The Cost of Referee Bias in Water Polo

James Graham and John Mayberry

January 10, 2018

Section 1. Official Bias

Official bias is now a well established phenomenon in almost all major sports. 40 years of quantitative research into this area has supported common stereotypes that referees tend to favor the home team (Nevill and Holder 1999), are adverse to sequential calls against the same team (Anderson and Pierce 2009; Noecker and Roback 2012), and in some situations, supportive of the underdog (Brymer, Holcomb, and Rodenberg 2015). In addition, referees in many sports seem to have an innate tendency to favor the losing team. For example, in NCAA basketball, it has been shown that foul calls are more likely to go against the losing team (Anderson and Pierce 2009) and in soccer, it has been shown that losing teams are more likely to receive penalty kicks (Plessner and Betsch 2001). In baseball, a similar effect is also present in strike calling patterns: strike zones tend to increase when the batter is up in the count and decrease when the batter is down (Moskowitz and Wertheim 2012; Green and Daniels 2014).

Elite water polo provides a particularly interesting forum for investigating officiating for two reasons. First of all, as Figure 1 demonstrates, more than half of all goals in men's water polo (and almost half of all goals in women's) result directly from major defensive fouls, which can take on one of two forms in the sport

1. *Exclusions* in which a player on the defensive team is temporarily suspended for 20 seconds giving the offense a six on five 'power play' advantage, or
2. *Penalty Shots* in which a severe goal-preventing infraction is committed within five meters of the goal resulting in a penalty shot at the goal.

Both exclusions and penalty shots provide the offense with significant advantages and hence, any biases in the rates at which they are called could largely impact the outcomes of games.

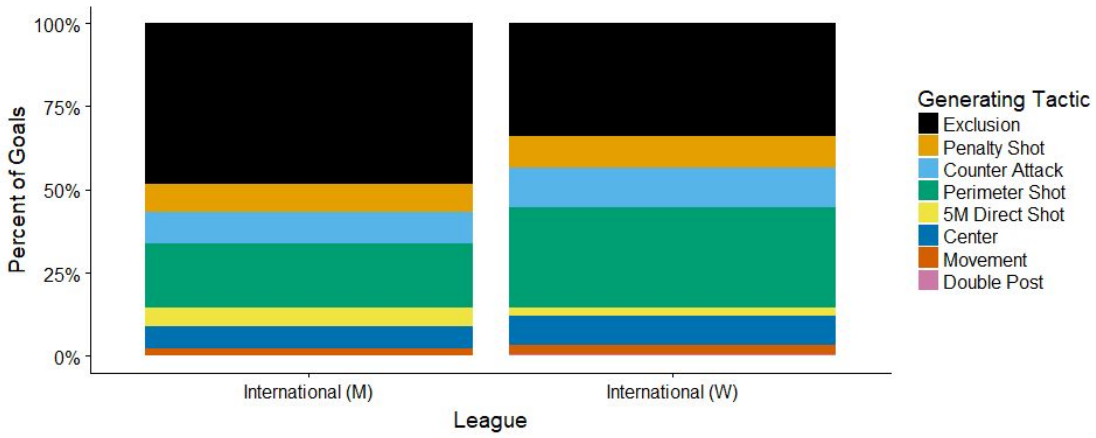


Figure 1. Breakdown of goal generating scenarios and tactics for 176 (97 men, 78 women) elite water polo contests from 2012-2015. For more details on the dataset see Section 2.

A second reason for studying water polo officiating stems from the following paradox first pointed out in (Graham and Mayberry 2014): although most goals result from major defensive fouls, the winning team rarely gets more such opportunities. Figure 2 shows the percentage of games in which the winning team received more scoring opportunities from defensive fouls as opposed to games in which the losing team had more opportunities. In men's contests, the winning team had more opportunities in only 37% of all contests while in women's, they had more opportunities in only 27%. In contrast, (Graham and Mayberry 2014) showed that if one looks at *Exclusion Conversion Rate*, defined as the fraction of power play opportunities which are converted to goals, the winning team had a higher value in almost 90% of all contests, even if we restrict our attention to 'close' games (those decided by 3 or fewer goals). These results again highlight the importance of exclusion opportunities while also providing evidence of a foul calling bias in favor of losing teams.

