2013

Biased Impartiality Among National Hockey League Referees

Michael J. Lopez
Skidmore College

Kevin Snyder

International Journal of Sport Finance, 2013, **8**, 208-223, © 2013 West Virginia University

Biased Impartiality Among National Hockey League Referees

Michael J. Lopez[1] and Kevin Snyder[2]

[1]Brown University
[2]Southern New Hampshire University

Michael J. Lopez is a PhD candidate in the Department of Biostatistics in the Brown University School of Public Health. His research interests include the application of statistical methods to identify trends and responses to incentives in sport.

Kevin Snyder is an assistant professor of sport management at Southern New Hampshire University. His interests include sport management and business strategy in the innovation and knowledge services.

Abstract

This paper builds an economic model of referee behavior in the National Hockey League using period-specific, in-game data. Recognizing that referees are influenced by a desire for perceived fairness, this model isolates situations where a referee is more likely to call a penalty on one team. While prior research has focused on a systematic bias in favor of the home team, we find that referee bias also depends upon game-specific conditions that incentivize an evening of penalty calls. Refereeing games in this fashion maintains the integrity of the game, thus benefiting spectator perceptions and opportunities for financial returns.

Keywords: biased impartiality, National Hockey League, referees

Introduction

Fans, coaches, and players often complain that referees make decisions based on factors other than what occurred on the playing surface. Although support for match fixing is scant, spectators frequently accuse officials of bias against their team. In reality, the bias usually rests with the irate fans or participants, while referees are loath to discuss the matter. However, anecdotal evidence from referees suggests that they are aware of how games are being called and work to balance infractions. Adding to the debate on make-up calls is former NHL senior referee, Kerry Fraser, who stated in 2011,

When a referee realizes he made a call in error, the most difficult thing to overcome is "human nature"—the natural tendency to attempt to fix it or make it right with an even-up call. The referee wants to be fair and recognizes something he did just wasn't fair.

Fraser went on to say, "It occurs in all jobs and all walks of life. It can be enhanced when under pressure."

Combined with fan reactions, these quotes suggest the possibility of temporary bias existing in the officiating of a sporting game. As the individuals charged with enforcing the rules of the game and policing player behaviors, referees are in place as an impartial party to maintain order and ensure a safe, fair outcome. However, officials are also faced with numerous pressures of their own including from spectators in the building (Boyko, Boyko, & Boyko, 2007; Buraimo, Forrest, & Simmons, 2009; Moskowitz & Wertheim, 2011; Nevil, Balmer, & Williams, 2002; Scoppa, 2008; Sutter & Kochera, 2004), from their peers, and from the league offices that determine their employment. Significant financial ramifications also ride on perceptions of a fairly called game. Balancing these pressures and maintaining strong working relationships with each group is a high wire act, delicately walked by referees each game. Given these expectations, how do referees maintain the integrity of the game while balancing numerous influences? Also, when would a referee be most likely to exhibit temporary bias to further the broader goal of the integrity of the match?

Although several scholars have explored officiating bias under different conditions (Boyko et al., 2007; Moskowitz & Wertheim, 2011; Mongeon & Mittelhammer, 2011), few studies have considered that biases may change throughout the course of a game. By looking at games and seasons in their totality, previous research has assumed that biases are consistent throughout an entire contest. Further, this assumption fails to allow for variation in referee behavior based on game conditions or other situational factors. These assumptions have led researchers to conclude that home or popular teams are always more likely to receive a favorable call from an official (Moskowitz & Wertheim, 2011; Price, Remer, & Stone, 2012). However, this paper highlights how referees subconsciously shift their bias as the game progresses based on the flow of the game. The model developed here illustrates when referees have incentives to call penalties on each team. By using a sample of National Hockey League (NHL) data, this research answers the questions of when referees allow for "make-up" calls within a game.

The paper starts by reviewing background information on the National Hockey League (NHL) and the role of referees. Second, areas where "biased impartiality" may occur within the sport of hockey are examined, followed by a review of literature on player actions. A description of the sample used in this study follows, along with our analysis. This paper concludes with a discussion of the results, limitations of our work, and avenues for future research.

