Estimating Implicit Bias Among WNBA Referees

Simi Edeki, Alex Arnell, and Jo Jiao

Spring 2025

Abstract :

We investigate whether racial bias exists in foul calls made by referees in the WNBA. Using doubly robust augmented inverse probability weighting (AIPTW), we analyze play-by-play, referee assignment, and player demographic data to estimate the causal impact of referee race on foul-calling behavior. Preliminary findings indicate no evidence of same race bias. However, our conclusions are contingent upon the validity of the model and our causal assumptions.

1 Introduction

Over the past 18 months, the WNBA has experienced a surge in popularity, primarily due to the rise in popularity of NCAA women's basketball, led by notable players such as Caitlin Clark, Angel Reese, Paige Bueckers, Haley Van Lith, Juju Watkins, and Cameron Brink, among others. In the 2024 season, the WNBA delivered its most-watched regular season in 24 years, finished with its highest attendance in 22 years, and set records for digital consumption and merchandise sales (WNBA, 2024). With this growing attention has come heightened scrutiny, particularly regarding the quality of officiating. One of the most controversial moments occurred in the decisive game of the 2024 WNBA Finals between the New York Liberty and the Minnesota Lynx. With the Lynx up by two points and only five seconds remaining, officials called Lynx center Alanna Smith for a foul on Liberty star Breanna Stewart, awarding Stewart the free throws needed to tie the game. The Liberty ultimately secured their first-ever championship in overtime. However, the victory was clouded by widespread allegations of poor officiating and potential bias in favor of the Liberty.

Against this backdrop, we set out to investigate what types of bias might exist among WNBA referees. Specifically, given the entrance of high-profile players like Caitlin Clark and her media-framed rival Angel Reese, racial dynamics between players, fans, and the organization itself have become especially charged. This raises the critical question: could the racial composition of refereeing crews influence foul-calling patterns in the WNBA?

To explore this question, we draw inspiration from the work by Price et al., 2010, who investigated racial discrimination among NBA referees. Their study explored whether players received more fouls when officiated by referees of a different race, using detailed player- and game-level data across multiple NBA seasons. They found that players received about 0.12 to 0.21 more fouls per 48 minutes (a 2.5% to 4% increase) when the number of opposite-race referees increased from zero to three. These effects persisted even after controlling for detailed player, referee, and game characteristics.

In this analysis, we shift the focus to the WNBA, a league that lacks detailed research into referee bias. We estimate the average treatment effect of the majority of the officials mismatching a player's race on the number of fouls a player receives. Specifically, we combine player-level foul data with referee crew composition data to construct a binary treatment variable indicating whether the majority of the referees in a game share the player's race. To estimate the causal effect, we apply doubly robust AIPTW methods, controlling for player characteristics and game characteristics. This allows us to isolate the impact of racial match or mismatch between players and officials on foul outcomes, minimizing bias from confounding factors.

2 Background

2.1 Referee Pipeline

WNBA officials are recruited through the combined National Basketball Referees Association recruiting pipeline, which represents officials in the NBA, WNBA, and the NBA G League (Dalzell, 2025). The

NBA Scouting Group selects candidates from national tryouts, the NCAA basketball, FIBA, and other competitive basketball leagues. The top candidates are hired to participate in the NBA G League. G League referees are then evaluated and recommended for hire into the NBA and WNBA. Because the pay is much higher in the NBA, top WNBA referees often move to the NBA. This talent gap between the two leagues leaves WNBA officials more susceptible to implicit bias than their NBA counterparts, potentially leading to a greater racial bias in the WNBA.

2.2 *Referee Assignment*

The 2024 WNBA–NBRA Collective Bargaining Agreement gives the league office complete control over scheduling: every game gets a three-person crew chosen by the league, and assignments are released 14-24 days in advance (SB Nation, 2025). Officials may request "closed dates," but those requests are granted only if enough referees remain available; no rule ties specific refs to specific teams or players.

In practice, the composition of the crew is driven by logistical variables that ensure efficient travel loops. Additionally, referees are not assigned to the same teams for consecutive games to avoid any perception of favoritism or bias. There are no fixed crews, so most referees do not work with the same peers from game to game.

