires in the sample — 135 of which are White, while eleven are Black and eleven are Hispanic. As seen in the table, recent years have clearly seen growth in the amount of minority umpires, as multiple umpires from Hispanic countries have debuted in recent seasons (CloseCallSports 2020a, 2020b, 2020c, 2015, 2016).

Looking at overall rates of over and under-recognition by race-matches in Table 3, we see some slight patterns that could potentially suggest discrimination is present

Table 2. Yearly breakdown of pitches called by umpire race

UMPIRE RACE	**Black**	**Hispanic**	**White**
2008	5.68%	5.29%	89.03%
2009	5.06%	4.23%	90.71%
2010	4.55%	5.40%	90.05%
2011	5.42%	5.39%	89.18%
2012	5.24%	6.03%	88.73%
2013	5.12%	5.37%	89.52%
2014	4.81%	6.11%	89.08%
2015	4.69%	6.78%	88.52%
2016	4.57%	8.74%	86.68%
2017	4.48%	8.54%	86.98%
2018	5.73%	8.76%	85.51%
2019	6.17%	9.49%	84.33%
2020	6.38%	15.19%	78.43%
Observations	174,426	233,457	2,971,352

in the sample. Over-recognition (pitcher favoritism), occurs .08 percentage points more often when the umpire's race only matches the pitcher's, compared to when the umpire's race only matches the batter's. For under-recognition (batter favoritism), a similar trend follows — the bad calls happen .32 percentage points more often when the umpire's race only matches the batter's race, versus matching only the pitcher's race. This is not a robust way of proving discrimination, nevertheless, it is useful to examine these trends before diving into the analytical approach of the study. Many other variables have the potential to drive the probability of a bad call being made by an umpire, thus, an econometric model is necessary to disentangle the effects of race on these bad calls.

Table 3. Over and under-recognition by race matches, entire sample

	Over-recognition (hurts batter)	**Under-recognition (hurts pitcher)**
Umpire-pitcher race match, no umpire-batter match	13.46% (574,134)	17.87% (416,951)
Umpire-batter race match, no umpire-pitcher match	13.38% (310,569)	18.19% (224,461)
Umpire race different from batter and pitcher	13.18% (362,586)	17.72% (260,606)
Umpire, pitcher, batter all have same race	13.82% (711,974)	18.30% (517,954)

Selected sample sizes denoted in parentheses.

5 Empirical Methodology

Two probabilistic models are developed to investigate if and to what extent umpires favor players of the same race. The first model predicts the probability of a strike erroneously being called a ball (batter favoritism) and the second model predicts probability of a ball erroneously being called a strike (pitcher favoritism).

As mentioned before, the main covariates of interest in this study are the indicator variables for an umpire-batter race match (UBM) and umpire-pitcher race match (UPM). An interactive term (UPM x UBM) is also included to capture the effect of all races being the same for a given pitch.

Other covariates are also included to account for other causes of variation in umpire pitch-calling behavior. To capture the most important determinant of good calls and bad calls, a dist_x variable is included that measures the distance to the closest edge of the plate, along with dist_z, which measures the distance to the top or bottom of the zone, whichever is closest. Pitcher and batter handedness are also included, as pitch location and flight paths are significantly different depending on which hands the batters and pitchers throw and pitch with. Controls for the pitchers' and batters' "star power" are also considered, as past research suggests that umpires make more favorable calls toward players with higher status. This is included through the use of wins-above-replacement (WAR)[2], a stat that captures the total effect a player has on his team winning games, along with variables that count how many All-Star Games the batter and pitcher had played in the years before the pitch was

2. Baseball Reference's version of WAR is used — Fangraphs also calculates WAR but uses a slightly different formula.

thrown. Game situation is accounted for by including covariates for runners on base, inning and run differential, in order to capture the effect of leverage of a single pitch. A dummy variable for a runner stealing during the pitch is also included, as catchers are unable to "frame" borderline pitches in these cases (catch them in a way that makes the pitch look like a strike). Finally, yearly indicator variables are included to account for changes across seasons in how umpires call pitches.

As discussed previously, two dependent variables are used in the analysis: over-recognition and under-recognition. This follows the framework outlined by Kim and King (2014). When an umpire calls a pitch a strike that should have been a ball, he is over-recognizing the pitch's quality as it pertains to the strike zone. The opposite is true for under-recognition. By using the covariates explained above to predict the likelihood of over-recognition and under-recognition, any potential bias by MLB umpires can be identified. If one/any of the race matches are a significant predictor of under or over-recognition, then evidence of discrimination is present in the analysis.

A logistic regression model is employed. The latent specification for over-recognition is:

$$y_i^* = UPM_i\beta_1 + UPM_i\beta_2 + \text{controls}_i'\beta_3 + \varepsilon_i \tag{1}$$

where $\varepsilon_i \sim \lambda$ (logistic pdf). The mapping from latent y_i^* to the observed y_i:

$$y_i = \begin{cases} 1 & \text{if called strike, pitch is ball} \\ 0 & \text{if called ball, pitch is ball,} \end{cases} \tag{2}$$

leaving the probabilistic formulation:

$$\Pr(\text{Called Strike but is Ball}|X, \beta) = \Lambda(UPM_i\beta_1 + UPM_i\beta_2 + \text{controls}_i'\beta_3). \tag{3}$$

For under-recognition, the latent specification is:

$$y_i^* = UPM_i\beta_1 + UPM_i\beta_2 + \text{controls}_i'\beta_3 + \varepsilon_i \tag{4}$$

where $\varepsilon_i \sim \lambda$ (logistic pdf). The mapping from latent y_i^* to the observed y_i:

$$y_i = \begin{cases} 1 & \text{if called ball, pitch is strike} \\ 0 & \text{if called strike, pitch is strike,} \end{cases} \tag{5}$$

leaving the probabilistic formulation:

$$\Pr(\text{Called Strike but is Ball}|X, \beta) = \Lambda(UPM_i\beta_1 + UPM_i\beta_2 + \text{controls}_i'\beta_3). \tag{6}$$

The logistic models are estimated using maximum likelihood:

$$\hat{\beta}_{MLE} = argmax_{\beta} lnf(y|\beta). \tag{7}$$

Several additional analyses are conducted to further understand the effects of race on umpires' decisions. In order to dive deeper into the discrimination faced by separate races, the model is also run for the subsamples of White umpires, Hispanic umpires, and Black umpires separately. By completing these sub-sample analyses, it can be identified which races face particular discrimination by different umpires. Further, time trend analyses are also conducted to examine the overall trend of discrimination in the sample. In these specifications, UPM and UBM are interacted with a time trend variable, which converts the year indicator terms into a continuous linear trend.