Background

The National Hockey League

Based on revenues, the National Hockey League is the fourth largest professional sporting league in North America. The NHL consists of 30 teams with 23 located in the United States and 7 in Canada. Revenues for the 2012 season are estimated at $3.2 billion, representing a 50% increase over the past 7 years (Staples, 2012). As with most other professional sport leagues, a disproportionate share of these revenues are earned in the postseason (Leeds & von Allmen, 2004; Robst, VanGilder, Berri, & Vance, 2011). However, differing from the NFL, NBA, and MLB, the NHL's primary source of revenue is ticket sales. The variability of this revenue stream creates greater incentives for the NHL to create close games, perhaps even favoring home teams, with high levels of outcome uncertainty across games and seasons (Coates & Humphreys, 2012; Fort, 2011; Staudohar, 2005). In playoff games, these incentives are enhanced due to a larger audience following through numerous media outlets.

The primary responsibility of hockey officials (two referees and two linesmen) is to monitor the play of the teams, watching for violations and penalties as a way of providing legitimacy to the contest. The referees are the only ones charged with calling penalties, though they can consult with the linesmen when necessary. Violations, such as offsides or icing, are mostly clear cut with little judgment necessary to determine where the puck was in relation to an offensive player. However, there is significant discretion in awarding penalties given the rule book's latitude in defining undesirable behavior. Typical penalties include stick violations of slashing, hooking, or high-sticking, as well as physical infringements such as boarding, roughing, interference, elbowing, and fighting. Additionally, teams are assessed a penalty for having too many players on the ice and delaying the game, though these are easier to identify and called with near uniformity. When a penalty is called, the offending team loses a player on the ice for the shorter of 2 minutes or until the other team scores. More severe penalties can be 5 minutes or disqualification from the game. Losing a player is a significant disadvantage. In the 2011-12 season, teams scored on 17.0% of their opportunities with an extra skater, and 20.9% of all goals were scored via a power play (www.nhl.com). In a game where winning teams typically score between three and four goals, penalties can play a significant role in determining the outcome.

In an attempt to call a fair game, referees may possess an unconscious bias towards evening up penalty calls. Subconsciously influenced by a desire to avoid directly favoring one team, officials may also inadvertently demonstrate bias by ignoring borderline acts that could be deemed in violation of a rule. This slight shift in standard occurs due to the degree of interpretation in the role and is not indicative of illicit actions from a referee.

The practice of evening up calls can have both positive and negative impacts on the game. One clear advantage for spectators is the likelihood of a close, competitive game. Exciting, unpredictable finishes make for good television and entertainment. However, the negative impacts of evening up calls outweigh the positive benefits. When violations are ignored or inconsistently called, participants struggle to regulate their actions to the desired behaviors. As a result, greater incentives exist for retaliation and dangerous play. Additionally, inconsistent officiating has the potential to undermine the public's confidence in the fairness of the game and the independence of the referees.

Other models of referee behavior

Although there may be many reasons for home advantage (Koyama & Reade, 2008), recent academic work has identified officials' behavior as an explanation for the greater success of home teams (Boyko et al, 2007; Pettersson-Lidbom & Priks, 2009; Leard & Doyle, 2010). Appealing to an implicit desire to appease the partisan crowd leads to a bias in the distribution of calls favoring the home team. Impacts of this behavior are illustrated in the effect of crowd noise (Nevill, Balmer, & Williams, 2002), added time in soccer matches (Sutter & Kochera, 2004) and games played in empty stadia (Pettersson-Lidbom & Priks, 2009). These studies have found favoritism effects for an aggregate group of officials, as well as individual referees. Other scholars link biases and behaviors of referees in response to crowd behaviors (Scoppa, 2008; Garicano, Palacios-Huerta, & Prendergast, 2005; Nevill et al, 2002; Boyko et al, 2006). Buraimo et al. (2009) represents an exception within the literature to the assumption of consistent biases favoring the home team, finding evidence instead of make-up calls in European football leagues. Ultimately, home teams win approximately 55% of games, a difference of 5% from the expected mean of 50% if all games were played at a neutral site. This bias has been connected to favorable home team calls and is consistent across numerous sports, including baseball, football, basketball, hockey, soccer, and the Olympic Games (Moskowitz & Wertheim, 2011; Nevill et al., 2002).

Models of player behavior

Penalty and goal outcomes depend upon both referee and player behaviors. Numerous studies have been conducted to assess the impact of player aggression on the outcome of games (Buraimo et al., 2009; Jewell, 2009; Widmeyer & Birch, 1984; Widmeyer & McGuire, 1997). In each instance, as in our model, aggression is viewed as actions beyond the boundary of ordinarily accepted activities required in the sport and is punished through the incursion of a penalty or foul.

