# Are NBA Referees Biased Toward Certain Players?

Shari Tian, Conner Byrd, Mounika Adepu, and Nagaprasad Rudrapatna

November 1, 2022

## Introduction

### Background

A big question of concern for NBA fans is whether referee calls are biased or not. While the NBA has publicly released the data of all referee calls since the second half of 2015-2016 NBA season, they have not published their analyses of these data. This case study conducts analyses of the official NBA Officiating Last Two Minute (L2M) Reports, which are released after each game. These reports are a post-game assessment of officiating events that occurred in the last two minutes of each NBA game that were at or within 6 points at any point during this time period. Each officiating event is categorized as one of four call-types: CC (correct call), CNC (correct non-call), IC (incorrect call), and INC (incorrect non-call). CCs and CNCs are officiating events that were handled correctly by the referees as a call or non-call respectively. INCs are officiating events that were not called by the referee but were actually violations. Likewise, ICs are officiating events that were called by the referee but were actually not violations. Cleaned by `atlhawksfanatic` and maintained on a public Git repository, the raw L2M dataset includes 63,283 observations and 42 columns (variables) where each row is a unique call. This dataset includes variables, such as the specific call or non-call type, whether or not the referee decision is correct upon post-game review, the names of committing and disadvantaged players, time stamp, and other game details. This manuscript will use a modified version of the L2M dataset.

### Motivation

Referee bias towards a player can pose a very dramatic butterfly effect on future games. For instance, if a referee consistently favors a player by not calling a foul on that player in the last few minutes of a game, it may provide that player's team an advantage and allow that team to unfairly win the game. One or multiple unfair wins would allow a team to have a more successful overall NBA season and might even impact who makes it to the playoffs and wins an NBA championship. In addition, more success in an NBA season is often associated with increased ticket and merchandise sales, better player trade deals, and other financial incentives. Given this context, we are interested in exploring whether referee calls are biased in favor of certain players. Based on our analysis, we hope to provide appropriate suggestions on how bias in referee calls (if any) can be improved to ensure the fairness of each NBA season, and by extension, the league.

To start, this manuscript takes inspiration from Russell Goldenberg's "NBA Last Two Minute Report" article in which he presents various summary statistics and visualizations for the L2M dataset. Specifically, Goldenberg created a double-sided bar graph to show how certain calls favor either the committing or disadvantaged player in the L2M dataset. While this visualization was insightful in showing which players had the highest total favors overall, conclusive statistical statements could not be made on this relationship because the visualization simply illustrated summary statistics (counts) related to player favors. Building on Goldenberg's previous work, this manuscript begins by exploring multivariate relationships between favors towards committing players and factors related to a referee call (names of referees, number of minutes that a committing player played prior to the last two minutes of a game, type of call, game attendance, whether the call occurred within the final minute of the game, and whether the game the call was made was during a playoff game). Our inferential analysis examines the accuracy and potential bias of referee calls involving specific players. Given the wide variation in player prestige and experience, we suspect that there may be different levels of referee bias toward different players.

### Data Cleaning

Given that our analysis revolves around potential referee bias toward specific players, we filtered out observations with missing values for the variable that marks the post-game judgment of the referee call. We also created an indicator variable for whether a call favored the committing player. A call is considered to favor the committing player if the decision is an INC as shown in Goldenberg's analysis. This variable will serve as the response variable in our analysis. We had initially planned to also explore whether a call favored the disadvantaged player (i.e., decision is an IC); however, since less than one percent of calls were favorable to disadvantaged players, we decided to limit our analysis to favors for committing players. We also created an indicator variable for whether or not the call was made within the final minute of the game. This was because, although we acknowledged that accounting for the timing of the call was important (i.e., awarding the opposition a pair of free throws with five seconds left is much more important than if this decision was made a minute earlier), we felt that the existing quantitative time variables were inadequate for building an easily interpretable model. Due to the data containing only the last two minutes of regulation but all five minutes of overtime for games that went to overtime, we chose to call incidents that did not happen in the final minute of either "not final minute" rather than something more descriptive.

Next, we excluded the following variables from the dataset: the quarter when the call was made, the NBA season, all remaining time variables, comments on the play, game details, all variables containing webscraping information, all variables containing dates, all variables containing scores, all variables containing game ID information, and all team-based variables. These variables were omitted because they were either redundant or irrelevant to our modeling approach. More in-depth justification for removing these variables is discussed in the Data Cleaning section of the Appendix. We then mean-imputed the missing observations for continuous variables–the number of minutes played by the committing player (1,792 values imputed) and the attendance of a game (5,567 values imputed)–and filtered out the missing observations for categorical variables. We selected this procedure to handle the missingness because we felt that the data may be missing at random (MAR). We think information about attendance and committing player minutes could be missing due to errors during data entry (when the raw L2M dataset was created) or improper data collection at certain games/stadiums.

Our decisions to explore certain variables were largely based on our domain knowledge about the NBA. We considered the attendance and playoff status of games because we believe that the size of the crowd (which correlates to the intensity of crowd noise) and the objective importance of the game (i.e., playoff games are more important) could be associated with a referee's accuracy and potential bias on calls, particularly near the end of close games. We also suspect that the number of minutes played by the player judged as committing a violation before "crunch time" (the last two minutes) may influence a referee's decision at the end of tight games. It is well-known that top players play more minutes before "crunch time", and we think that referee bias in calls may be associated with a committing player's prestige. Thus, as a proxy for player prestige (since we do not have access to this information), we consider committing player minutes. Moreover, we considered the type of violation (call) because it is well-known that NBA referees are more likely to misjudge certain incidents (e.g., charges versus legal blocks). To make the call types easier to interpret, we changed the variable representing a standardized classification of referee call types to group all violations that were not shooting, offensive, personal, or loose-ball fouls into an "other" category. This decision is motivated by the fact that shooting, offensive, personal, and loose-ball fouls are the four most common call types in the dataset and can more easily be pinned as advantaging a specific player (and his team). Lastly, we wanted to control for the effects of different referee trios. To do so, we first performed simple data wrangling to extract a list of unique referee names based on the given referee variables. We then transformed each unique name into an indicator variable that identified whether a particular referee was on duty during the game when a call was made. As a result, every play (observation) had exactly three 1's across all the referee dummy variables.

### Exploratory Data Analysis

To investigate our suspicion that there may be referee bias toward certain players, we first examined the distribution of favors for committing players (i.e., distribution of the response variable). We found that 4,030 (6.8%) calls favored the committing player (see Table 2). This clearly indicates an imbalance in the data (see Tables 2 and 3). We then determined the total number of committing favors for the 25 players with the most calls as the committing player. We segmented this by whether or not the call occurred in the final minute of the game and by the type of violation.