6 Interpretation

Before diving into the results, a quick guide is presented to interpreting the output from the logistic regression specifications. For each specification described, there is one that predicts over-recognition, and one that predicts under-recognition. For the over-recognition models, only pitches thrown outside the zone are considered. For the outcome variable to equal one, the pitch had to be called a strike incorrectly. Over-recognition is an occurrence that helps pitchers and hurts batters — any pitch that was over-recognized should have resulted in a ball, but instead was called a strike. As a result, the signs of the UPM and UBM race match variables have opposite effects. For over-recognition, evidence of umpire discrimination against players with different races exists if the UPM term is positive, and the UBM term is negative. If these are the outcomes, the model predicts that when an umpire is calling a pitch for a player that has the same race, he will make more of those over-recognized calls for same-race pitchers, and fewer for same-race batters.

The opposite is true for under-recognition. When a pitch inside the zone is under-recognized (called a ball when it should be a strike), the umpire's mistake hurts the pitcher and helps the batter. Consequently, if discrimination is present, the UPM term is negative and the UBM term is positive. More same-race batters will have strikes called balls, and fewer same-race pitchers will have strikes called balls.

Lastly, it is important to note that these effects flip when including pitcher and batter race dummy variables instead of the UPM and UBM terms. This is the case in Section 7.4, when the sample is filtered for one umpire race at a time, and the effects of race are decomposed to determine which races of players are discriminated against by which umpires. To illustrate, if for an over-recognition specification of only White umpires the indicator term for Hispanic pitchers is negative, then that is evidence of discrimination against Hispanic pitchers, as less balls outside the zone are called strikes for Hispanic pitchers compared to White pitchers.

The logistic regression output has been converted from the default log odds to average marginal effects, computed through maximum likelihood estimation. The resulting outputs can be interpreted as the effects of a one-unit change in the covariate on the probability of over or under-recognition, depending on the specification. A coefficient of .005, for example, would mean that a one-unit increase in that covariate causes, on average, a .5 percentage point increase in the probability of over or under-recognition. Marginal effects are discussed instead of log odds due to ease of interpretation, and the ability to measure yearly effects by evaluating the time trend interaction terms at each year.

7 Results

7.1 Main Results

Table 4: Marginal covariate effects on over-recognition

Covariate	dy/dx	Std. Err.
UPM	0.0032292***	0.0004545
UBM	-0.0028134***	0.0004359
1 Ball	0.0176974***	0.0005124
2 Balls	0.0465601***	0.000804
3 Balls	0.0839497***	0.0013492
1 Strike	-0.0582044***	0.0005103
2 Strikes	-0.0890718***	0.0006223
1 Out	0.0009956*	0.0005357
2 Outs	-0.00905***	0.0005243
2nd Inning	-0.0035336***	0.000875
3rd Inning	-0.0024652***	0.0008757
4th Inning	-0.0004463	0.0008827
5th Inning	0.0009895	0.0008919
6th Inning	0.000632	0.0008899
7th Inning	0.0023464***	0.0008999
8th Inning	0.0024267***	0.0009054
9th Inning	0.0096272***	0.0009937
10th Inning	-0.001084	0.0023723
Runner Stealing	-0.0285549***	0.0023997
Top of Inning	0.0027854***	0.0004317
Run Margin	0.000558***	0.0000707
Right-Handed Batter	-0.0329657***	0.0004516
Right-Handed Pitcher	0.0045131***	0.000507
Dist_x	0.2861215***	0.0007699
Dist_z	0.3385679***	0.0006192
Bases Loaded	-0.085896***	0.0009977
Men On Base	-0.0605972***	0.0005334
Runner(s) in Scoring Position	-0.0507458***	0.0005648
Pitcher WAR	0.0007063***	0.0001195
Batter WAR	-0.0004909***	0.0001052

Pitcher All-Star Appearances	0.0046991***	0.0001875
Batter All-Star Appearances	-0.0008194***	0.0001257
2009	0.0019092*	0.0011349
2010	-0.0024182**	0.0011313
2011	-0.0156969***	0.0011055
2012	-0.0194362***	0.0011079
2013	-0.0293378***	0.0010936
2014	-0.025384***	0.0011055
2015	-0.0331787***	0.0010923
2016	-0.0431078***	0.0010749
2017	-0.0406842***	0.0010785
2018	-0.0490136***	0.0010687
2019	-0.0394066***	0.0010936
2020	-0.040765***	0.001502
Classification		87.69%
Observations		1,959,263

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

First we examine the results of the baseline specifications, as seen in Tables 4 and 5. Starting with Table 4, it is clear that a number of the covariates included have strong effects when predicting an umpire's over-recognition of a pitch's quality. The count when the pitch is thrown is a highly significant predictor of the umpire's likelihood of expanding the strike zone — umpires are far more likely to over-recognize a pitch's quality in counts that favor the batter, and less likely to do so in counts that favor the pitcher. This is in line with the consensus in baseball research, which has shown that umpires zones' are tighter in pitchers' counts and wider in hitters' counts (Walsh 2010; Green 2014). The results also suggest that a positive run margin for the batting team increases the chance of a call outside the zone hurting the batter — this can likely be interpreted as a similar effect of a batter being ahead in the count, as the umpire consciously or subconsciously tries to even the game.

Having one out in the inning is barely statistically different from zero outs, but having two outs is a strong negative predictor of over-recognition. Inning is also somewhat important for predicting over-recognition — compared to the first inning of the game, these bad calls generally happen more often in the late innings and less often in the early ones. A runner stealing a base during a pitch also has a negative effect on over-recognition, likely due to the fact that the catcher is unable to “frame” the pitch, since he has to focus on throwing the runner out. As expected, the terms that have the strongest effect on the chance of a bad call are the dist_x and dist_z variables, which measure the length to the closest horizontal and vertical edges of the zone. These terms are negative for pitches outside the zone, so a positive shift means the pitch is closer to being a strike. When a pitch is closer to the zone, there is logically a positive effect on the chance of that pitch being incorrectly called a strike. Additionally, over-recognition is more likely to happen in the top of the inning — consistent with research discussed previously that suggests away teams face an officiating disadvantage (away teams hit exclusively in the top of the inning in baseball). As mentioned earlier, past research has also demonstrated that the status of the batter or pitcher could lead to better calls; this trend follows in my analysis, with the pitcher’s wins-above-replacement and number of All-Star appearances leading to more calls in their favor, and the same for batters. The results also suggest that the handedness of the pitcher and batter have an effect — with over-recognition less likely for right-handed batters, and more likely for right-handed pitchers. The effect of having runners on base appears to tighten umpire’s zones as well — compared to having the bases empty, having men on base, runners in scoring position, or having the bases loaded all significantly decrease the probability of an umpire widening his zone for a borderline pitch. Additionally, it can be seen that compared to 2008 (the first year pitch tracking data was recorded), over-recognition has decreased over time — this demonstrates that umpires' calls are generally improving.