2.3 *Referee Feedback*

Referees receive feedback continuously through the league's internal platform, the Referee Engagement Performance System (REPS). Every WNBA game is uploaded to REPS, where staff tag and evaluate each call. According to VP of Officiating Monty McCutchen, "Each official gets hundreds of feedback points over the season, delivered daily, so improvement areas are flagged in real time rather than in a bi-weekly review" (WNBA, 2024).

3 Data Generation

We constructed a comprehensive player-level dataset from WNBA games that spans four seasons (2021-2024). Our raw data source was 879 unique files containing detailed play-by-play and statistical information for WNBA games from 2021-2024 seasons. From here, we extracted 20,699 player-game observations with 282 distinct players and 879 distinct games and 859 unique referee combinations. All numerical game data (fouls, points, etc.) was converted to a rate based on playtime. Each such rate describes occurrences per 40 minutes, the length of a WNBA game. Since we are working with rate variables, we drop any player-game rows for which a player played less than five minutes. This prevents outliers from players who foul or score points, but are only in the game for a short amount of time. For the generating code and the raw data set, see JoNeedsSleep, 2025.

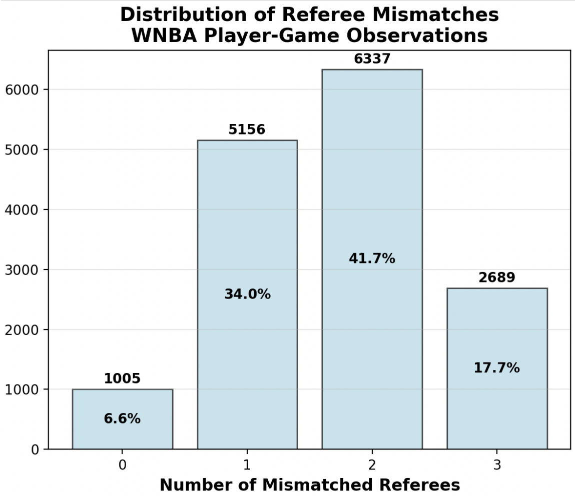


Figure 1. Histogram of race mismatch between player and the three referees.

3.1 *Identifying Player and Referee Referee Races*

The players and officials' races were identified based on their physical appearance and self-description if available. We included an Other category for players and referees were not identified as Black or white. This expands on the categorization used by Price et al., 2010, which coded other races as white. Of the 266 players and 55 referees in our data set, only 11 players and 4 referees were neither Black nor white. While it is necessary to keep these referees for assessing treatment, we drop the 11 players because of overlap concerns; these players have almost no chance of being treated.

3.2 *Treatment as Race Mismatch*

We generated our treatment variable A by comparing the race of the player with the races of the three referees. A score of 0, for example, indicates that the player is of the same race as all the three referees. See figure 1 for a histogram of our scores. We then transformed the categorical data into a binary data with 0 being a mismatch score of 0 and 1 and 1 being a mismatch score of 2 and 3.

4 Methods

4.1 Formal Estimand

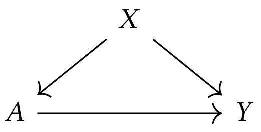
We are interested in estimating the average treatment effect (ATE) of being officiated by a majority non-matching referee crew (one or two out of three referees match the player's race) compared to a majority matching crew (two or three out of three referees match the player's race) on the number of fouls a player receives per game.

Let: $A_i = 1$ if player i is officiated by zero or one matching referees $A_i = 0$ if player i is officiated by two or three matching referees Y_i be the foul rate for player i, measured as fouls per minute played.

We define the ATE as

$$\text{ATE} = \mathbb{E}[Y|\text{do}(A=1)] - \mathbb{E}[Y|\text{do}(A=0)]$$

4.2 Causal DAG



A = Majority racial mismatch
Y = Foul rate of player
X = Other observed variables (team, opponent, player characteristics, player box score)

4.3 Assumptions and Identification

To properly identify this causal estimand, we must satisfy the overlap assumption and no unobserved confounding variables assumption.

The assignment of the referees and thus the treatment is not random and depends on co-variates such as date, home team, away team, etc. Referee mismatch further depends on player race. We include the plausibly confounding variables team, opponent, and race in training our propensity and outcome models.