As a context for understanding aggression and officiating, European soccer has been used to explain how referees manage yellow and red cards issued for aggressive fouls (Buraimo et al., 2009). Beyond the referee bias towards home teams found in several similar studies, players modify their behavior and become more aggressive when behind in a match (Buraimo et al., 2009). Aggressive actions by players may be the result of a tactical shift or frustration in how the game has progressed. Players may be more likely to commit a foul when the game is likely to be lost and the opportunity cost of disqualification is low (Jewell, 2009). However, when referees are added, hockey players are not deterred from committing penalties but more penalties are observed due to greater enforcement efforts (Heckelman & Yates, 2003). The combination of these studies suggests that players' actions are consistent with the team's strategy and have little variance in relation to referee behaviors.

Aggression is also studied as a strategy to increase the likelihood of success. Through an examination of hockey results, Widmeyer and Birch (1984) examine the impact of first period penalties on future performance. Although counterintuitive, committing penalties in the beginning of a game may help increase the chances of a victory through intimidation of the opponent (Widmeyer & Birch, 1984). Aggression may be more prevalent with familiarity of the opponent as penalties increase between frequent rivals seeking to carry the intimidation advantage over to multiple games (Widmeyer & McGuire, 1997).

Building on prior frameworks of referee bias and player aggression, our paper continues in pursuit of referee tendencies to balance penalty calls. Whereas prior research has sought a bias towards specific teams (Price et al., 2012), we model the behavior of referees with each team treated equally in search of a balance of penalty calls. The following section outlines how data was collected to answer these questions. The paper continues by displaying the results and discussing the conclusions drawn based on the incentives of hockey referees.

Theory and Calculation

The nature of hockey officiating creates a desirable laboratory for examining referee behavior. In a given game, hockey referees assign an average of fewer than 10 total penalties. This number is lower than other sports, as basketball referees may assign 25-30 fouls per game and football officials spot 12-15 violations. The greater ease of tracking referee behavior, combined with the greater ability to affect the outcome (due to the severity of the punishment and the closeness of games), provides a strong data set for empirically testing the referee's issuance of penalties.

Our interest lies in measuring NHL referee behavior in both playoff and regular season games. To account for the disproportionate amount of revenue earned by the NHL in the playoffs (Leeds & von Allmen, 2004; Robst et al., 2011), the sample includes all playoff games from 2006 - 2012 (n=599) and a randomly selected group of regular season games (n=450) from the same seasons. The websites www.nhl.com and www.espn.com were used for data collection.

Of primary interest is the number of power-play creating penalties per period. A period based analysis of penalty counts is selected for three reasons. First, we are concerned with the balancing of all penalty calls, including those warranted and those that may be questionable. Perceptions of fairness move beyond debatable calls that go each way, and referees recognize that the totality of the game must be managed sufficiently, not in constant response to a mistaken call. Second, with two intermissions, officials have two opportunities to reflect upon how the game has progressed. While this may not necessarily dictate how referees call games, the breaks provide for analysis over a larger period of time than the previous play-by-play data. Penalties are awarded as the play dictates and referees are likely to be unable and unwilling to alternate power play opportunities. Finally, penalty calls need not be evened up with the same type of penalty. While other studies of aggression measure the type of foul committed, balancing penalty calls need not match a slashing call with a similar slash by an opponent. In collecting data, matching minors or matching fighting penalties are excluded as these do not give one side an advantage and frequently involve lower degrees of judgment.

Our hypothesis is that teams with more penalties called on them early in the game will receive fewer penalties later in the game, both in terms of expected penalties compared to the opposition and the frequency of playing future periods with more penalties. The dependent variables for our analysis are second and third period penalty counts, for each model respectively.

Each model uses a unique set of independent variables. For modeling second period outcomes, information on penalties and score at the conclusion of the first period was used. This includes a variable for absolute goal differential (Tied, 1 goal, 2 goals, 3+ goals), if the team is ahead (Yes/No), and the teams' penalty differential entering the period. For modeling third period data, we use score and penalty statistics from the end of the second period. Postseason fixed effects also include indicators for the 7th game of a series (Yes/No) or if the game was played in the Stanley Cup Finals (Yes/No). For regular season models, a binary variable for attendance (High/Low) is determined by the arena drawing at least 95% of its capacity for a particular game. This variable is excluded from the playoff models due to all playoff games being sold out. Additional fixed effects are included for team and opponent in both populations to adjust for the tendency of certain teams to give or receive higher or lower penalty counts. Each model is fit on the penalty data from all teams, and, as in Buraimo et al. (2009), fits are also estimated separately for home and away teams.