With respect to the main variables of interest for this study — umpire-batter race match and umpire-pitcher race match — this specification clearly demonstrates that umpires treat same-race players better than different-race players when calling pitches outside the strike zone. If the batter's race matches the umpire's race, all else equal, the batter is .28 percentage points less likely to have a ball called a strike on them. By multiplying this effect by the sample size, we find that the model predicts that the batter's race matching the umpire's caused roughly 5,512 fewer incorrect calls for pitches outside the strike zone. Conversely, if the pitcher's race matches the home plate umpire's race, all else equal, the pitcher is .32 percentage points more likely to have one of their pitches be called a strike, when it should be a ball. This accounts for roughly 6,327 more missed calls in the sample.

Table 5: Marginal covariate effects on under-recognition

Covariate	**dy/dx**	**Std. Err.**
UPM	-0.0016171***	0.0005721
UBM	0.0032253***	0.0005444
1 Ball	-0.0202838***	0.0006778
2 Balls	-0.0490812***	0.0008833
3 Balls	-0.0852144***	0.0010102
1 Strike	0.0740611***	0.0007136
2 Strikes	0.1472923***	0.0014778

1 Out	-0.0001409	0.000648
2 Outs	0.0156134***	0.0006738
2nd Inning	0.0057694***	0.0011003
3rd Inning	0.0042618***	0.0010982
4th Inning	0.0042608***	0.0011155
5th Inning	0.0029593***	0.0011175
6th Inning	0.0048308***	0.0011199
7th Inning	0.0056801***	0.0011194
8th Inning	0.0062738***	0.0011222
9th Inning	0.0014981	0.0012118
10th Inning	0.0166167***	0.0030258
Runner Stealing	0.0458703***	0.0025334
Top of Inning	-0.0020213***	0.00054
Run Margin	-0.00017*	0.0000879
Right-Handed Batter	0.0016738***	0.0005552
Right-Handed Pitcher	-0.0022635***	0.0006152
Dist_x	-0.2198308***	0.0011155
Dist_z	-0.5408094***	0.0008926
Bases Loaded	0.1205555***	0.0020203
Men On Base	0.0663865***	0.0007294
Runner(s) in Scoring Position	0.062251***	0.0007593
Pitcher WAR	-0.0002976**	0.0001506
Batter WAR	0.0015676***	0.0001304
Pitcher All-Star Appearances	-0.0038818***	0.0002575
Batter All-Star Appearances	0.0002799*	0.0001559
2009	-0.0259599***	0.0014596
2010	-0.0556211***	0.0014308
2011	-0.0481128***	0.0014452
2012	-0.0572415***	0.0014489
2013	-0.0706468***	0.001441
2014	-0.0887801***	0.0014214
2015	-0.0918921***	0.0014292
2016	-0.0902373***	0.0014249
2017	-0.1202492***	0.0013829
2018	-0.1347736***	0.0013617
2019	-0.1480808***	0.0013436
2020	-0.1523477***	0.0017317
Classification		85.66%
Observations		1,419,972

**** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$*

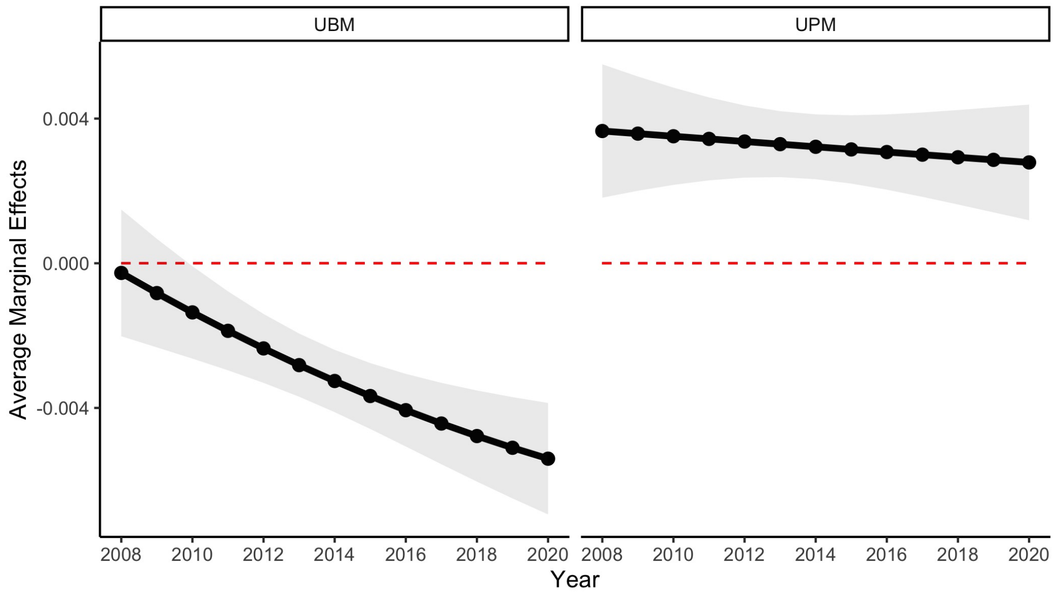
Looking at the main specification for under-recognition, the effects of our covariates of interest generally hold for pitches inside the strike zone — as expected, the signs of the marginal effects of the covariates are all flipped from over-recognition. The only effect that does not flip are the yearly indicator terms — the likelihood of a strike being called a ball also decreases with time. This suggests that umpire quality has increased overall in the last 13 years, as pitches inside and outside the zone are more likely to be called correctly. This is a logical result, as umpires currently face more scrutiny than ever — both from MLB and from fans, due to the widespread publication of pitch-tracking data.