The overlap assumption states that all groups have a non-zero chance of receiving treatment. The histogram of propensity scores show that this condition is well met. This distribution is roughly bimodal, clustering around 0.64 and 0.45, representing Black athletes and white athletes, respectively. Note that Black athletes are more likely to be treated, as there are fewer Black referees than white referees, despite the majority of WNBA players being Black.

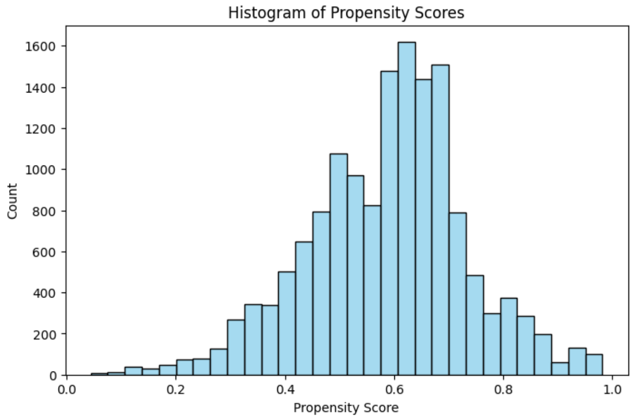


Figure 2. Propensity Scores cluster around 0.64 and 0.45.

4.4 Estimation

We estimate the ATE using the doubly-robust augmented inverse probability of treatment weighted estimator (AIPTW) and a double-machine learning. For the outcome model, we used a gradient boosting regression model to build $\hat{Q}$ to predict Y from X and A. For the propensity score function, we used a gradient boosting classification model to estimate $\hat{g}$ predict A from player race, home team, and away team. Gradient boosting was chosen instead of random forest because it performs better on structured tabular data and usually captures subtle interactions better; it was chosen over low-variance/high-bias models such as Logistic Regression to capture flexible interactions between the dataset, and since we have a relatively large dataset. In addition, we use k-fold cross-fitting to improve the error penalty in AIPTW. Plugging our data and these models into the AIPTW equation, we have:

$$\hat{\tau}^{\text{AIPTW}} \triangleq \frac{1}{n}\sum_{i=1}^{n}\Big[\hat{Q}(1,X_i) - \hat{Q}(0,X_i)$$

$$+ A_i\,\frac{Y_i - \hat{Q}(1,X_i)}{\hat{g}(X_i)} - (1-A_i)\,\frac{Y_i - \hat{Q}(0,X_i)}{1-\hat{g}(X_i)}\Big].$$

5 Results

We found the average treatment effect of referee mismatch to be 0.038 ± 0.108 fouls per 40 minutes. This is not significant and indicates no trend of opposite-race bias. To look at the effect of mismatch on white and Black players separately, we calculated the conditional average treatment effect (CATE) of referee mismatch by conditioning on player race. The CATE of referee mismatch for Black players is 0.086 ± 0.105, and the CATE for white players is -0.104 ± 0.118. Neither of these figures are significant.

5.1 *Sensitivity Analysis*

We seek to verify that our results are robust to unobserved confounding variables through sensitivity analysis. This estimates how large of an effect an unobserved confounder would have to have, in comparison to observed confounders, to meaningfully change our ATE estimation. The Austen plot shows that an unobserved confounder would have to have a stronger effect than player race and position on treatment and outcome to meaningfully change our ATE. This seems unlikely, as race is a strong predictor of treatment and position is a strong predictor of fouls. Thus, our analysis is robust to unobserved confounding factors.

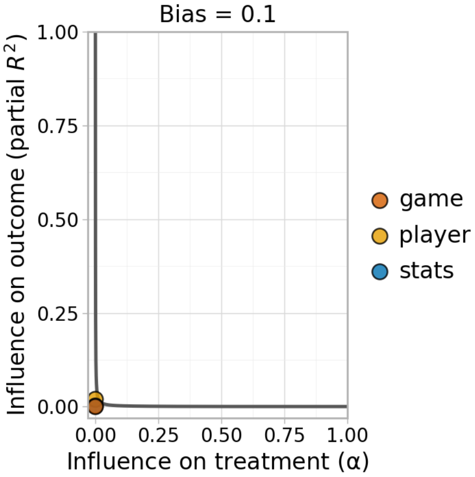


Figure 3. In this Austen plot, the player categorization encompasses player race and position. An unobserved confounder would have to have a stronger effect than this categorization to change our results.