Penalties and score are recorded and tracked by period for the home and road team, and an indicator for referee pairing is included. We compare mean penalties per period given location, goal differential, and penalty differential at the start of the period. Goal differential is defined as the difference in the number of goals each team had scored when the period began. A team's penalty differential is defined as the difference between the number of power-play inducing penalties previously called on that team and their opponent. For example, a second-period penalty differential of three indicates that a team was called for three more penalties than the opposition in first period. Also, the frequency of finishing the second and third periods with higher penalties, given previous penalty differentials, is calculated.

While previous work on penalties per period is scarce, Alan Ryder (2004) suggested hockey outcomes can be approximated by a Poisson distribution. A count variable falls under the Poisson distribution if the observation time is fixed and the events operate independently and at a constant rate over time. In hockey, penalties can be called at all moments of a game, and the per-period length among both our populations is 20 minutes. A Pearson χ^2 test for over- or underdispersion can identify whether a data set appears to be drawn from a Poisson distribution, and is used with our penalties per period information.

A generalized linear mixed effects model for Poisson data (McCulloch & Neuhaus, 2001) is used for four dependent variables, second and third period penalties in both the regular season and the postseason. For periods showing over- or underdispersion, the variance parameter of the Poisson fit is adjusted using a quasi-likelihood fit (Wedderburn, 1974). In all models, we use random intercepts for each referee pairing to account for the fact that the same set of referees preside over multiple NHL games

Table 1: Mean (SE) Penalties per Period

		2nd period		3rd period	
		Postseason	Regular season	Postseason	Regular season
Team	Home	1.44 (0.05)	1.44 (0.05)	1.27 (0.05)	1.17 (0.05)
	Away	1.70 (0.05)	1.52 (0.05)	1.42 (0.05)	1.29 (0.05)
Penalty differential at start of period	Fewer penalties	1.80 (0.06)	1.71 (0.06)	1.44 (0.05)	1.31 (0.05)
	Even penalties	1.63 (0.06)	1.35 (0.06)	1.26 (0.07)	1.07 (0.07)
	More penalties	1.46 (0.06)	1.37 (0.06)	1.30 (0.06)	1.19 (0.06)
Goal differential at start of period	Down 3+ goals	2.43 (0.32)	1.64 (0.27)	1.99 (0.18)	1.54 (0.15)
	Down 2 goals	1.57 (0.11)	1.36 (0.12)	1.54 (0.12)	1.18 (0.11)
	Down 1 goal	1.59 (0.07)	1.44 (0.08)	1.20 (0.08)	1.20 (0.08)
	Tied	1.53 (0.06)	1.49 (0.06)	1.07 (0.06)	1.04 (0.07)
	Ahead 1 goal	1.70 (0.07)	1.40 (0.08)	1.26 (0.07)	1.29 (0.07)
	Ahead 2 goals	1.60 (0.13)	1.68 (0.12)	1.53 (0.10)	1.09 (0.09)
	Ahead 3+ goals	2.20 (0.27)	1.88 (0.21)	1.73 (0.14)	1.48 (0.14)

together. With 147 unique referee pairs in our data, and because we are not interested in the specific behavior of individual referees, this term is not considered as a fixed effect.

Results

Penalty Calls by Period

Mean 2nd and 3rd period penalties per team are shown in Table 1. In postseason games, teams with a higher number of first period penalties finished the second period with fewer penalties nearly twice as often as the team with fewer first period penalties (46% vs. 26%, with the remaining 28% an equal number of 2nd period penalties). This difference is larger for the home team (50% vs. 21%), teams participating in the Stanley Cup Finals (51% vs. 20%), and teams participating in postseason Game 7 (58% vs. 5%).

Teams with more first period penalties in regular season games also finished the second period with fewer penalties more often than their opponents (43% vs. 29%). This difference was slightly higher in the games reaching our 95% attendance cutoff (48% vs. 25%).

Figure 1 shows the effect of an unequal end of first-period penalty differential on mean second period penalties per team. The strongest effect is shown in postseason Game 7s, where teams whistled for more first period penalties receive an average of 0.75 fewer second period penalties.