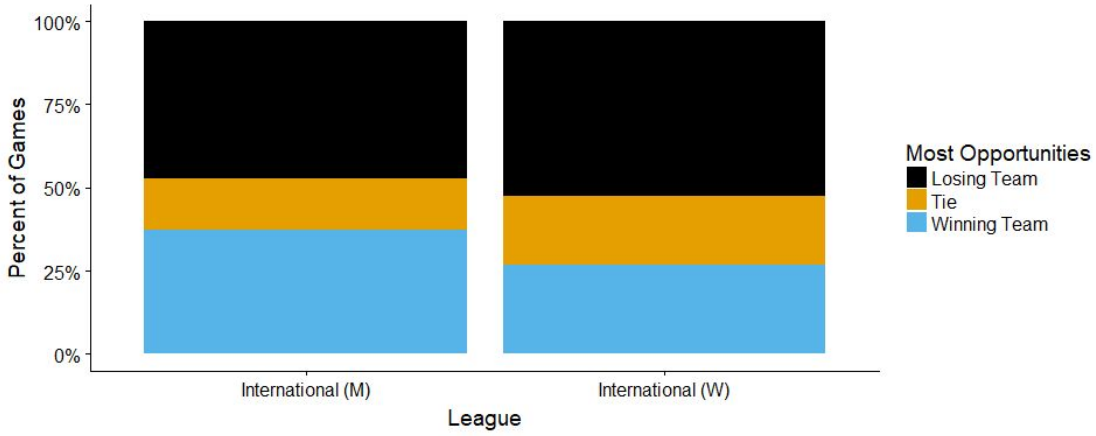


Figure 2. Distribution of which team had the most scoring opportunities from defensive fouls based on the game outcome. Here tie could mean either that both teams received the same number of opportunities or that the game ended in a tie (shootouts excluded).

(Graham and Mayberry 2016) further studied this phenomenon in elite men's water polo on a possession by possession basis. They defined two new statistics: the *Defensive Foul Rate* (DFR), defined as the probability of being awarded an exclusion or penalty shot opportunity on a given offensive possession, and the *Offensive Foul Rate* (OFR), the probability of getting called for an offensive foul resulting in a turnover of possession to the opposing team. Using hierarchical logistic regression, they studied the impact of various game state factors (eg. sign and magnitude of offensive team's lead, sequential fouls, scoring momentum) on both the DFR and OFR. In particular, they showed that there is statistically significant evidence of losing team bias in water polo officiating with the odds of drawing a major defensive foul decreasing by about 27% when the offensive team is winning or the game is tied. They also showed that this losing team bias persists even after accounting for differences in playing style (i.e.~offensive and defensive tactics), game-score (close vs unbalanced games), event (Olympics, World Championships, European Championships), team, and game-time. Figure 3 illustrates the overall losing team bias in defensive foul calling rates.