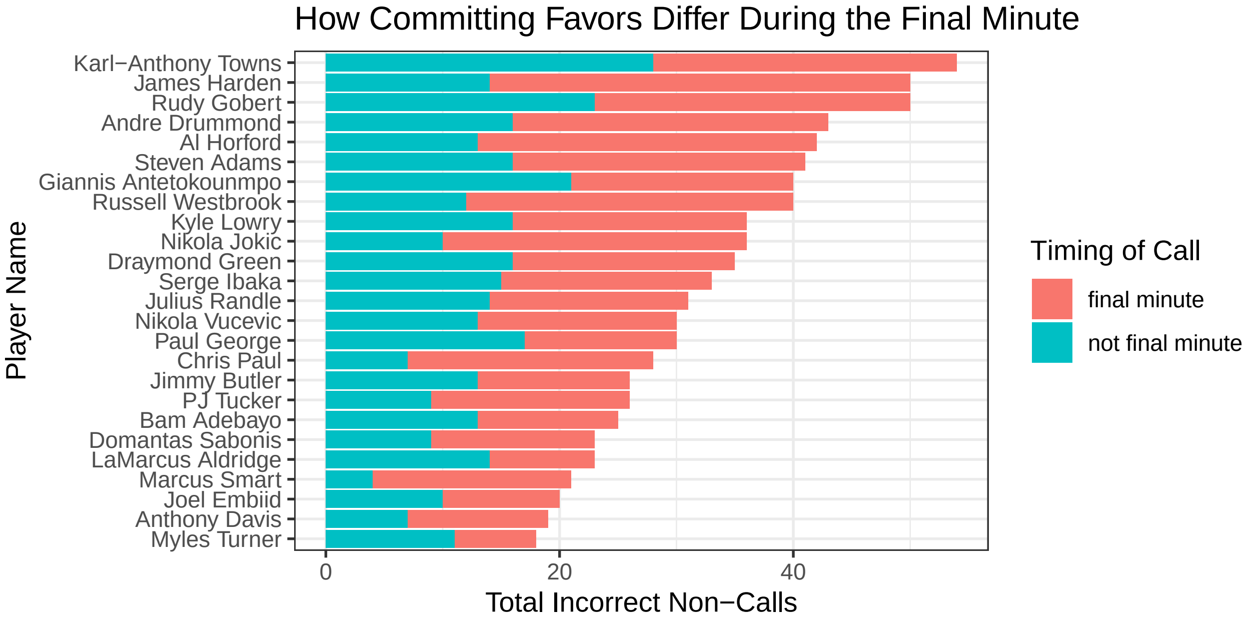

Figure 1: Favors for Committing Players Before and During the Final Minute

Figure 1 illustrates the relationship between whether the call was made in the final minute or not and the total number of INCs for the ten players with the highest ratio of committing favors per minute played. It is interesting to see how the proportion of calls made in the final minute changes across players. For instance, James Harden has a higher proportion of calls made in the final minute than Karl-Anthony Towns. It is important to note that the relationships depicted above and described in this section (including Figure 2) are conditional on the ten players with the most committing favors as they are the only ones represented.

Figure 2 illustrates the relationship between the type of call that was made and the total number of INCs for the ten players with the highest ratios of committing favors per minute played. Collectively, the visualizations suggest a relationship between the type and timing, respectively, of a call and the number of committing favors a particular player receives. Thus, the models we consider in the next section will include the final minute indicator and type of violation (as well as attendance, committing player minutes, and playoff status) as explanatory variables. Given our interest in building a player-level model, we consider the 25 players with the most violations as the committing player. We chose the top 25 players because this approximately accounts for the top 5% of players in the league (i.e., there are roughly 450 players in the NBA in any given season, assuming 30 teams and 15-man rosters). We decided to select the players based on the number of violations rather than the proportion of violations because we believe each player (in the top 25) has sufficient incidents (i.e., the sample sizes for each player are sufficiently large).

### Research Questions

This manuscript aims to answer two questions. Firstly, how do the odds of receiving a favorable call during “crunch time” differ between players (considering the 25 most frequent committing players) when they commit a violation, after accounting for the effects of referees (specifically those who officiated at least 100 plays in the last two minutes), attendance, the number of minutes that the player played before the last two minutes, the playoff status of the game, and whether the call was made in the final minute? Secondly, after accounting for the effects of referees (those who officiated at least 100 plays in “crunch time”) and the 25 most frequent committing players, how are player and game-level characteristics (number of minutes a committing player has played, type of violation, whether a call was made in the final minute, attendance in the stadium, and whether the call was made during a playoff game) associated with the odds of receiving a favorable call in "crunch time"? To address our research questions, we decided to fit a binary logistic regression model. The response variable was whether a given decision (made by the game's referee trio) favored the committing player or not.

## Methodology

Before building models, we performed data cleaning again. Following our research questions, we created a categorical variable with 25 levels identifying each player of interest by name. To do this, we first found the players with the most violations committed and then filtered the data so that the categorical variable indicating the name of the committing player only contains the names of the 25 players with the most violations. We assigned Karl-Anthony Towns to the baseline category because he has the most favorable calls as the committing player (see Figures 1 and 2). We also decided to restrict our analysis to plays in games that attracted crowds of more than a certain number of people. We selected a threshold of 10,000 people after examining the distribution of attendance across the games in which one of the 25 players of interest was characterized as the offender (see Figure 3). The motivation for this decision was excluding abnormally low attendance games (which were relatively rare and likely due to social distancing measures or reduced ticket availability during the COVID pandemic), as we noticed that they violated the linearity assumption in our empirical logit plots. Furthermore, we decided to only account for referees who officiated at least 100 plays (in which one of the 25 players of interest were characterized as the offender) during the last two minutes of games in our model. We selected this threshold because we thought that 100 plays (assuming 10 plays per game, this represents 10 games of experience with these players of interest) would be a sufficiently large sample to accurately assess a referee's decision patterns in "crunch time". Thus, out of a potential 94 referees (who officiated at least one play involving a committing player of interest), we decided to control for the effect of 76 referees. After completing the second stage of data cleaning, we created the final dataset used to fit our models. This final dataset includes 9,809 observations and 83 columns (variables), namely whether the call favored the committing player, the number of minutes a committing player was on the court prior to "crunch time", the name of the committing player, attendance, whether the game occurred during the playoffs, whether each referee of interest officiated the play, and whether the call was made in the final minute. One important feature of this dataset (a by-product of how the 25 players were chosen) is that there are sufficient data points (plays) for each player of interest.

We built our initial binary logistic regression model exclusively based on our knowledge of the NBA; in the previous section, we justify the inclusion of each explanatory variable. We believe a binary logistic regression model is appropriate for this analysis for multiple reasons. Firstly, this modeling technique is suitable for binary response variables such as whether the call was favoring the committing player. Additionally, since our research questions are inferential in nature, fitting a simpler model is preferable for better interpretations. As mentioned in the previous section, we decided to recode the variable indicating the type of violation associated with each play. Collapsing this variable improved not only the interpretability of our models but also decreased the variance inflation factor (VIF) associated with this variable. We also conducted an analysis with VIFs and found that all VIFs were below 5. Due to insufficient evidence of high multicollinearity, we proceeded with the analysis.

We next considered adding additional complexity to the model in the form of interaction terms. To comply with our objective of developing interpretable models, we limited our exploration to interaction effects between categorical and quantitative explanatory variables and obeyed the hierarchical principle, which states that the main effects included in an interaction term must also appear separately in the model. In particular, we explored an interaction between attendance at a game and whether the game occurred in the playoffs and another between the number of minutes a committing player was on the court prior to "crunch time" and whether the game occurred in the playoffs. We thought the conditional relationship between crowd size and referee decisions (and hence player favor) might differ between playoff and regular season games. For instance, larger crowds (louder crowd noise) may have a greater association with player favor in regular season games since referees are usually more lenient in low-stakes games. Furthermore, we thought the conditional relationship between court time and referee decisions (and hence player favor) might differ between playoff and regular season games. For instance, playing more minutes (especially in the playoffs, teams rely on the lengthy contributions of their best players) may have a greater association with player favor in playoff games since referees often hesitate to eject top players during the playoffs. To further investigate these interactions, we created multivariate exploratory data analysis plots (see Figures 4 and 5). Based on Figure 4, there is insufficient evidence of an interaction between attendance at a game and whether the game occurred in the playoffs (little to no difference between the means of each level of playoff status). On the other hand, Figure 5 provides better visual evidence of an interaction between the number of minutes a committing player was on the court prior to "crunch time" and whether the game occurred in the playoffs. Therefore, we decided to add the interaction between committing player minutes and playoff status to the model. We, however, did not include the interaction between game attendance and playoff status (which was not supported by EDA) because we wanted to keep the model as small as possible (as our research questions are inference-based). With the addition of this interaction term, we obtained the final model. Although we noticed that several terms were not statistically significant (based on their respective p-values) based on an alpha level of 0.05, we decided to keep all terms in the final model since we were confident that NBA enthusiasts would be interested in understanding the associations between these variables and player favor (of course, we kept all referee dummy variables in the model, regardless of significance, since we wanted to control for all referees with sufficient "crunch time" experience). For this final model, we found that all of the assumptions for logistic regression were reasonably satisfied (see Assessing Model Assumptions in the Appendix). Furthermore, based on the plot of the binned deviance residuals, we suspect that our final model may be misspecified (see Model Diagnostics in the Appendix).