6 Discussion

We found that there was no significant difference between player foul rates when the majority of officials had a matching race. This result is not particularly surprising, considering the volume of feedback WNBA officials receive. We found no statistically detectable effect of majority-race-matched crews on personal-foul rates. This null result is consistent with the WNBA's modern officiating infrastructure, which makes systematic bias harder to sustain.

First, every official is subject to daily video review through the Referee Engagement Performance System (REPS). Clips of every foul or no call are tagged 'correct' or 'incorrect' within 24 hours, and season-long accuracy scores determine playoff assignments and contract renewals. Continuous monitoring reduces the opportunities for unconscious biases to translate into persistent patterns of behavior.

Second, crew assignment is highly mixed and rotational: under the current travel-loop algorithm, fully matched three-official crews are rare and short-lived. Officials, therefore, work with a wide variety of partners and teams, limiting repeated interactions that could reinforce own-race comfort effects.

Third, the WNBA referee pool is nearly gender-balanced, a strong contrast to the NBA, where there were only six female referees in 2022. Some studies have found that women have less implicit bias. Specifically, Assari, 2018 found that male participants exhibited higher levels of implicit anti-Black bias and that white men demonstrated the highest levels of implicit bias against Black individuals. Given this fact, it is feasible that the WNBA data would not find the same significant effect as Price et al., 2010.

7 Conclusion

We set out to test whether WNBA players are whistled differently when most of the officials on the floor share their race. Leveraging four full seasons of merged play-by-play, roster, and referee-roster data (20,699 player-games) and a doubly-robust AIPTW estimator with cross-fitted gradient-boosting nuisance models, we obtain an average treatment effect of 0.038 fouls per 40 minutes (95 percent CI: -0.070 to 0.178). The null finding is credible in light of the league's modern officiating ecosystem: continuous video grading through the REPS platform, centrally controlled referee assignments, and a nearly gender-balanced referee pool all act to suppress systematic own-race bias.

7.1 Limitations

Without whistle-level ref IDs, we cannot probe individual-referee heterogeneity. Obtaining the internal NBA/WNBA "precision timing" feed would enable referee-specific causal analysis. Additionally, the binary treatment (majority racial match) oversimplifies nuanced interactions—future work should test continuous or stratified specifications (e.g., zero vs. three same-race referees). Finally, we lack referee-level covariates (e.g., experience, gender-specific effects) that might moderate bias. These limitations highlight trade-offs between causal identification and real-world complexity but do not invalidate our core null finding.

Acknowledgements

This paper is written as part of Prof. Victor Veitch's STAT 27420 Introduction to Causality with Machine Learning class.

References

Assari, Shervin (2018). *Interaction Between Race and Gender and Effect on Implicit Racial Bias Against Blacks.* DOI: 10.15171/ijer.2018.10.

Dalzell, Noa (Jan. 3, 2025). *How WNBA referees are scouted, trained, and held accountable, explained.* URL: https://www.sbnation.com/2025/1/3/24330571/wnba-officiating-referees-womens-basketball-cheryl-reeve-liberty-lynx-caitlin-clark-monty-mccutchen (visited on 05/30/2025).

JoNeedsSleep (2025). *causal_inference_wnba.* GitHub repository. URL: https://github.com/JoNeedsSleep/causal_inference_wnba (visited on 05/30/2025).

Price, Joseph and Justin Wolfers (2010). "Racial Discrimination Among NBA Referees". In: *The Quarterly Journal of Economics* 125.4, pp. 1859–1887. DOI: 10.1162/qjec.2010.125.4.1859.

SB Nation (2025). *WNBA Officiating Referees in Women's Basketball.* URL: https://www.sbnation.com/2025/1/3/24330571/wnba-officiating-referees-womens-basketball-cheryl-reeve-liberty-lynx-caitlin-clark-monty-mccutchen (visited on 05/30/2025).

WNBA (2024). *WNBA Delivers Record-Setting 2024 Season.* URL: https://www.wnba.com/news/wnba-delivers-record-setting-2024-season (visited on 05/30/2025).