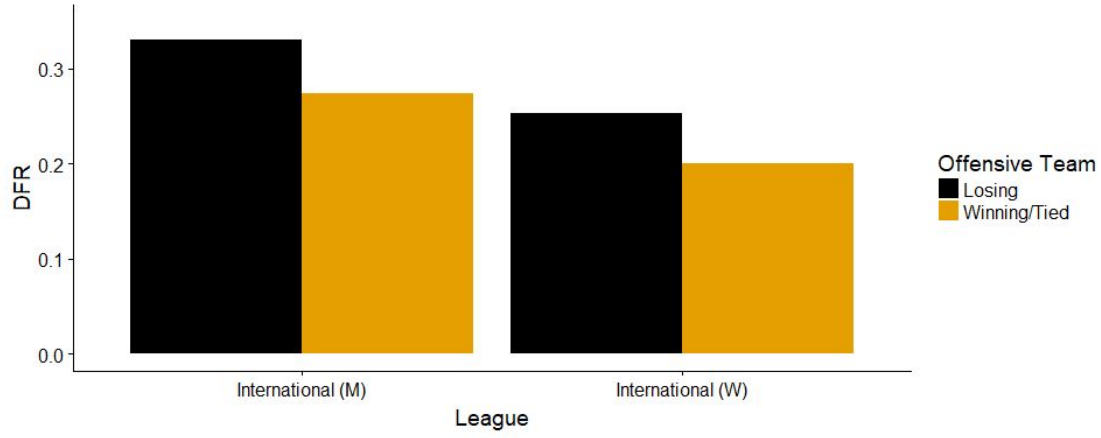


Figure 3. Differences in the probabilities of the offensive team drawing a defensive foul (DFR) between losing and winning/tied teams.

What is unclear from (Graham and Mayberry 2016) or other investigations of foul calling biases is the extent to which such biases may actually impact the outcome of a game. Evaluating the impact of losing team bias in real water polo contests is more challenging than validating its existence. It is difficult to isolate the effect of a single foul call on the final outcome of a game and predict how the outcome would have changed if a particular foul had not been called at a particular time. Instead, we will investigate the impact of losing team bias here by generating our own games via random simulation. Although this igrnores potential psychological effects on 'game momentum', parallel game simulations will at least allow us to assess the objective cost of an additional foul call (or lack of call) on the games' final score.

To explain our approach, imagine a contest between two equally matched elite men's teams which we will refer to as 'Dark' and 'White'. In the real world, these two teams would only be able to play once, but in the virtual world, we can generate two parallel games which are

identical in all respects except that in Game 1, there is no losing team bias in foul calling while in Game 2, the losing team bias is present. By comparing the pattern of goals scored by Dark in Game 1 with the pattern of goals scored by Dark in Game 2 (and similarly, compare White scoring patterns between the two games), any differences will be goals scored (or lost) as a result of losing team bias.

The rest of this chapter will be organized as follows. In Section 2, we will give a more precise description of our simulation process and the data set used for callibrating parameters. In Section 3, we will summarize the results of our simulations and the estimated cost of losing team bias. Section 4 is dedicated to a more in depth study of the symmetric case in which the losing team advantage is the same as the winning team disadvantage. In this scenario, we can obtain some explicit formulas based on the use of binomial probabilities. Finally, we conclude in Section 5 with some remarks and questions for furter investigations.

Section 2: Model Description and Parameters

To simulate the cost of referee bias, we generate $M = 10,000$ coupled pairs of games, one of which incorporates a losing team bias in foul calling and one in which foul calling rates are independent of the game-state. The evolution of each game is based on a discrete time Markov chain model in which each step represents a change in possession. A *possession* is defined as the period of time from when a particular team takes offensive control of the ball until offensive control returns to the opposing team. This possession orieted approach to studying water polo was first proposed in (Graham and Mayberry 2014) inspired by a similar approach earlier applied to model basketball games (Kubatko et al. 2007).

The state of the unbiased game during game possession k is a triple (D_k, W_k, O_k) where D_k, W_k tracks the current game score and O_k tracks who is currently in possession of the ball. During each possession, a goal is scored with probability g independent of the current score. Possession then switches to the opposing team and the process continues for N possessions where N is a random variable. The team who gets the ball first is determined by a fair coin flip. The unbiased game is then ``coupled'' with the biased game in the following way:

- If the offensive team is losing during possession k, then a goal is scored in both games with probability g and only in the biased game with probability b_ℓ for some parameter $b_\ell > 0$.
- If the offensive team is winning or the game is tied during possession k, then a goal is scored in both games with probability $g - b_w$ and only in the unbiased game with probability b_w for some parameter $b_w > 0$.

The parameters b_ℓ, b_w represent the respective boost and reduction in goal scoring rates resulting from losing team bias.