## Results

Using mathematical notation, we can write the final model as

$$\begin{aligned}\log(\frac{\hat{p}_j}{1-\hat{p}_j}) &= \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + \hat{\beta}_3 x_{3j} + \hat{\beta}_4 x_{4j} + \hat{\beta}_5 x_{5j} \\ &\quad + \hat{\beta}_6 x_{6j} + \hat{\beta}_7 x_{7j} + \hat{\beta}_8 x_{8j} + \hat{\beta}_9 x_{2j} x_{7j} + \hat{\beta}_{\text{player}_i} x_{ij} + \hat{\beta}_{\text{referee}_k} x_{kj}\end{aligned}$$

where $j$ is the index for the call, $i$ is the index for the committing player (ranging from 1 to 24; Andre Drummond to Steven Adams), $k$ is the index for the referee (ranging from 1 to 76; James Capers Jr. to Gary Zielinski; each referee that was included officiated at least 100 plays), $\hat{p}_j$ is the predicted probability of the $j$th call favoring a committing player, $\hat{\beta}_{\text{player}_i}$ is the estimated regression coefficient for the $i$th committing player, and $\hat{\beta}_{\text{referee}_k}$ is the estimated regression coefficient for the $k$th referee. In terms of explanatory variables, $x_{1j}$ is the attendance of the game in which the $j$th call occurred, $x_{2j}$ is the number of minutes that the committing player played prior to the last two minutes of the game in which the $j$th call occurred, $x_{3j}$ through $x_{6j}$ indicate the type of violation associated with the $j$th call, $x_{7j}$ indicates whether the $j$th call occurred in a playoff game, and $x_{8j}$ indicates whether the $j$th call occurred in the final minute of a game. $\hat{\beta}_9$ is the estimated regression coefficient for the interaction effect between $x_{2j}$ and $x_{7j}$.

The regression coefficients were estimated using the `glm()` function. Table 1 displays exponentiated coefficient estimates, as well as the corresponding 95% confidence intervals and p-values (in other words, to obtain the $\hat{\beta}$'s for the equation written above, we would take the log of the values in Table 1).

Table 1: Exponentiated Final Model Coefficients and 95% Confidence Intervals

| | Coefficient | Lower Bound | Upper Bound | P-Value |
|---|---|---|---|---|
| (Intercept) | 0.26949 | 0.07203 | 1.00829 | 0.0515 |
| Attendance | 0.99997 | 0.99992 | 1.00001 | 0.1592 |
| Committing Player Minutes | 1.01559 | 0.99850 | 1.03297 | 0.0741 |
| Type of ViolationShooting | 0.18484 | 0.14660 | 0.23305 | <0.01 |
| Type of ViolationPersonal | 0.13549 | 0.10544 | 0.17409 | <0.01 |
| Type of ViolationLoose Ball | 0.34314 | 0.26714 | 0.44078 | <0.01 |

|  |  | Lower | Upper |  |
|---|---|---|---|---|
|  | Coefficient | Bound | Bound | P-Value |
| Type of ViolationOffensive | 0.20749 | 0.16732 | 0.25730 | <0.01 |
| PlayoffDuring Playoffs | 0.42945 | 0.06786 | 2.71758 | 0.3693 |
| During Final Minutenot final minute | 1.00336 | 0.85905 | 1.17191 | 0.9663 |
| Committing PlayerJames Harden | 1.10510 | 0.70786 | 1.72525 | 0.6602 |
| Committing PlayerRudy Gobert | 0.99884 | 0.64144 | 1.55539 | 0.996 |
| Committing PlayerAndre Drummond | 1.20065 | 0.75302 | 1.91435 | 0.4424 |
| Committing PlayerAl Horford | 1.09717 | 0.69239 | 1.73859 | 0.693 |
| Committing PlayerSteven Adams | 0.63423 | 0.39607 | 1.01560 | 0.0581 |
| Committing PlayerGiannis Antetokounmpo | 0.74115 | 0.45987 | 1.19447 | 0.2187 |
| Committing PlayerRussell Westbrook | 1.05786 | 0.65891 | 1.69836 | 0.8159 |
| Committing PlayerKyle Lowry | 0.97598 | 0.60185 | 1.58266 | 0.9215 |
| Committing PlayerNikola Jokic | 0.57025 | 0.35458 | 0.91711 | 0.0206 |
| Committing PlayerDraymond Green | 0.73500 | 0.44666 | 1.20947 | 0.2257 |
| Committing PlayerSerge Ibaka | 0.94018 | 0.57535 | 1.53635 | 0.8056 |
| Committing PlayerJulius Randle | 0.67283 | 0.39450 | 1.14752 | 0.1458 |
| Committing PlayerNikola Vucevic | 0.73470 | 0.44672 | 1.20832 | 0.2246 |
| Committing PlayerPaul George | 0.85693 | 0.51675 | 1.42105 | 0.5497 |
| Committing PlayerChris Paul | 0.91238 | 0.54294 | 1.53320 | 0.7292 |
| Committing PlayerJimmy Butler | 0.79420 | 0.46551 | 1.35499 | 0.3979 |
| Committing PlayerPJ Tucker | 0.82614 | 0.49110 | 1.38974 | 0.4717 |
| Committing PlayerBam Adebayo | 0.75227 | 0.44406 | 1.27439 | 0.2899 |
| Committing PlayerDomantas Sabonis | 0.73521 | 0.42408 | 1.27460 | 0.2733 |
| Committing PlayerLaMarcus Aldridge | 0.56453 | 0.32925 | 0.96795 | 0.0377 |
| Committing PlayerMarcus Smart | 0.61851 | 0.33857 | 1.12992 | 0.1182 |
| Committing PlayerJoel Embiid | 0.48063 | 0.26983 | 0.85613 | 0.0129 |
| Committing PlayerAnthony Davis | 0.53045 | 0.30039 | 0.93671 | 0.0289 |
| Committing PlayerMyles Turner | 0.50622 | 0.27848 | 0.92023 | 0.0256 |
| Committing Player Minutes:PlayoffDuring Playoffs | 1.01370 | 0.96782 | 1.06175 | 0.5648 |

∗ 5 digits are displayed to properly show coefficients (e.g., game attendance)

∗∗ The estimated coefficients for each referee dummy variable are not shown for brevity (these coefficients will not be interpreted anyway)