The ratios of the mean and variance of second period penalties in the postseason and regular season are 0.86 and 0.77, respectively, suggesting Poisson models be adjusted for underdispersion using a quasi-likelihood fit. The ratio of the mean and the variances of third period penalties are 1.09 and 1.02, close enough to 1, as judged by a lack of significance in the Pearson test for dispersion, to fit using a traditional like-

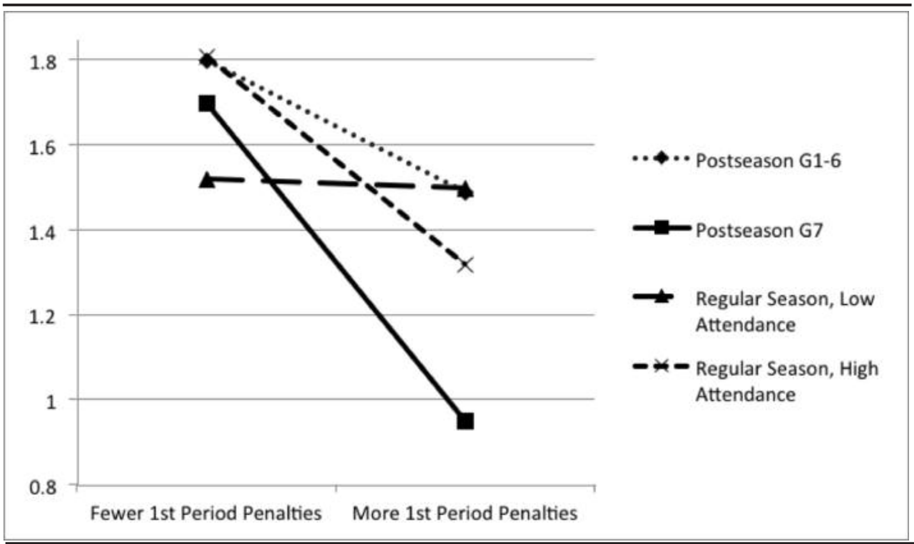

Figure 1: Mean second period penalties by first period penalty differential, game type

Table 2: Estimates (SE) from Generalized Linear Mixed Model (GLMM) of log(2nd Period Penalties) from Postseason Games

	Model		
	Overall (n = 1198)	**Home teams (n = 599)**	**Away teams (n =599)**
Intercept	0.974 (0.100)**	0.879 (0.142)**	0.921 (0.130)**
Penalty differential after P1	-0.075 (0.014)**	-0.062 (0.022)**	-0.097 (0.020)**
Ahead (yes vs. no)	0.009 (0.051)	0.014 (0.078)	0.004 (0.071)
Score differential (1 vs. tied) after P1	0.059 (0.054)	0.057 (0.084)	0.054 (0.072)
Score differential (2 vs. tied) after P1	0.062 (0.070)	-0.009 (0.107)	0.138 (0.093)
Score differential (3+ vs. tied) after P1	0.362 (0.092)**	0.247 (0.143)*	0.461 (0.120)*
Cup finals (yes vs. no)	-0.172 (0.092)*	-0.225 (0.133)*	-0.112 (0.116)
Game 7 (yes vs. no)	-0.183 (0.112)	-0.332 (0.180)*	-0.064 (0.139)
Home team (yes vs. no)	-0.121 (0.040)**	NA	NA

Note: All models include random effects for referee pairing and adjustments for team and opponent. Score differential represents the absolute value of the difference in the score between the two teams at the beginning of the period, with a tie game used as the reference category.
** $p < 0.01$
* $p < 0.10$

lihood approach. The random intercept for referee pairing is significant in all models, as judged by a likelihood ratio test.

Effect estimates from our mixed effect models of second and third period penalty counts, from the postseason and regular season groups, are shown in Tables 2-5 alongside their standard errors. First period penalty differential is a significant predictor for both home (rate ratio 0.94, 95% CI 0.90 - 0.98) and away (Rate Ratio 0.91, 95% CI 0.87 - 0.94) second period postseason penalties (Table 2). In a Poisson regression with explanatory variable X, a rate ratio is estimated by e^{β}, where β is the coefficient on X. In this case, X represents the first period penalty differential, and our estimate suggests that for each increase of 1 in the end of first period penalty differential between a home team and its opponent, that team will be whistled for 6% fewer second period penalties. The effect estimate for series round (Stanley Cup Finals vs. other) is borderline significant (p<0.10), suggesting referees may call fewer penalties when stakes are increased.