To determine appropriate values for the parameters g, b_w, b_ℓ and the distribution of N, we used game data from 68 elite men's water polo games including 23 from the 2012 London

Olympics (henceforth Oly), 25 from the 2013 World Championships (WC), and 20 from the 2014 European Chamionships (EC)[1]. This data set included all playoff games from the three tournaments as well as selected games from the preliminary rounds between competitive teams. The teams involved in these games are listed in Table 1 below. Games were filmed from mid-court by representatives from Team USA water polo. While camera position varied, all twelve players and the defending goalie were kept in frame at all times. The recorded tapes were later viewed by the first author or one of his assistants and play by play game logs were recorded summarizing the outcomes of all possessions in the contests. Information transcribed about the possession included the team on offense, offensive tactic(s) employed[2], any defensive fouls called, rebounds/new clocks, and the ultimate result of the possession (Goal, Missed Shot, Blocked Shot, Goalie Save, Turnover, or Offensive Foul).

Table 1. Teams involved in our dataset

Team	Number of Games
AUS	11
CAN	2
CHN	3
CRO	15
ESP	14
GER	2
GRE	13
HUN	16
ITA	15
MNE	14
ROM	1
ROU	7
SRB	15
USA	8

[1] The figures in Section 1 also included 29 men's games from various 2015 international tournaments and world qualifiers, but these were excluded from our model because of slight differences in tracking methods and the competitiveness of the events. We also excluded women's games from our model because of differences in game play and foul calling rates. Building a similar model for women's water polo would be an interesting project for future investigations although a complication is that winnings/tied teams tend to score non-exclusion goals at a higher rate than losing teams and hence, the model assumptions used here would be invalid.

[2] See (Graham and Mayberry 2014) for a further discussion of offensive tactic classifications.

Overall, our data set included 4625 possessions (1556 from Oly, 1766 from WC, and 1303 from EC). The distribution for the number of possessions per game was roughly symmetric (median = 68, mean= 68.9 possessions per game) with 50% of all games having between 65 and 73 possessions and 90% of all games having between 60 and 76.7 possessions. Figure 4 shows the distribution of possessions across games and this empirical distribution was used to bootstrap sample N in our simulations.

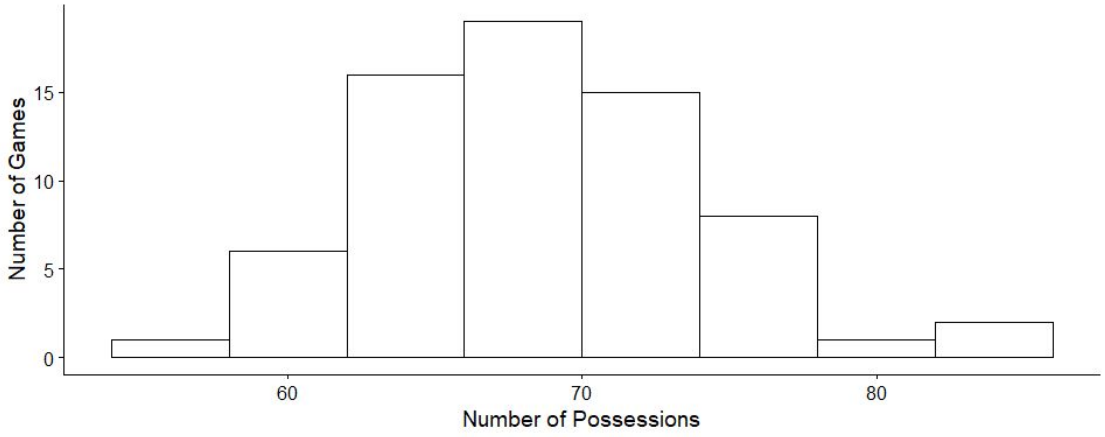


Figure 4. Distribution of Number of Possessions per Game

To determine the goal scoring probability g and the losing team bias, we categorized possessions according to the following four scenarios:

- G : A goal was scored with no foul being called
- P : A single penalty shot was granted resulting from a severe goal-preventing infraction.
- E : An exclusion foul was called resulting in a 20 second 6 on 5 power play advantage for the offense.
- C : A change of possession occurred without any of the three above outcomes. Change of possession could mean a turnover, blocked shot, saved shot, missed shot, offensive foul, or shot clock violation.

Table 2 summarizes the fraction of each possession ending in each outcome broken down by the offensive state at the start of the possesion(offensive team losing or winning/tied).