Again, we see strong evidence of discrimination when examining the average marginal effects for the race-match terms for under-recognition. If the pitcher's race is the same as the umpire's race, they have a .16 percentage point lower chance of receiving one of these incorrect calls which hurt them — a reduction of 2,296 incorrect calls inside the strike zone. If the batter's race is the same as the umpire's, the pitch is .32 percentage points more likely to be called a ball when it should be a strike, an outcome that benefits the batter. The model predicts that this caused 4,580 additional missed calls. In the main under-recognition and over-recognition specifications, it is clear umpires are more likely to make calls that favor same-race players.

7.2 Trends in discrimination over time

The next subject of interest is whether or not these race-match effects change over time. To address this, additional specifications are run that include UBM and UPM interaction terms with a time trend variable, which converts the season of each

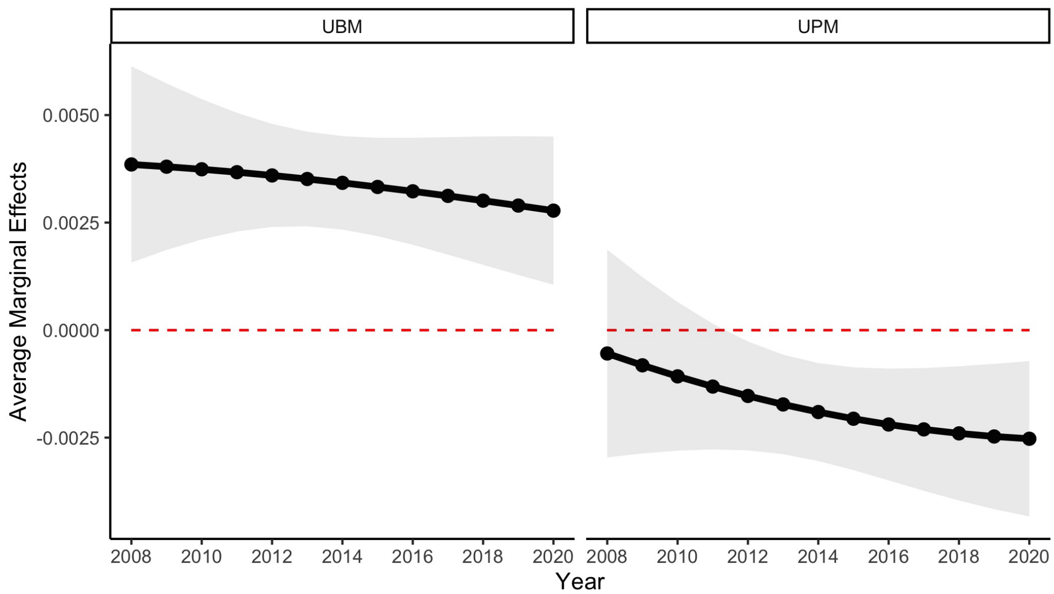
Figure 2. Umpire-player time trend race match effects: over-recognition



Marginal effects evaluated yearly, 95% confidence intervals included.

pitch into a continuous linear trend. For the results seen in Figure 2, this model is evaluated for average marginal effects at each season, computing the yearly effects of UPM and UBM on over-recognition. In this specification, we see that the effects of UPM are slightly decreasing overall, but that they stay strong and significant over the entirety of the sample. The effects of UBM are statistically insignificant in the first two years, and become more negative over time. This suggests that for over-recognition, umpire discrimination has not significantly diminished, even though pitch-calling has become more accurate overall over the last 13 years. If there was evidence that discrimination was decreasing, the two lines in Figure 2 would trend toward zero.

Figure 3. Umpire-player time trend race match effects: under-recognition



Marginal effects evaluated yearly, 95% confidence intervals included.

When looking at the yearly effects of UPM and UBM on under-recognition, the results are similar. The average effects of UPM are insignificant in the first three years of the sample, but become increasingly negative over time. The effects of UBM do decrease slightly, but remain strong and significant throughout the entire sample.

7.3 Heterogeneity across umpire races

After determining that racial bias exists in pitch calling consistently, we now look to determine who is driving this discrimination. In order to disentangle potential heterogeneity, sub-sample analyses by umpire race are considered. In Table 6 the results of this sub-sample analysis are shown for over-recognition.

When looking only at White umpires, which make up 86% of the umpires in the sample, the marginal effects are significant and slightly stronger than the base

Table 6. Umpire-player race match effects on over-recognition per umpire race

UMPIRE RACE	Black	Hispanic	White
UPM	-0.00598 (0.00469)	-0.00654*** (0.00188)	0.00373*** (0.000513)
UBM	-0.00635** (0.00262)	0.00456** (0.00179)	-0.00376*** (0.000470)
Observations	101,234	136,305	1,721,724

Std. Errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

specification with all races included, suggesting that White umpires could be at fault for a disproportionate amount of this discriminatory behavior. Indeed, there is no evidence that Hispanic umpires discriminate against non-Hispanic players on pitches outside the zone; in fact, evidence of the *opposite effect* is present. On pitches that should be called balls, Hispanic umpires are *more likely* to make calls that hurt Hispanic players than non-Hispanics. This is perhaps attributable to the recent increase in MLB umpires from Hispanic countries, who could potentially fear the consequences of appearing biased toward Hispanic players. This result further suggests that White umpires could be the driving force behind discrimination in MLB, as the overall effects are muted by these opposite effects for Hispanic umpires. When looking only at Black umpires, there is not any evidence of discrimination toward pitchers, and only slight evidence that they discriminate against non-Black batters. The other results for our control variables generally hold across these specifications.

We see somewhat similar results for under-recognition. Looking again only at White umpires, the effect of umpire-batter race match is stronger than in the base

Table 7. Umpire-player race match effects on under-recognition per umpire race

UMPIRE RACE	**Black**	**Hispanic**	**White**
UPM	0.0154** (0.00612)	-0.00337 (0.00242)	0.00000931 (0.000646)
UBM	0.00357 (0.00340)	0.000538 (0.00226)	0.00431*** (0.000584)
Observations	73,192	97,152	1,249,628

Std. Errors in parentheses
**** p<0.01, **p<0.05, * p<0.1*

specification. However, there exists no evidence from the model that White umpires discriminate against non-White pitchers on pitches that should be called strikes. For Hispanic umpires, under-recognition does not appear to be driven by batter or pitcher race, as neither the UPM or UBM terms are statistically significant. Slight evidence of discrimination on Black players is present for Black umpires on pitchers, another unusual result, while the UBM term is not significant for Black umpires' calls.

Ultimately, while the UPM term is indeed insignificant for White umpires, the overall evidence presented in this sub-sample analysis suggests that the majority of the discrimination in MLB is driven by White umpires, directed toward non-White players.