All player coefficients in the model should be interpreted after having accounted for all referees who officiated at least 100 plays, game attendance, type of violation, number of minutes the committing player has played, whether the call occurred during a playoff game, and whether the call occurred in the final minute of the game. The committing player levels are ordered in terms of decreasing numbers of favorable calls, with Karl-Anthony Towns as the baseline category (see Figures 1 and 2). Since Towns has the most committing favors, we expect to see the other players having comparatively lower odds of being favored during “crunch time”, which enables easier and more interesting interpretations of the model output. Compared to Towns, there seem to be several committing players who show a statistically significant change in odds of receiving a favorable call at the 5% significance level (chosen by convention), after accounting for all the other factors listed previously. The players with significant coefficients are Nikola Jokic, LaMarcus Aldridge, Joel Embiid, Anthony Davis, and Myles Turner, for whom Table 1 shows evidence of having lower odds of receiving a favorable call during “crunch time” compared to Karl-Anthony Towns, after accounting for all other factors. For these players, we have sufficient evidence to reject the null hypothesis that there is no difference in the odds of being favored as the committing player between them and Towns, after accounting for all other factors. The odds that Nikola Jokic receives a favorable call compared to Towns are expected to be multiplied by 0.57025, after accounting for all other factors. The confidence interval for this exponentiated coefficient estimate means that we are 95% confident that the odds of Jokic being favored as the committing player compared to Towns are multiplied by a factor between 0.35458 and 0.91711, after accounting for all other factors. The interpretations for the other significant players can be found in the Model Interpretations section of the Appendix.

To determine how general player and game characteristics are associated with a favorable call for the committing player, we look at the regression coefficients corresponding to the number of minutes a committing player has played, the type of violation, whether a call was made in the final minute, game attendance, and whether the call was made during a playoff game. These regression coefficients should be interpreted after having accounted for all referees who officiated at least 100 plays, the 25 most frequent committing players, and all other player and game characteristics in the model. It appears that for every additional 100 attendees, the odds of receiving a favorable call are expected to be multiplied by 0.99700, after accounting for all other factors. The distribution of the attendance variable after filtering for attendance greater than 1000 showed that increments of 100 were appropriate for the interpretation of this regression coefficient. Furthermore, in regular season games, for every increase in five minutes played by the committing player prior to "crunch time", the odds of receiving a favorable call are expected to be multiplied by 1.08042, after accounting for all other factors. In playoff games, for every increase in five minutes played by the committing player prior to "crunch time", the odds of receiving a favorable call are expected to be multiplied by 1.15650, after accounting for all other factors. We interpreted the regression coefficient for minutes played by the committing play in increments of five due to the fact that differences in committing minutes tend to be minimal across players in one game. The odds of the committing player receiving a favorable call are expected to be multiplied by 1.00336 for calls not in the final minute compared to those made within the final minute, holding all other predictors constant. It is important to note that none of these regression coefficients are statistically significant at the 5% level, but they are still informative to interpret in order to address our explanatory modeling objective. However, the coefficients for all levels of the type of violation variable (baseline category: "other") are significant at the 5% level. This shows sufficient evidence to reject the null hypothesis that there is no difference in the odds of favoring the committing player during "crunch time" between each type of specific foul and "other" type of violation, after accounting for all other factors. Recall that the "other" type of violation includes all violations that are not shooting, personal, offensive, or loose-ball fouls. The odds that a committing player is favored are expected to be multiplied by 0.18484 if the call is a shooting violation compared to an "other" violation, after accounting for all other factors. The interpretations for the remaining types of violations can be found in the Model Interpretations section of the Appendix.

## Discussion

After accounting for all other variables in the final model, we notice that all non-baseline violation types (shooting, personal, loose ball, and offensive) appear to be significant predictors of player favor (at the 5% level). Compared to the miscellaneous violations categorized as "other", shooting, personal, loose-ball, and offensive fouls all tend to be quick, spur-of-the-moment fouls. This implies that there could be a correlation between the way NBA players commit these particular fouls and whether the referee actually penalizes them for it. In addition, we find that several referees are significant predictors of player favor at the 5% level. Although referee-specific information that can provide more insight as to why certain referees tend to be better predictors of player favor is difficult to find, it is reasonable to assume that differences in years of officiating experience, whether the referee played in the NBA or not, or refereeing style in general could account for some of this significance. In our model, we also find that the committing player being Nikola Jokic, LaMarcus Aldridge, Joel Embiid, Anthony Davis, or Myles Turner are all significant predictors of player favor compared to Karl-Anthony Towns at the 5% level. While this does not mean these specific players are being unfairly biased towards or against (not a causal statement), it could highlight certain aspects of their game or play style which result in more calls working in their favor. All five of these players were significant in the "unfair" direction (received a substantial amount of favorable calls less than Towns). It is also of note that all five of these players are well above average in height (all 6' 10" and above — the NBA mean is 6' 6"), weight (all 250 pounds and above — the NBA mean is 217 pounds), and wingspan (all 7' 3" and above — the NBA mean is 6' 10"). Combining this knowledge with the fact that specific violations mentioned above are also significant, it is possible that due to the increased frame of these players, they may have slower reactions during spur-of-the-moment or "crunch time" plays and could commit fouls more frequently in these kinds of incidents. Further analysis of the videos of officiating events in which these incorrect missed calls occurred could provide more insight as to why these particular types of violations tend to be problematic for the significant players. Also of note is the fact that the exponentiated coefficient estimate for attendance is very close to 1. In the context of our model, this means that the attendance of an NBA game has very little impact on the player favor of calls, accounting for all other variables in the model — a conclusion that subverted our initial suspicions based on the near horizontal pattern in the empirical logit plot for this numerical variable (see Figure 7).

### Limitations and Summary

When evaluating player favors within the data, the largest limitation was the number of possible confounding variables that were either too tedious to obtain or too difficult to calculate (absent from the L2M dataset). Examples of these confounding variables that could theoretically bias a referee's decision towards a specific player include biological statistics such as height, weight, and wingspan (see Discussion), the seniority of players (how many years they have played in the league), the prestige or reputation of players, the stadiums each team plays in, the time of day each game occurred, how long each game lasted, and many more. Additionally, missing values within the data presented several issues. We assumed the data was missing at random (MAR), which made mean imputing values for the quantitative variables an acceptable approach. We chose to mean impute values for the quantitative variables crucial to our analysis — committing player minutes and game attendance. One of the dangers of this approach is that prescribing incorrect (not representative) values to certain players may negatively impact results. For example, mean imputing committing player minutes and then performing an analysis on the 25 most frequent committing players (who might be more likely to play above-average minutes) may have skewed certain values in ways we were not able to account for. Also, if the missingness were due to another underlying mechanism, then our approach probably would not have been entirely appropriate. A potential ramification would be that our final model would include biased coefficient estimates and standard errors (and hence incorrect p-values, which can lead to incorrect assessments of variable significance).

Our analysis was hindered by several limitations that impacted our ability to adequately answer the research questions. We initially wanted to fit a hierarchical model, which in theory should have given us the most accurate results, but we ultimately abandoned this approach due to our group's general lack of knowledge about mixed models. More concretely, this alternative modeling approach would have involved fitting a 3-level logistic regression model (individual play, game, and season levels) and would have removed the assumption of independence. Since our final model only considers the 25 players with the most violations committed in "crunch time", we are only able to generalize our conclusions for these players in future seasons (assuming independence of referee decisions across seasons). We cannot generalize our conclusions to other NBA players, which is a major limitation of the utility of the model. One of the reasons why we decided to limit our analysis to the top 25 players was to reduce the number of coefficients in the model (there are approximately 900 unique players in the database) while keeping the sample size of the dataset reasonable. In hindsight, it may have been worthwhile to find another solution that incorporated more players and was more indicative of the full L2M dataset. The same can be said for the referees we selected. With the referees, we chose to consider only those referees who had officiated at least 100 plays. Perhaps it would have been better to select our players and referees using a more generalizable sampling scheme, such as stratified random sampling. Another limitation arose due to our decision to consider playoff games and games that went to overtime. Including plays from these games could bias player favor toward those who played more playoff or overtime games respectively. Another limitation of our data is that it is imbalanced, as referee decisions which favored the committing player were much less frequent than ones that did not. This meant that we were working with a smaller dataset than we would have liked, which could have affected the validity of our coefficient estimates and p-values. Finally, since our model may be misspecified (based on Figure 8; perhaps indicating poor model fit to the data), the validity of our conclusions is questionable.