Table 2. Percentage of possessions ending in each of the four possible outcomes G,P,E,C based on the starting state of the posession with respect to the offensive team (L = Offensive team losing, W/T = Offensive team winning or tied)

Offense	Count	G	P	E	C
L	1888	0.106	0.018	0.326	0.550
W/T	2800	0.109	0.019	0.257	0.615
Overall	4688	0.108	0.019	0.285	0.589

We estimate the probability g of scoring a goal in a given possession by taking a weighted average of the three outcomes which could result in a goal:

$$g = G_O + \varepsilon E_O + \rho P_O$$

where G_O, E_O, P_O are the overall fractions of possessions resulting in G, E, P and ε, ρ are the exclusion and penalty conversion rates, respectively. Estimates of ε and ρ from our database are provided in Table 3 below. Finally, we estimate the foul calling biases by

where E_{WT} and E_{ℓ} are the probabilities of drawing an exclusion when the offensive team is winning/tied or losing respectively. Since the goal scoring and penalty shot rates did not differ significantly between losing and winning/tied teams, we leave these factors identical in the biased and unbiased games. Therefore, the winning and losing team goal scoring biases depend only on differences in exclusion calling rates between the two scenarios.

Table 3. Defensive Foul Statistics including mean number of fouls per game (Mean), conversion rates (CR), and margin of error in conversion rate estimates at a 95% confidence level (ME)

Type	Mean	CR	ME
Exclusion	20.7	0.46	0.03
Penalty Shot	1.3	0.75	0.11

Section 3: Simulation Results

We employ two metrics to quantify the impact of losing team bias in our simulations:

1. *Difference in goals scores*, defined as the difference between the total number of goals scored in the biased game and the total number of goals scored in the unbiased games.
2. *Alteration of Final Outcome*, defined as a binary variable which takes on the value of 1 if the outcome of the biased game differed from the outcome of the unbiased game. This includes situations in which one game was tied and the other yielded a clear victory for one team.

We illustrate the computation of these metrics in Figure 5 below which shows one coupled simulation of a single game between teams Dark and White. Comparing Dark 1 (the goals scored by team Dark in the unbiased game) with Dark 2 (goals scored by Dark in the biased game), we can see that there are no differences until Possession 27. At this point, the Dark team is leading in both games, but due to the presence of losing team bias in Game 2, they score a goal in Game 1 and not in Game 2. During Possession 41, the Dark team is hurt again in Game 2 when the game is tied. In contrast, the White team scores lie on the same trajectory until Possession 66. At this point, the White team is leading in Game 2 and is then hurt by a losing team bias. So to summarize the impact of losing team bias in this game, we can say that it cost the Dark team two goals and cost the White team 1, changing the game outcome from a 13-12 win for Dark to an 11-11 tie. In terms of our metrics, the difference in goals scored was 3 and the alteration of final outcome was 1.