7.4 Race-specific discrimination decomposition

To examine which races of players are explicitly discriminated against the most, we examine Table 8, which includes the results from model specifications for all three umpire races, and indicator terms for pitcher and batter race. By examining these marginal effects, we can further disentangle the overall UPM and UBM effects discussed in Section 7.3.

Table 8. Player race-specific effects on over-recognition per umpire race

UMPIRE RACE	**Black**	**Hispanic**	**White**
Black Batter		-0.00525** (0.00256)	-0.00034 (0.00067)
Hispanic Batter	0.00777** (0.00301)		0.00594*** (0.0005386)
White Batter	0.00563** (0.00273)	-0.00438** (0.00186)	
Black Pitcher		0.00764* (0.00464)	-0.00772*** (0.0011997)
Hispanic Pitcher	0.00821 (0.00501)		-0.00309*** (0.0005431)
White Pitcher	0.00520 (0.00472)	0.00650*** (0.00189)	
Observations	101,234	136,305	1,721,724

Std. Errors in parentheses
**** $p<0.01$, **$p<0.05$, * $p<0.1$*

For White umpires, the results suggest that for pitches outside the zone, Hispanic batters, along with Black and Hispanic pitchers, face umpire discrimination. The effect of the batter being Black is insignificant. For Hispanic umpires, the reverse-bias effect mentioned in Section 7.3 holds for both Black and White batters and pitchers. Lastly, the bias against non-Black batters by Black umpires is relatively consistent for both Hispanic and White batters.

Shifting the focus to under-recognition in Table 9, White umpires appear to have significant bias against Hispanic batters for pitches in the zone, but not against

Table 9. Player race-specific effects on under-recognition per umpire race

UMPIRE RACE	**Black**	**Hispanic**	**White**
Black Batter		0.00013 (0.00329)	-0.00113 (0.00085)
Hispanic Batter	-0.00575 (0.00384)		-0.00595*** (0.00066)
White Batter	-0.00225 (0.00352)	-0.00075 (0.00234)	
Black Pitcher		0.00897 (0.00607)	0.00461*** (0.00154)
Hispanic Pitcher	-0.01635** (0.00651)		-0.00076 (0.00068)
White Pitcher	-0.01498** (0.00616)	0.00318 (0.00243)	
Observations	73,192	97,152	1,249,628

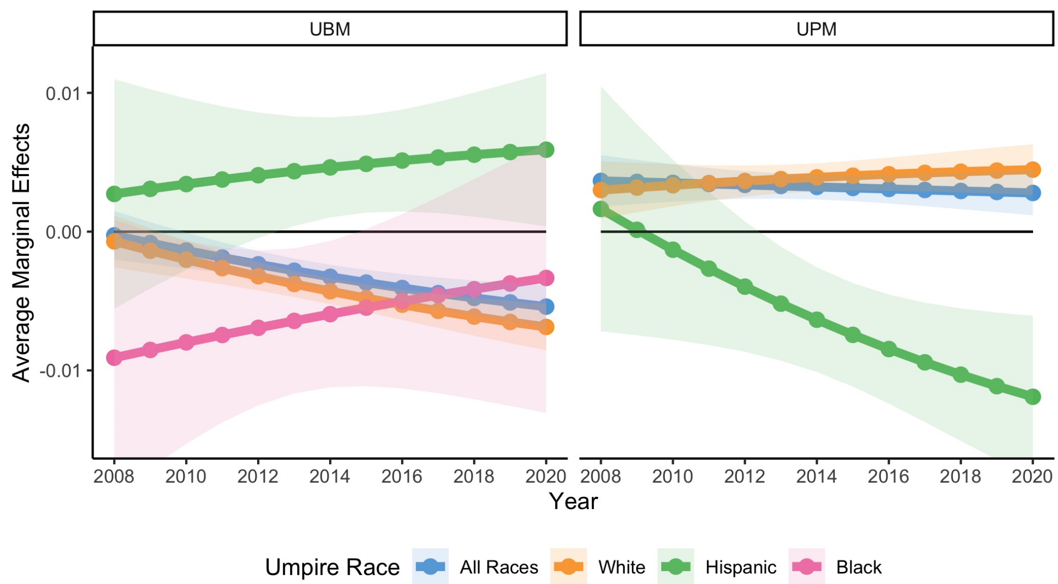
Std. Errors in parentheses
**** p<0.01, **p<0.05, * p<0.1*

Black batters. This breakdown also gives context to the insignificant effect of UPM for White umpires mentioned in Section 7.3, as discrimination against Black pitchers is actually positive and significant, while it is slightly negative but insignificant for Hispanic pitchers. It appears these effects potentially offset one another, leaving the net UPM effect insignificant from zero. None of the race indicators suggest significant evidence of discrimination for under-recognition by Hispanic umpires, while the favoritism toward non-Black pitchers by Black umpires is fairly equal for both Hispanic and White pitchers.

7.5 Time trends in umpire race sub-samples

We now investigate the trends of the umpire-race specific discrimination outlined in Sections 7.3 and 7.4. To do this, the samples are again reduced to only White umpires, only Hispanic umpires, and only Black umpires, and time trend interaction terms are included to capture the time-variant effects of discrimination. By computing the average marginal effects of UPM and UBM at each year in the sample and then plotting the results, we see the trend of discrimination over time.

Figure 4. Umpire-player time trend race match effects per umpire race: over-recognition



Black-umpire specification not shown for UPM due to high standard errors.