Our analysis uncovered some underlying trends that may be associated with the odds of receiving a favorable call in "crunch time". The trend of big men being significant predictors of player favor is quite interesting and may describe the association between the response variable and how players act during quick or high-pressure situations. Additionally, certain referees being significant may highlight an association between the response and officiating styles — a relationship we would love to explore in future work.

# Appendix

## Data Cleaning: Justification for Excluding Certain Variables

As mentioned previously, we excluded the following variables from the dataset: the quarter when the call was made, the NBA season, all remaining time variables, comments on the play, game details, all variables containing webscraping information, all variables containing dates, all variables containing scores, all variables containing game ID information, and all team-based variables. These variables were omitted because they were either redundant or irrelevant to our modeling approach. For instance, since we chose to create a variable indicating whether the call was made in the final minute of the game or not, all other (quantitative) time variables are redundant. We also do not believe that the date or season of the game would affect favors enough to warrant including them in the final model as referee calls should be consistent over time. Furthermore, we removed the away and home team scores because these represent the results of each game and would not have impacted favors made during the game itself. Since we were interested in performing a player-level analysis, we chose to omit all team-level variables.

## Exploratory Data Analysis: Favors for Committing Players by Type of Violation

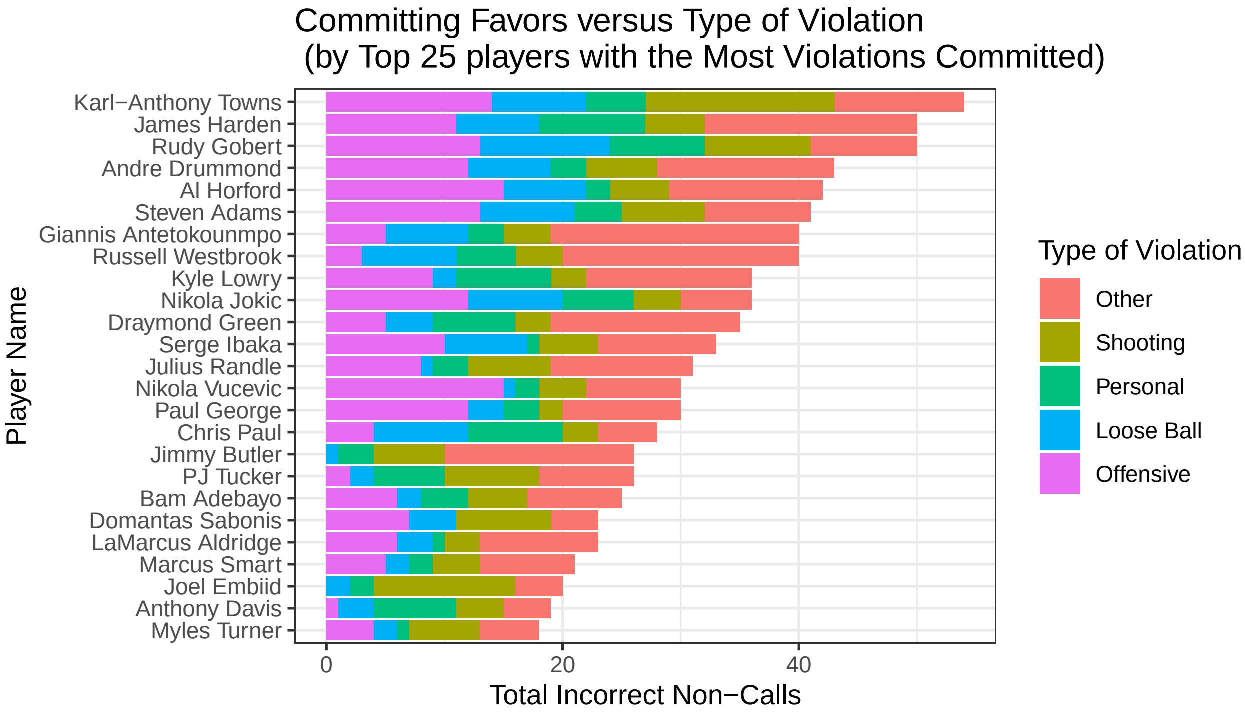

Figure 2: Favors for Committing Players by Type of Violation

Here, we see how the proportions of different types of calls vary across players as well. For example, Nikola Vucevic has the highest proportion of favors on offensive calls, while Karl-Anthony Towns has the highest proportion of favors on shooting calls. It interesting to note how the size and height of these players has a relationship with call type as well. Therefore, we suspect that the type of call may be associated with how the call favored a particular player.

## Distribution of Response Variable

Table 2: Distribution of Player Favor (Raw Dataset)

| Favoring the Committing Player | Count | Proportion |
|---:|---:|---:|
| 0 | 54886 | 0.932 |
| 1 | 4030 | 0.068 |

Table 3: Distribution of Player Favor (Final Dataset)

| Favoring the Committing Player | Count | Proportion |
|---:|---:|---:|
| 0 | 9028 | 0.92 |
| 1 | 781 | 0.08 |

### Distribution of Game Attendance

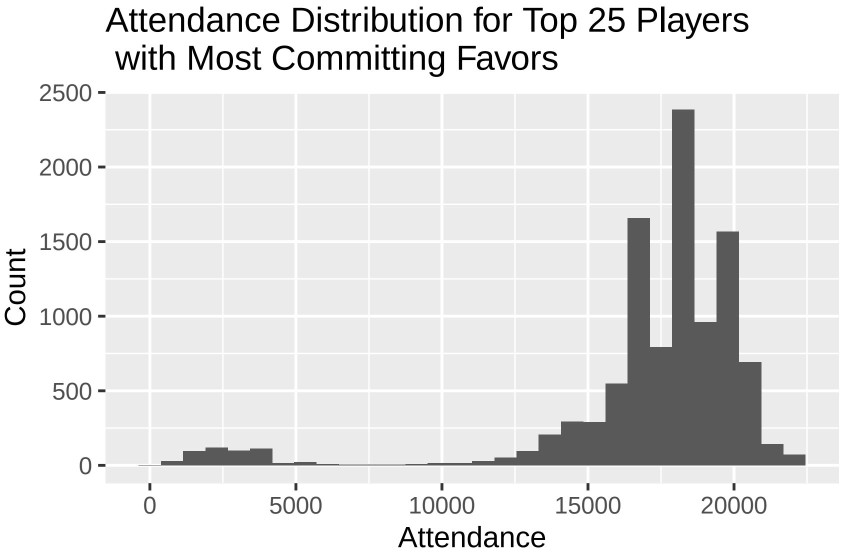

Figure 3: Distribution of Attendance (Filtered for Top 25 Players)