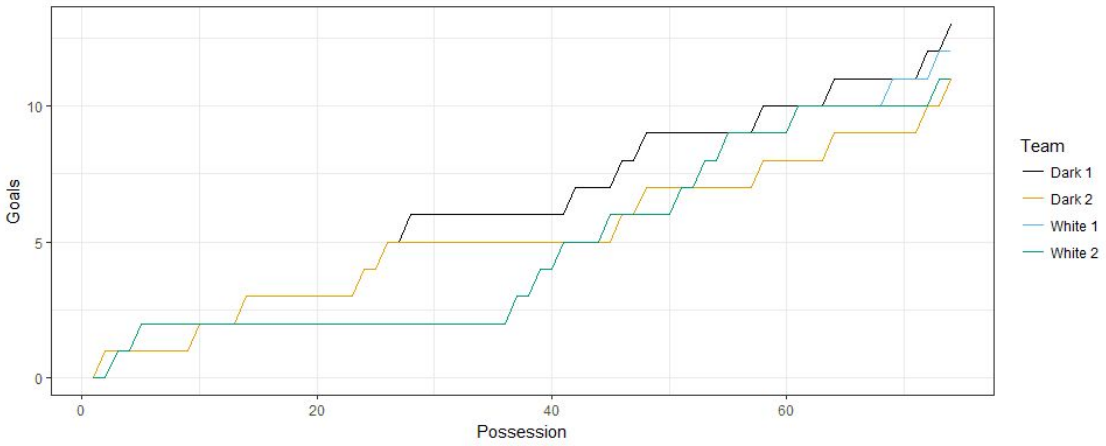
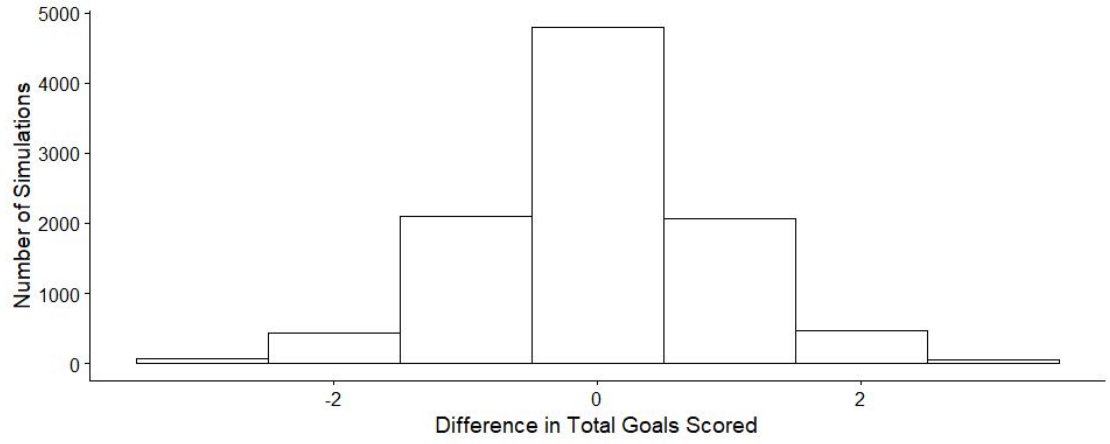


Figure 5. Sample of a coupled game simulation demonstrating the difference between the biased (Dark 2 vs White 2) and unbiased (Dark 1 vs White 1) games.

Figure 6 shows the distribution of the difference in goals scored across all 10,000 simulated games. On average, losing team bias did not affect the total number of goals scored and approximately 48% of all simulations ended with no difference between the two games. Another 42% of all games ended with a total goal difference of 1, however, the symmetry of the distribution implies that the total number of goals was just as likely to be higher in the biased games as in the unbiased games. There were a few outliers (like the game shown in Figure 5) in which the total number of goals in the two games differed by as much as three.



In contrast to Figure 6, Figure 7 shows the impact of losing team bias on game outcome in our database. Overall, about 14% of all game outcomes were altered by the presence of losing team bias. The most common alteration was from a clear victory for one team in the unbiased game to a tie in the biased game. Only 1.3% of all alterations actually switched the winner of the game from Dark to White (or vica-versa).

Figure 7. Distribution of game alterations.

Section 4: Symmetric Bias Approximation

To demonstrate the sensitivity of our simulation results to parameter selection, we include a discussion of how the fraction f of all games altered changes as a function of the amount of bias present in a game. For simplicity, we restrict our attention to the the symmetric case where the boost in foul calling biases for the losing team is equal to the reduction in foul calling rates for winning/tied teams. Figure 8 shows the results of 10000 simulated games at various foul calling biases ranging from 0.005 to 0.1. From Table 2, we can see that the foul calling bias in men's water polo is not quite symmetric: losing teams get a boost in exclusion calling rates of around 0.041 per possession while winning/tied teams get a reduction of only 0.028. Nevertheless, using a symmetric bias of between 0.03 and 0.04 provides a rough approximation to this scenario.