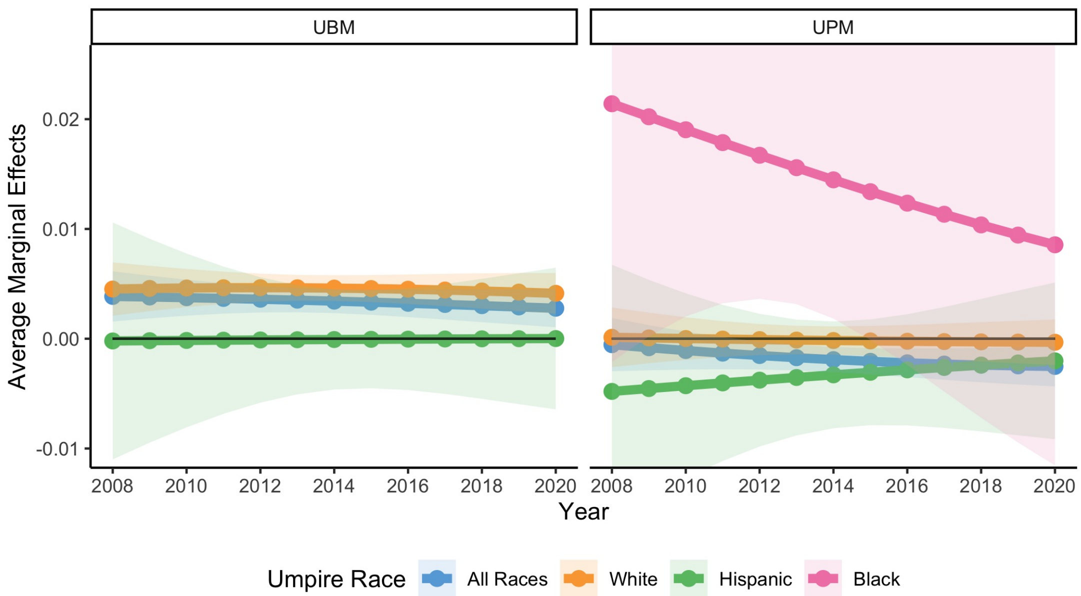
In Figure 4, this analysis is shown for over-recognition. In the line plot for UBM, a few main results stand out. For one, discrimination by White umpires is slightly stronger than in the base specification (including all umpires). This is feasibly somewhat due to the increasing bias exhibited by Hispanic umpires against Hispanic batters, which is also evident in the figure. As hypothesized in Section 7.3, this trend could be due to the recent increase in Hispanic umpires in MLB. The effect of UBM for Black umpires is only statistically significant for a few years in the sample and trends toward zero — providing somewhat limited evidence of widespread batter-level discrimination.

Focusing on the umpire-pitcher match portion of the figure, the results are relatively similar. While the base specification suggests pitcher-level discrimination on pitches outside the zone is falling slightly over time, the effect of UPM for White umpires is slowly rising. Again, this could be attributable to the UPM trend for Hispanic umpires, which rapidly falls throughout the sample — more evidence that the reverse-bias exhibited by Hispanic umpires is likely growing. The UPM effects for Black umpires are not included in this figure, as the standard errors were overwhelmingly large and the effects were statistically insignificant across the sample.

Regarding these breakdowns for under-recognition, we now examine Figure 5. For UBM, only White umpires exhibit significant discrimination towards batters for balls inside the zone, and the size of the effects hold relatively constant over time. Again, Black umpires are not included as their standard errors were well outside the bounds of the figure. As touched on before, the effect of UPM on pitches inside the zone is largely insignificant when breaking down the effects by umpire race, and the time breakdown does not illustrate any significant trends that further explain that result.

As done in Section 7.4, the race-specific yearly effects outlined above can be decomposed further to examine which races are affected the most. While the full

Figure 5. Umpire-player time trend race match effects per umpire race: under-recognition



Black-umpire specification not shown for UBM due to high standard errors.

results are on display below in Figures 6 and 7, a few notable results can be identified. For instance, for over-recognition and White umpires, the bias exhibited against non-White pitchers seems to be slightly stronger against Black pitchers than Hispanic ones, although the effects are converging with time. For the bias against non-White batters, Hispanic players are more likely to be harmed by bad calls than Black players, although again, the effects seem to be converging. Notably, the insignificant effects of UPM for under-recognition and White umpires can be partially explained by the opposite effects for Hispanic and Black pitchers. Black pitchers were significantly harmed early in the sample, with Hispanic pitchers receiving slight benefits. These terms converge over time, and their mirrored effects likely result in the overall effects being insignificant from zero. The trends for Hispanic and Black umpires do not provide any additional insights that further explain the results discussed previously.

Figure 6. Player race-specific time trend effects per umpire race: over-recognition