## Exploratory Data Analysis: Potential Interaction Effects

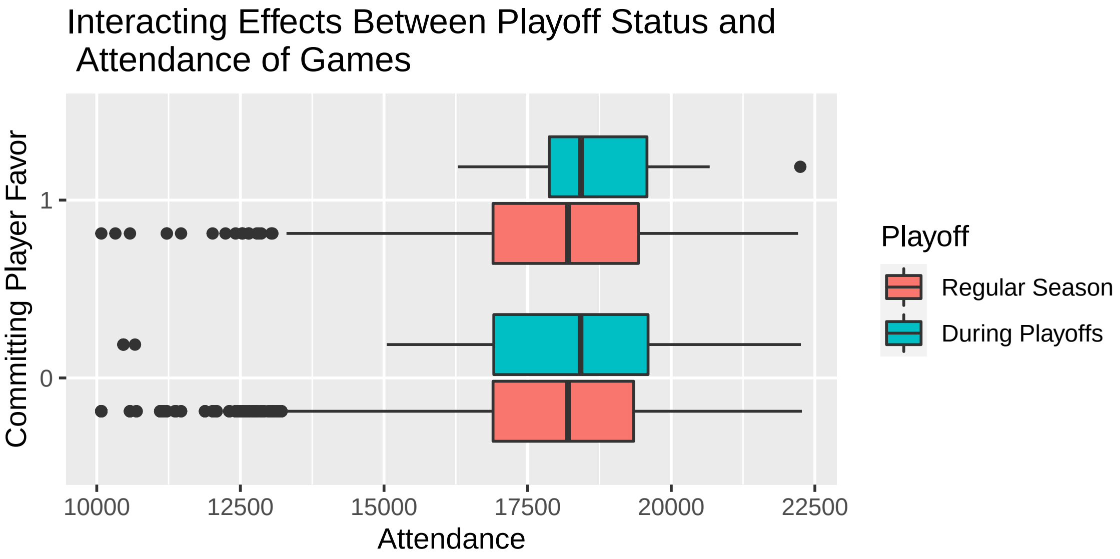

Figure 4: The Relationship Between Playoff Status and Attendance of Games

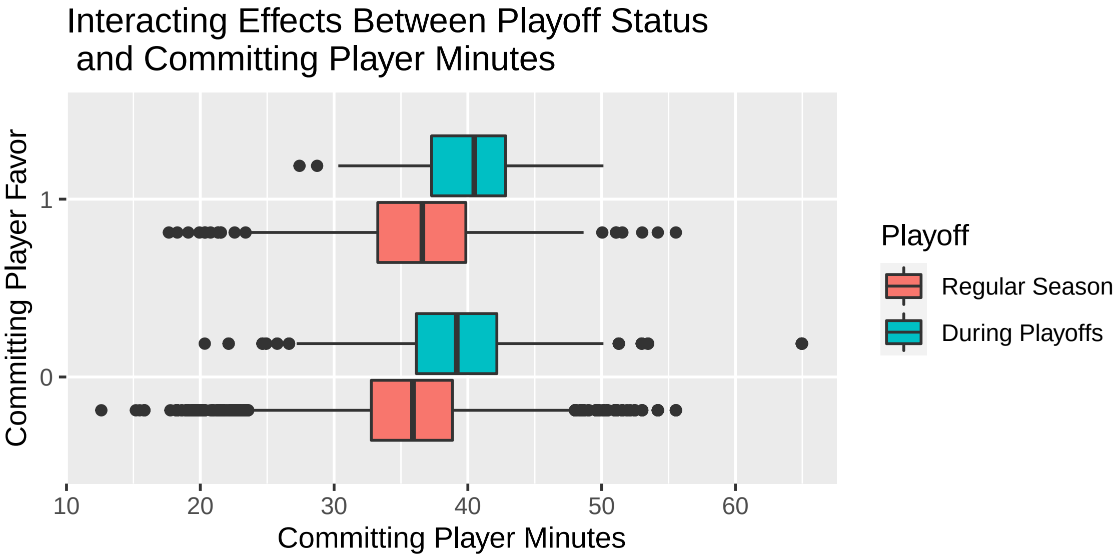

Figure 5: The Relationship Between Playoff Status and Committing Player Minutes

## Assessing Model Assumptions

We then checked the assumptions for our binary logistic regression model, namely that the log of the odds ratio is a linear function of the quantitative explanatory variables and that the observations are independent of each other (we already mentioned that the response variable is binary, that, before adding the interaction term, the model did not suffer from severe multicollinearity between the explanatory variables, and that the sample size of the final dataset is sufficiently large). To check the first assumption, we created empirical logit plots for each quantitative explanatory variable (game attendance and committing player minutes).

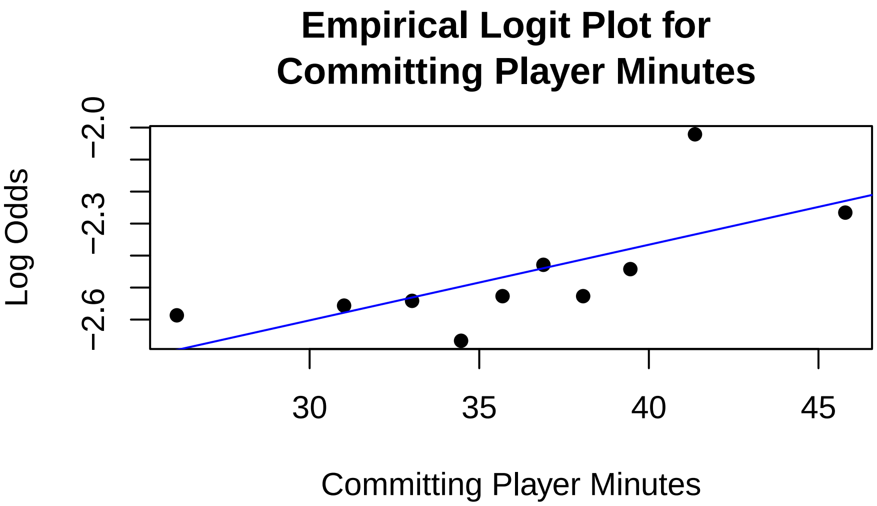

Figure 6: Assessing Linearity for Committing Player Minutes

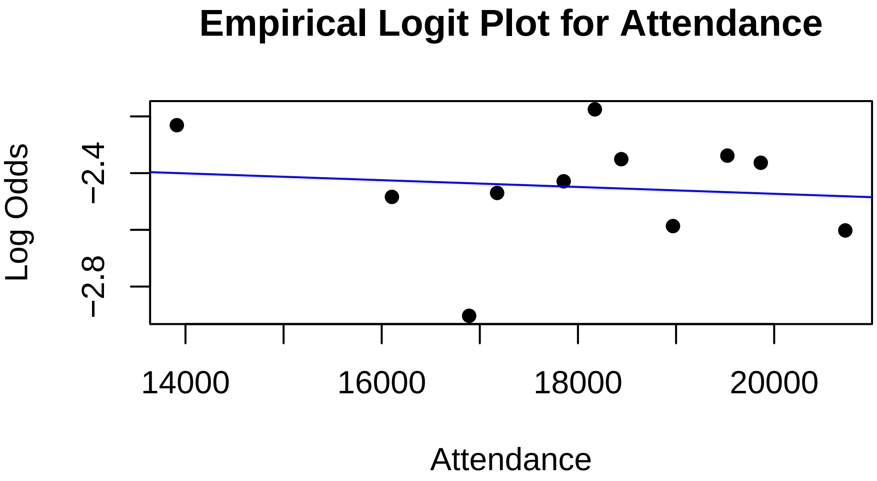

Figure 7: Assessing Linearity for Attendance

Figures 6 and 7 indicate a roughly linear relationship (the discrepancies are likely due to the presence of influential points; we tried to address this issue with our second stage of data cleaning) between the aforementioned numeric variables and the log of the odds that a referee decision favored the committing player. Thus, we concluded that the linearity condition was reasonably satisfied. A fundamental assumption underlying our modeling approach is that we can treat the referee decisions on different plays as independent. Our rationale is first and foremost that referee decisions should not carry over from game to game and, by extension, from season to season (i.e., any patterns in player favor should ideally apply only within the same game). We attempted to address the correlation in referee decisions on plays occurring in the same game by including the referee identifier in our model. Since we found that the assumptions were reasonably met, we did not feel that sensitivity analyses were warranted.