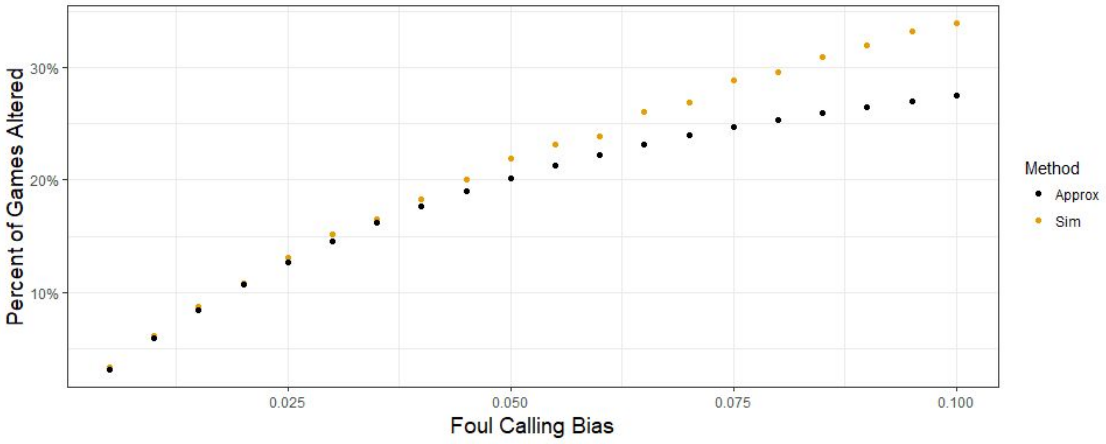


Figure 8. Comparison of symmetric bias with theoretical approximations.

Figure 8 shows the results of these simulations. When the foul calling bias is small, the shape of this curve can be explained by direct computation. During each possession, the

probability of a goal being scored in one game but not the other is $b_w = b_\ell = b$. Therefore, the probability of at least one extra goal being scored in the biased game is

$$1-(1-b)^n$$

Of course, this is not enough to actually change the outcome of a game, but when b is small, the probability of more than one extra goal being scored is relatively small and hence, most game altering situations will fall into one of the following two categories:

1. The unbiased games ends in a tie, but one team scores at least one extra goal in the biased game.
2. The unbiased game score differential is 1, but the losing team scores at least one extra goal in the biased game.

Since goal scoring events are independent, the final scores in the unbiased game can be approximated by independent Binomial random variables D and W, each with approximately $\nu/2$ trials and success probability g where ν is the mean number of possessions per game. Therefore the probability of scenario 1 above is approximately

$$P(D = W) \times (1-(1-b)^{\nu})$$

and the probability of 2 is

$$P(|D-W| = 1) \times (1-(1-b)^{\nu/2})$$

Therefore, we can approximate the fraction $f(b)$ of altered games as

$$f(b) \approx P(D = W) \times (1-(1-b)^{\nu}) + P(|D-W| = 1) \times (1-(1-b)^{\nu/2}).$$

ν can be computed from the data (see Section 2 and Figure 4) while from standard formulas for Binomial probabilities, we can easily compute

$$P(D = W) = \sum_{j=0}^{\nu/2} \left[\binom{\nu/2}{j} g^j (1-g)^{\nu/2-j} \right]^2$$

and

$$P(|D-W| = 1) = \sum_{j=0}^{\nu/2} \binom{\nu/2}{j} g^j (1-g)^{\nu/2-j} \times \left[\binom{\nu/2}{j-1} g^{j-1} (1-g)^{\nu/2-j+1} + \binom{\nu/2}{j+1} g^{j+1} (1-g)^{\nu/2-j-1} \right]$$

The resulting approximation is shown in Figure 8 along with the simulated values and we see it provides a good approximation for foul calling biases of 0.05 or smaller which contains the case of international men's water polo[3]. For larger b, the approximation is invalid because it ignores the now relevant percentage of all games which differ by more

[3] In international women's water polo, the foul calling bias is smaller ranging from a 0.03 boost for losing teams to a 0.02 reduction for winning/tied teams.

than one goal in the unbiased scenario as well as the games in which both teams could have received multiple offsetting calls in the biased game.