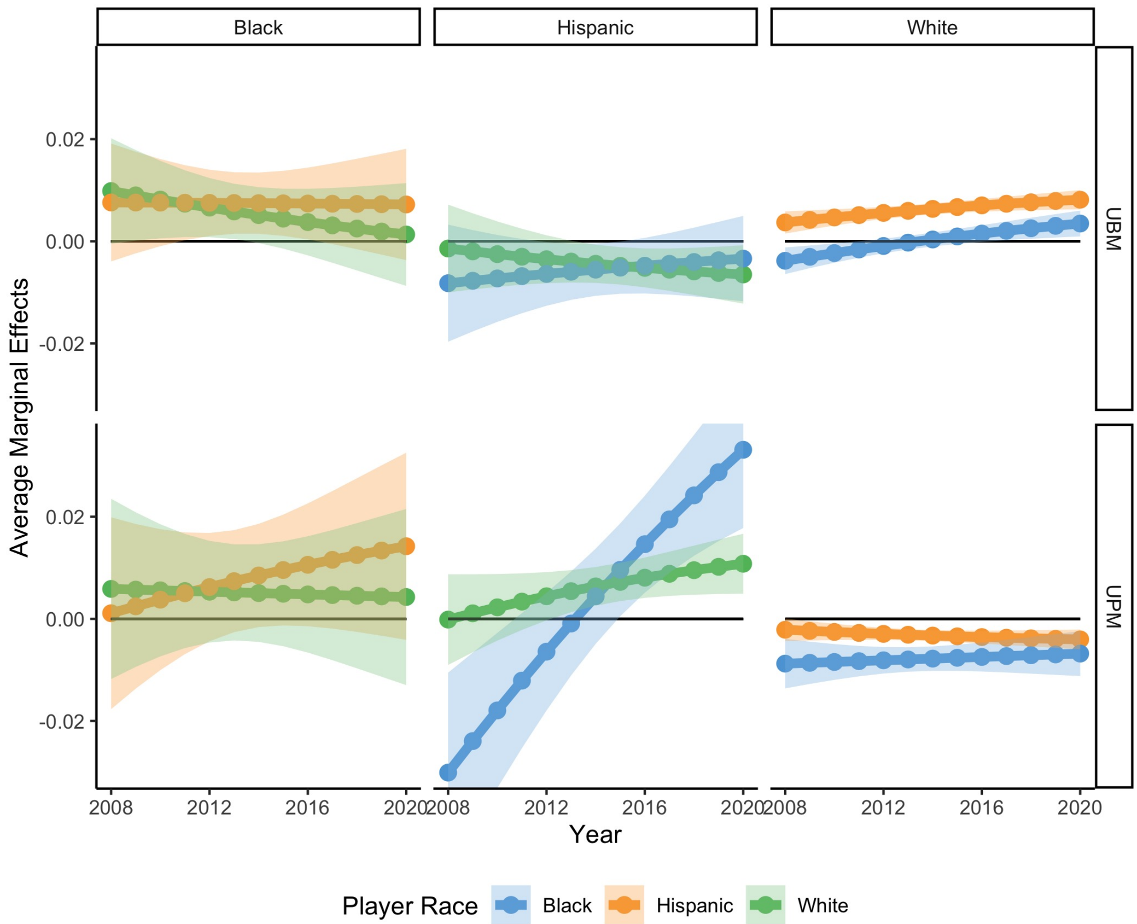
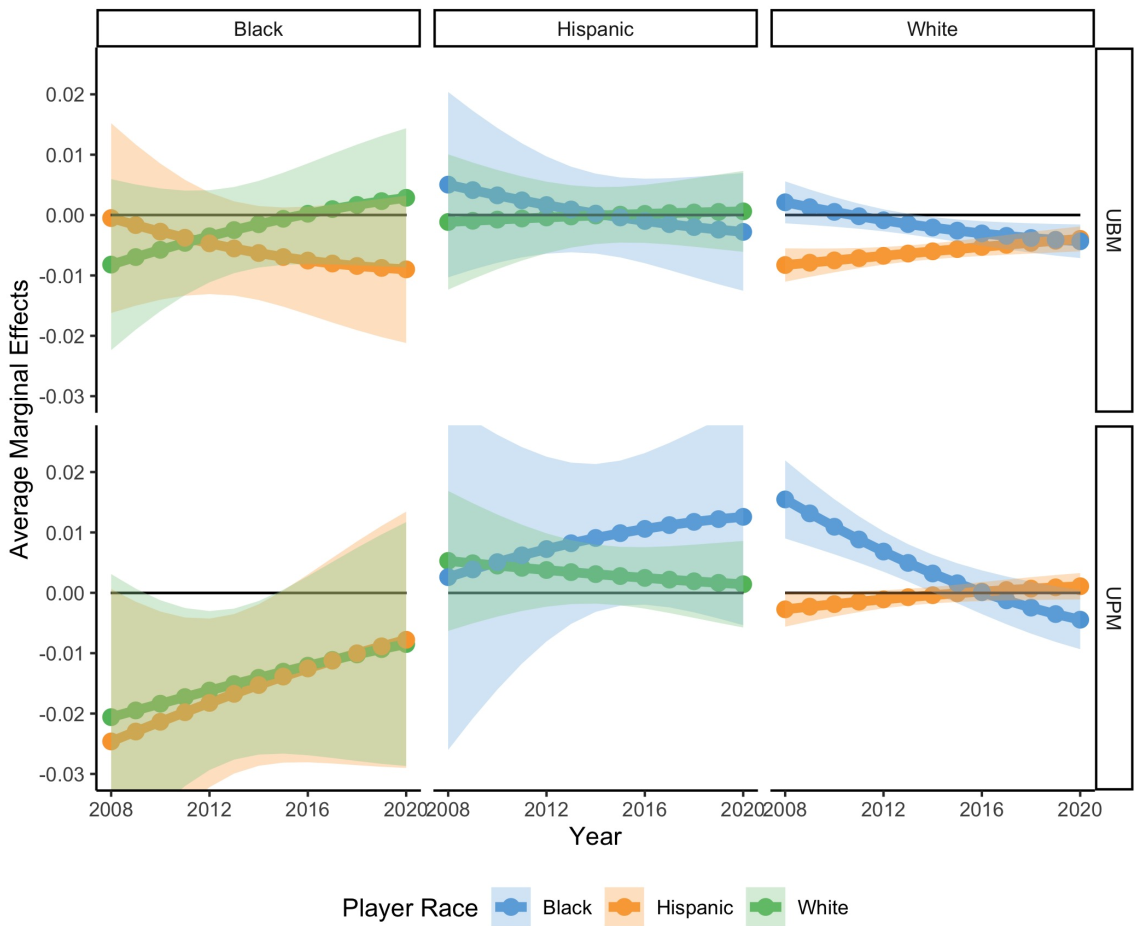


Figure 7. Player race-specific time trend effects per umpire race: under-recognition



8 Limitations

Before offering a discussion of these results, a few limitations of this work are as follows. First, attendance was not accounted for in the modeling, despite past work that shows officials' decisions can be driven by the amount of fans attending a game. This was an intentional omission — although using attendance data would likely help make the models stronger, it would make it impossible to use any pitches from the COVID-altered 2020 season, which was played entirely without fans in the stands. Rather than omit these observations, I chose to not include attendance as a covariate. To ensure this was not markedly changing the race-match effects found in the results, I ran the base specifications from 2008-2019 and included attendance data — no significant differences to the effects of UPM and UBM were found.

Moreover, the potential exists for some of the race categorizations to be incorrect. Determining players' races is the opposite of an exact science — my approach used as many checks as possible to make sure the races were correct, but it is almost inevitable that there are errors. It is unknown how this could affect the results, as it is unknown if/how much/in what direction errors were made. Be that as it may, I do believe my categorization strategy is more thorough and intensive than past efforts, and I am ultimately confident in its accuracy.

9 Conclusion

In summary, this analysis presents evidence that from 2008-2020, race has played a significant role in how umpires call balls and strikes. In the full sample of all years and all umpire races, more favorable calls are made for players whose races match the umpires', all else equal. It is also clear that this discrimination does not diminish over time. While overall pitch-calling quality has improved in the last thirteen seasons, the predictive effects of batters' and pitchers' races have not substantially declined.

Significantly, this study also suggests that White umpires account for a disproportionate amount of said discrimination. When restricting the sample to only White umpires, the average marginal effects of UPM and UBM are larger than in the base specifications in almost all cases, with the exception of UPM and under-recognition. Further, for over-recognition, while the overall trend of discrimination does indeed look to be slightly decreasing over time, the trend for White umpires looks to be slightly *increasing*. This is feasibly somewhat due to the reverse-bias exhibited by Hispanic umpires, most present in the later half of the sample. When considering these findings together with the fact that a vast majority of MLB umpires are White, it is evident White players benefit the most from umpire discrimination in the league.

Admittedly, the observed marginal effects on the probability of a missed call are somewhat small. Be that as it may, there is reason to believe that these effects do actually cause significant real-world impacts. For one, the occurrence of a missed call is somewhat rare in the sample. For pitches in the zone, the umpire misses the call only 22% of the time. For pitches out of the zone, missed calls make up only 16% of all pitches. An increase of .373 percentage points of the likelihood of a missed call (the effect of UPM on over-recognition for White umpires) is a significant amount for such a rare event — this effect accounts for roughly 6,422 additional helpful calls for White pitchers in the sample.