## Model Diagnostics

With regard to model diagnostics, we decided to only examine residual plots in this analysis. This decision was informed by the inferential nature of our research questions; traditional diagnostics, like ROC (receiver operator characteristic) or PR (precision-recall) curves, and even confusion matrices (i.e., picking a discrimination threshold for classification) are only relevant for prediction. We believe that a residual plot, along with an assessment of the model assumptions, can provide a rough sense of the model fit. In particular, we decided to examine a plot of the binned deviance residuals for the final model. We decided to use deviance residuals rather than standardized residuals because our final model is a generalized linear model (GLM) and the deviance is a useful goodness-of-fit statistic for this class of models. We decided to plot the binned residuals (divide data into bins based on fitted values; plot average residual versus average fitted value for each bin) as opposed to the raw residuals because the response variable is binary. We chose a small bin width (0.01) to ensure that there were many bins, allowing us to see more local patterns in the deviance residuals. Typically, we expect the binned deviance residuals to be randomly scattered around zero.

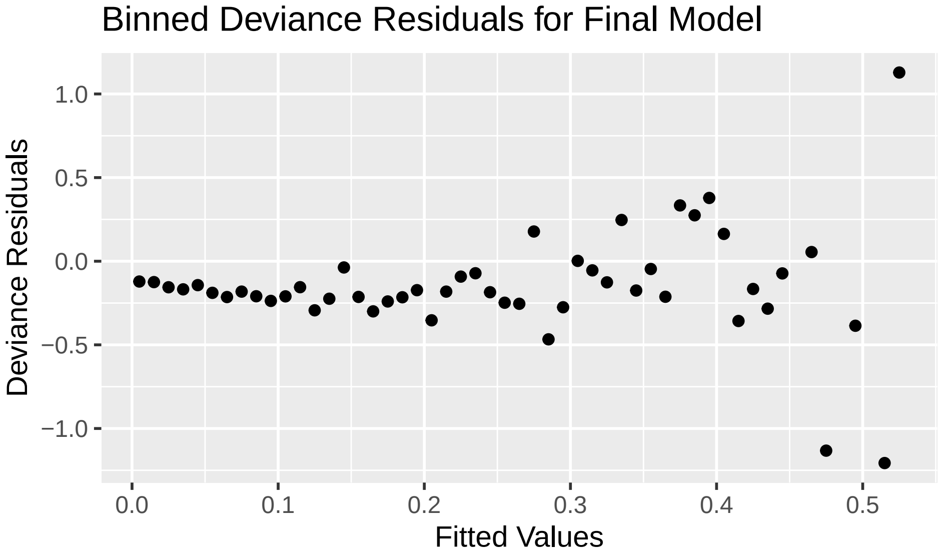

Figure 8: Assessing Residuals of Final Model

In Figure 8, we observe a “fanning” pattern, with lower fitted values tending to have deviance residuals less than and closer to 0 and higher fitted values having many more deviance residuals greater than 0. Although binned residual plots are just one way to assess goodness-of-fit, it is important to acknowledge that there may be a misspecification in our final model. One potential reason for this pattern is that our final model is missing important explanatory variables–those in the L2M dataset that we chose to exclude based on our domain knowledge and intuition and those that are not available in the L2M dataset at all (see Limitations for further discussion of potential confounding variables).

## Model Interpretations

### First Research Question

The odds that LaMarcus Aldridge receives a favorable call compared to Towns are expected to be multiplied by 0.56453, after accounting for all other factors. We are 95% confident that the odds of Aldridge being favored as the committing player compared to Towns are multiplied by a factor between 0.32925 and 0.96795, after accounting for all other factors. Similarly, the odds that Joel Embiid receives a favorable call compared to Towns are expected to be multiplied by 0.48063, after accounting for all other factors. We are 95% confident that the odds of Embiid being favored as the committing player compared to Towns are multiplied by a factor between 0.26983 and 0.85613, after accounting for all other factors. The odds that Anthony Davis receives a favorable call compared to Towns are expected to be multiplied by 0.53045, after accounting for all other factors. We are 95% confident that the odds of Davis being favored as the committing player compared to Towns are multiplied by a factor between 0.30039 and 0.93671, after accounting for all other factors. Finally, the odds that Myles Turner receives a favorable call compared to Towns are expected to be multiplied by 0.50622, after accounting for all other factors. We are 95% confident that the odds of Turner being favored as the committing player compared to Towns are multiplied by a factor between 0.32028 and 0.86083, after accounting for all other factors.

### Second Research Question

The odds that a committing player is favored are expected to be multiplied by 0.13549 if the call is a personal foul compared to an “other” violation, after accounting for all other factors. The odds that a committing player is favored are expected to be multiplied by 0.34314 if the call is a loose ball foul compared to an “other” violation, after accounting for all other factors. Lastly, the odds that a committing player is favored are expected to be multiplied by 0.20749 if the call is an offensive foul compared to an “other” violation, after accounting for all other factors.

# Formal Response to Peer Review Feedback

***An overall comment I have is that your write-up failed to stay within the page limit, but there was a lot of wasted space (I think you actually could have fit everything in if you were more efficient/judicious in what you included).***

We fixed this issue when we decided to pursue one model instead of two models. There was less output overall to include. We modified our tables to only include coefficients that are interesting and relevant to our results section.

Introduction: 15/20

***I liked the context/background about the L2M report and the reference to previously conducted research, but would have appreciated a bit more motivation beyond "we wanted to do better than Goldenberg's article." In the absence of this article, why should people care about referee bias? Why is this potentially interesting research? I like the narrow research focus on player-specific calls in looking at whether some might have differential outcomes in terms of calls against them. This focus on a specific inferential/explanatory question rather than prediction is appropriate.***

We added more context in our introduction section (specifically motivation) to elaborate on why referee bias for players is important. We specifically talk about how incorrect calls on one player can create a monumental butterfly effect that impacts who makes it to the playoffs within a season and who wins championships, along with how these outcomes impact profits and sales for each team.

***The description of the data-cleaning process is clear and I liked that you didn't use raw variable names in your final manuscript (you'd be surprised at how many groups did. . . ). However, I would have appreciated more justification/rationale for /why/ certain decisions were made. For instance, you ignored many variables - why was that ok? Were they not potentially relevant to your modeling approach? As well, decisions such as mean imputation carry with them implicit assumptions (e.g., about the missingness mechanism). Were they appropriate? What are some potential ramifications of these choices, and before even getting there, how much data were even imputed? Quick note - avoid non-standard language like "mutated" when describing your data manipulation process.***

We made sure not to use the word mutated and specifically stated that we created new variables and added them to the dataset or transformed the variable/column that exists in the dataset. We also added a paragraph of justifications for variables we include and exclude that are based on our team's domain knowledge of basketball and our research question. We also added how many mean imputations were done in our writeup, along with what the ramifications are in our limitations section (noting specifically implications of biased coefficients and incorrect p-values). We also note our assumption that the data is missing MAR.