Section 5: Discussion

Parallel game simulations provide a novel approach for investigating the impact of foul calling biases. By running a game twice, once with and once without a particular type of bias, we can isolate the impact of this factor on the evolution of the game. Here, we focus on losing team bias, the established fact that a losing team in water polo has a better chance of getting a foul call in their favor than the winning team. Our simulations suggest that the presence of this bias will alter the total number of goals scored in about half of all games between typical, equally matched elite teams and could be altering the outcome in about 14% of all such games. A small fraction (just over 1%) of these alterations actually switched the winner of the contest, but the most common alteration was switching from a clear victory for one team in the independent game to a tie in the biased game. This means that the main effect of losing team bias is producing more overtime games than there should be. We also were able to examine the dependence of the fraction of games altered on the size of the losing team bias which may help understand the impact in leagues where the amount of bias is stronger (or less strong) than in elite men's water polo.

One improvement which could be made to our simulations is in the method by which game length was determined. We determined game length at the start of a simulation by randomly sampling a value from the empirical distribution of possessions and then running the game. But in reality, the number of possessions is also impacted by the number of exclusion fouls called since such fouls add to the length of a possession. A more robust model would be to give a fixed game time and allow possession lengths to vary according to a random variable with additional time added after each called exclusion. Unfortunately, our data set does not provide information on the length of possessions so we were unable to compare how this alteration would impact our results. Another improvement which could be made to our simulations is the inclusion of other factors which have been shown to affect foul calling rates such as sequential foul call biasing.

In (Graham and Mayberry 2016), it is shown that losing team bias is present more strongly in close games than blowouts and persists across different offensive and defensive tactical choices. Here, we also show that the direct goal scoring rates for losing and winning teams are similar (see Table 2) and that the differences in foul calling rates feed into differences in turnover rates. Together, these investigations suggest that it is referees, and not teams, who provide the primary source of losing team bias in the sport. These observations are consistent with an old adage that referees prefer to be 'fair' (giving equal opportunities to both teams) as opposed to being 'objective' (calling fouls based on severity of infractions alone) (Askins 1978). With the help of parallel simulations, we have provided evidence that this principle is affecting a significant fraction of games and generating more excitement (or anxiety) than their should be in elite international water polo.

References

Anderson, Kyle J, and David A Pierce. 2009. “Officiating Bias: The Effect of Foul Differential on Foul Calls in Ncaa Basketball.” *Journal of Sports Sciences* 27 (7). Routledge: 687–94.

Askins, RL. 1978. “The Official Reacting to Pressure.” *Referee* 3: 17–20.

Brymer, Rhett, Tim R Holcomb, and Ryan M Rodenberg. 2015. “Referee Analytics: Bias in Major College Football Officiating.” In *2015 Mit Sloan Sports Analytics Conference*.

Graham, James, and John Mayberry. 2014. “Measures of Tactical Efficiency in Water Polo.” *Journal of Quantitative Analysis in Sports* 10 (1): 67–79.

———. 2016. “The Ebb and Flow of Official Calls in Water Polo.” *Journal of Sports Analytics* 2 (2). IOS Press: 61–71.

Green, Etan, and David P Daniels. 2014. “What Does It Take to Call a Strike? Three Biases in Umpire Decision Making.” In *2014 Mit Sloan Sports Analytics Conference*.

Kubatko, Justin, Dean Oliver, Kevin Pelton, and Dan T Rosenbaum. 2007. “A Starting Point for Analyzing Basketball Statistics.” *Journal of Quantitative Analysis in Sports* 3 (3).

Moskowitz, Tobias, and L Jon Wertheim. 2012. *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won*. Three Rivers Press.

Nevill, Alan M, and Roger L Holder. 1999. “Home Advantage in Sport.” *Sports Medicine* 28 (4). Springer: 221–36.

Noecker, Cecilia A, and Paul Roback. 2012. “New Insights on the Tendency of Ncaa Basketball Officials to Even Out Foul Calls.” *Journal of Quantitative Analysis in Sports* 8 (3).

Plessner, Henning, and Tilmann Betsch. 2001. “Sequential Effects in Important Referee Decisions: The Case of Penalties in Soccer.” *Journal of Sport and Exercise Psychology* 23 (3): 254–59.