Furthermore, baseball is a game that operates on increasingly thin margins as teams continue to apply a more empirical, Sabermetric approach to roster-construction — the effect that one missed call may have on a player's career can be substantial. While one incorrect called strike may not have a significant impact on the career of Mike Trout, perennial All-Star and consensus best player in baseball, one missed call could have disastrous effects for a player looking to break into the Major Leagues. One strikeout could be the difference between a batter playing another day or getting sent back to the minor leagues. Almost 5% of MLB players in history played in *only one game* [3] (BaseballReference, n.d.; BaseballAlmanac, n.d.) — if an umpire makes a bad call due to a player's race, that decision could potentially make or break a player's career.

In terms of the policy implications of these results, MLB could update their umpire evaluation systems to include race information. The effects of discrimination have clearly not decreased in the last thirteen seasons — perhaps this is due to a lack of internal recognition of the problem by the league. Further, public efforts by websites and fans to evaluate umpires (such as the Twitter account @UmpScorecards)

3. Since 1876, the first year of the National League.

could also focus on how umpires call pitches for same-race and different-race players, as an increased public pressure would also likely cause discrimination to diminish over time.

A more effective approach would be to implement robotic umpires league-wide in order to eliminate pitch-calling biases altogether. MLB has tested electronic strike zones in its lower-level partner Atlantic League, and has long had the technology to completely remove pitch-calling from umpires' responsibilities. While this would represent a marked change from the version of baseball fans have known forever, it may be the only way to ensure players are not victims of racial discrimination at the plate. If a goal of the league is to encourage minority representation, then continuing with human umpires who potentially limit the ability of Black and Hispanic players to succeed would be in direct conflict to that mission.

Bibliography

BaseballAlmanac. n.d. "Every Major League Baseball Player in History." https://www.baseball-almanac.com/players/ballplayer.shtml.

BaseballReference. n.d. "Cup of Coffee Players." https://www.baseball-reference.com/friv/coffee.shtml.

Becker, Gary S, et al. 1971. "The Economics of Discrimination." *University of Chicago Press Economics Books.*

Birnbaum, Phil. 2010. "Is There Racial Bias Among Umpires?" *Fall 2010 Baseball Research Journal,* 65–70.

CloseCallSports. 2015. "MLB Debut of Umpire Carlos Torres," July 15, 2015. https://www.closecallsports.com/2015/07/major-league-debut-of-umpire-carlos.html.

———. 2016. "MLB Debut of Umpire Ramon De Jesus," April 22, 2016. https://www.closecallsports.com/2016/04/major-league-debut-of-umpire-ramon-de.html.

———. 2020a. "MLB Debut of Umpire David Arrieta," August 21, 2020. https://www.closecallsports.com/2020/08/mlb-debut-of-umpire-david-arrieta.html.

———. 2020b. "MLB Debut of Umpire Jose Edwin Moscoso," July 23, 2020. https://www.closecallsports.com/2020/07/mlb-debut-of-edwin-moscoso.html.

———. 2020c. "MLB Debut of Umpire Jose Navas," July 24, 2020. https://www.closecallsports.com/2020/07/mlb-debut-of-umpire-jose-navas.html.

Gabriel, Paul E, Curtis Johnson, and Timothy J Stanton. 1995. "An Examination of Customer Racial Discrimination in the Market for Baseball Memorabilia." *The Journal of Business* 68, no. 2 (April): 215–230. https://doi.org/10.1086/296661. https://ideas.repec.org/a/ucp/jnlbus/v68y1995i2p215-30.html.

Garicano, Luis, Ignacio Palacios-Huerta, and Canice Prendergast. 2005. "Favoritism under social pressure." *Review of Economics and Statistics* 87 (2): 208–216.

Green, Etan. 2014. "Four Strikes and You're Out." *FiveThirtyEight* (April 3, 2014). https://fivethirtyeight.com/features/four-strikes-and-youre-out/.

Gwartney, James, and Charles Haworth. 1974. "Employer costs and discrimination: The case of baseball." *Journal of Political Economy* 82 (4): 873–881.

Irwin, Will. 2004. "Racial Discrimination in Major League Baseball: Can it Exist When Productivity is Crystal Clear?" Bachelor's Thesis, Illinois Wesleyan University.

Kim, Jerry W, and Brayden G King. 2014. "Seeing stars: Matthew effects and status bias in major league baseball umpiring." *Management Science* 60 (11): 2619–2644.

Mills, Brian M. 2014. "Social pressure at the plate: Inequality aversion, status, and mere exposure." *Managerial and Decision Economics* 35 (6): 387–403.

Nardinelli, Clark, and Curtis Simon. 1990. "Customer racial discrimination in the market for memorabilia: The case of baseball." *The Quarterly Journal of Economics* 105 (3): 575–595.

Nevill, Alan M, Sue M Newell, and Sally Gale. 1996. "Factors associated with home advantage in English and Scottish soccer matches." *Journal of sports sciences* 14 (2): 181–186.

Parsons, Christopher A, Johan Sulaeman, Michael C Yates, and Daniel S Hamermesh. 2011. "Strike three: Discrimination, incentives, and evaluation." *American Economic Review* 101 (4): 1410–35.

Pascal, Anthony H., and L. A. Rapping. 1970. *Racial Discrimination in Organized Baseball.* Santa Monica, CA: RAND Corporation.

Price, Joseph, and Justin Wolfers. 2010. "Racial discrimination among NBA referees." *The Quarterly journal of economics* 125 (4): 1859–1887.

Sommers, Paul M, and Noel Quinton. 1982. "Pay and performance in major league baseball: The case of the first family of free agents." *The Journal of Human Resources* 17 (3): 426–436.

Sutter, Matthias, and Martin G Kocher. 2004. "Favoritism of agents–the case of referees' home bias." *Journal of Economic Psychology* 25 (4): 461–469.

Tainsky, Scott, Brian M Mills, and Jason A Winfree. 2015. "Further examination of potential discrimination among MLB umpires." *Journal of Sports Economics* 16 (4): 353–374.

Verducci, Ben. 2017. "Racial Discrimination in Baseball: A Reinvestigation of MLB Umpire Bias." Bachelor's Thesis, Haverford College.

Walsh, John. 2010. "The Compassionate Umpire." *The Hardball Times* (April 7, 2010). https://tht.fangraphs.com/the-compassionate-umpire/.