***I did not find the tables in the EDA section to be useful - they take up lots of space to essentially provide one statistic each. As well, the tables themselves were not formatted professionally, with raw header names kept in the table. You could have simply written something along the lines of "518 (0.9%) of calls were in favor to the disadvantaged player" and get across the same information as the table. As well, the two tables displayed different information, but have the same title.***

We moved these tables to our Appendix and reference these tables in our narrative. We also added a caption/label to these tables in our Appendix to make sure they are formatted nicely.

***The visualizations on page 3 were more effective, but I'm not sure what "final five" refers to (it is not referenced in the text. Why are five minute calls included?). As well, these plots were only among the top ten players, so make sure that your descriptions are specifically only among these players. It's possible that there are other players with higher offensive violation incorrect non-calls than Gortat, but they are not displayed here.***

We clarified that this variable (final minute) should have two levels: in the final minute or not in the final minute. The original "final five" level refers to the last five minutes of overtime, but since not all games go into overtime, this was not an appropriate label for the non-final minute level. Instead, we left this as a binary variable between being in the final minute of the game or not in the final minute of the game. We also make it clear in our EDA that our conclusions about the visualizations are only relevant to the top 25 players with the most violations committed.

Methodology: 22/40

***Upon closer reading of your manuscript, there were substantial issues with the methodology, and unfortunately they're directly related to your exact research questions. From the introduction, I was under the impression that you aimed to identify players who may be differentially favored or "unfairly called against," but it actually looks like you're only looking /within specific players/ and looking at characteristics of their calls. I think that these specific models, while aimed at answering these specific research questions, miss the overall point as implied originally in the introduction (and also are less interesting overall - why should we care about these two players only?).***

We changed our research questions so that our model is relatively more generalizable. Our research questions now generalize to the top 25 players with the most committing violations rather than looking at a specific player. We justify these choices as well, and our research question now aims to identify players who may be differentially favored among the top 25 we identified.

***In terms of variable selection I appreciate that you didn't use some sort of black-box likelihood-based technique due to your specified goal, but again I would have appreciated some context-specific rationale for why the specific variables were chosen. The rationale for including interaction terms was not satisfactory - interaction terms should be included if there is reason to suspect that the conditional relationship between a predictor and the response depends on the value of another predictor. Given your previously stated goal of an exploratory/inferential model vs. a prediction-based one, use of solely a p-value to make this decision was not the most convincing rationale.***

We changed a few things about our model. To start, instead of making models for specific players, we created a model for the top 25 players with the most violations and made this a categorical variable. We also included variables based on our domain knowledge and whether they would be appropriate for our research question. Since we are focusing on the players who committed violations, we eliminated variables related to teams, the disadvantaged player, and identifier variables. Due to the new subset of data, we were able to use a binomial model as we did not have dispersion issues in this data subset. We also reworded our rationale for including interaction terms.

***The choice of a quasibinomial model was fine, although I admit being a bit disappointed with the rationale "it was the only thing that fit." This might be related to the fact that you focused only within these players and only considered some of the variables. It's possible that you had some sort of perfect separation issue here. Did you create contingency tables examining crosstabs of your categorical variables? My suspicious is that there are some empty, or very-nearly empty cells, especially with the interaction terms.***

Resolved via regrade.

***The assumptions for the model are reasonably assessed, but I didn't find the use of misclassification rate to be the best metric of model performance. Accuracy is simply the sum of the correct decisions; given your very imbalanced classes, a simply classifier that always returns the more prevalent class would also have high accuracy. As well, as presented on the top of***

***page 6, you waste a lot of space displaying results that are already mentioned in the text (the two tables are not labeled either).***

We did not include a misclassification rate, confusion matrix, or ROC curve (as well as assigning a discrimination threshold at all) because we realized that these model diagnostics are only appropriate for prediction-based research questions. We instead assessed the assumptions for a logistic regression model, examined a binned deviance residual plot to evaluate the model fit, and justified our choice of binned deviance residuals.

Results: 8/20

***In displaying the results, be sure to specifically provide what the predicted probability is of, and also to give what you are indexing over (the index of observations). The model is additionally not quite correctly written - if you have 48 separate beta terms corresponding to each referee, you should also have dummy terms corresponding to them (basically, there are missing "x" terms in the model - can you write this in a more correct way that addresses this while still being compact?). Finally, your outcome is given as a predicted logit, but the use of the raw beta term in the model without a hat implies a population parameter on the RHS instead of the predicted beta term.***

We provided a clearer description of $\hat{p}_j$ and carefully rewrote the model with predicted $\beta$ terms and correct indexing. We also discussed in detail what each $\beta$ and $x$ term represented.

***The table displaying the results is not formatted well. In addition to having raw variable names listed, you provide results to an unreasonable number of decimal points, erroneously implying a degree of precision which is not achieved in the model (this also applies when interpreting the coefficients numerically). As well, having all numerical results in scientific notation is distracting and unclear. Finally, it appears that you are displaying exponentiated coefficients in the table, but this is not clearly stated.***

We made sure that we did not include variable names in our interaction terms and main effects as well. We also decreased the number of digits we include from 7 to 5, with the reason that this precision was necessary to ensure that values were not erroneously displayed. We also noted in our table caption that the coefficients are exponentiated. In our language, we made sure to specify "exponentiated coefficient" instead of "coefficient." Lastly, our new model outputs were a lot more reasonable than our model output in the first draft (i.e., there were no numbers that were extremely large or small), so we did not format any of our values with scientific notation.

***I'm a bit confused as to why you are making interpretations at the alpha = 0.10 level, which is non-standard (and does not line up with the fact that you provide 95% confidence intervals in the table). It's fine for things not to be significant. Changing around your alpha just so things can be significant is inappropriate. As a minor note in interpretation, the sentence "there is no evidence to suggest that there is a relationship. . . " is incorrect - there is /some/ evidence, it just wasn't enough for you to reject the null hypothesis.***

The mismatch between the alpha level and the confidence interval range was a mistake in our first draft. We decided to stay with the convention of interpreting results at the 0.05 alpha level and using 95% confidence intervals as we have no reason to change it otherwise.

***Unfortunately, the interpretations themselves are also not entirely correct. You have interaction terms in your model, but you are interpreting the main effects only of some variables which have interactions. The relationship between type of violation and log-odds of a favorable call depend on the "committing player minutes" - it is not simply 3.5E-10 across the board.***

Incorrectly interpreting the predictors involved in interaction terms was a mistake in our first draft. We corrected this error in our final draft. For example, our new model contains an interaction between minutes played by the committing player and whether the call occurred during a playoff game. Given this, we made sure to interpret the regression coefficient for minutes played by the committing player in the context of whether the call occurred during a playoff or regular season game.

Discussion: 17/20

***The discussion section summarized your model results and addressed potential limitations of the analysis. One quick note is that you appear to conflate model selection with variable selection in the discussion of the model limitations. In general, the limitations are discussed soberly - unfortunately I think the overall research question/approach might have some major fundamental issues, but it may be too late to "change course," so to speak.***

All language conflating model selection with variable selection was edited or removed.

***As another comment, ending on "these are some problems with our approach" isn't a very satisfying ending - instead, provide a brief, high-level summary of what you've learned, where your research fits in with other work, and why what you've done is useful and meaningful. A final note is to remember to keep the same high standards in tables/output in your appendix as you do in the main text.***

A summary paragraph was added to the end of the limitations section and the section itself was retitled to Limitations and Summary. The Appendix was cleaned up and reorganized heavily.