Table 3 presents shows estimates from our fit of third period postseason penalties. Teams with more cumulative penalties than their opponents entering the third period of postseason games are called for fewer infractions than their opponents, with this effect slightly stronger for home teams (RR 0.95, 95% CI 0.92-0.99) than away teams (RR 0.97, 95% CI 0.93-1.01). Effect estimates for the absolute goal differential terms

Table 3: Estimates (SE) from Generalized Linear Mixed Model (GLMM) of log(3rd Period Penalties) from Postseason Games

	Model		
	Overall (n = 1198)	**Home teams (n = 599)**	**Away teams (n =599)**
Intercept	0.630 (0.130)**	0.405 (0.178)**	0.686 (0.175)**
Penalty differential after P2	-0.041 (0.013)**	-0.053 (0.019)**	-0.032 (0.019)*
Ahead (yes vs. no)	-0.051 (0.056)	-0.043 (0.079)	-0.022 (0.082)
Score differential (1 vs. tied) after P2	0.123 (0.071)*	0.203 (0.102)*	0.099 (0.101)
Score differential (2 vs. tied) after P2	0.347 (0.079)**	0.350 (0.114)**	0.369 (0.110)**
Score differential (3+ vs. tied) after P2	0.528 (0.080)**	0.554 (0.116)**	0.545 (0.111)**
Cup finals (yes vs. no)	-0.140 (0.113)	-0.277 (0.157)*	-0.035 (0.151)
Game 7 (yes vs. no)	-0.161 (0.132)	-0.074 (0.183)	-0.205 (0.191)
Home team (yes vs. no)	-0.135 (0.050)**	NA	NA

Note: All models include random effects for referee pairing and adjustments for team and opponent.
** $p < 0.01$
* $p < 0.10$

Table 4: Estimates (SE) from Generalized Linear Mixed Model (GLMM) of log(2nd Period Penalties) from Regular Season Games

	Model		
	Overall (n = 900)	**Home teams (n = 450)**	**Away teams (n = 450)**
Intercept	0.426 (0.059)**	0.269 (0.190)	0.600 (0.200)**
Penalty differential after P1	-0.075 (0.018)**	-0.092 (0.026)**	-0.060 (0.031)*
Ahead (yes vs. no)	-0.077 (0.051)	-0.075 (0.073)	-0.053 (0.089)
Score differential (1 vs. tied) after P1	-0.049 (0.056)	-0.046 (0.081)	-0.059 (0.084)
Score differential (2 vs. tied) after P1	0.020 (0.073)	0.029 (0.106)	-0.028 (0.118)
Score differential (3+ vs. tied) after P1	0.172 (0.104)*	-0.003 (0.164)	0.298 (0.153)*
Attendance > 95%	0.037 (0.052)	0.122 (0.101)	-0.023 (0.084)
Home team (Yes vs. No)	-0.055 (0.049)	NA	NA

Note: All models include random effects for referee pairing and adjustments for team and opponent.
** $p < 0.01$
* $p < 0.10$

Table 5: Estimates (SE) from Generalized Linear Mixed Model (GLMM) of log(3nd Period Penalties) from Regular Season Games

	Model		
	Overall (n = 900)	**Home teams (n = 450)**	**Away teams (n = 450)**
Intercept	-0.008 (0.174)	-0.220 (0.233)	0.094 (0.260)
Penalty differential after P2	-0.035 (0.015)*	-0.053 (0.020)**	-0.010 (0.023)
Ahead (yes vs. no)	0.006 (0.063)	0.104 (0.092)	-0.041 (0.17)
Score differential (1 vs. tied) after P2	0.151 (0.088)*	0.054 (0.122)	0.243 (0.131)
Score differential (2 vs. tied) after P2	0.044 (0.098)	-0.035 (0.139)	0.081 (0.141)
Score differential (3+ vs. tied) after P2	0.357 (0.098)**	0.309 (0.139)*	0.370 (0.149)*
Attendance > 95%	-0.035 (0.066)	-0.040 (0.109)	0.007 (0.098)
Home team (yes vs. no)	-0.099 (0.057)*	NA	NA

Note: All models include random effects for referee pairing and adjustments for team and opponent.
** $p < 0.01$
* $p < 0.10$

are large and statistically significant, suggesting games entering the period with a larger absolute score differential result in a significantly higher number of penalties than games tied at the beginning of the period. This is consistent with prior research indicating that fouls of aggression are more common when the opportunity cost of winning the match is lower (Jewell, 2009).

Tables 4 and 5 show model estimates for regular season second and third period penalties, respectively. First period penalty differential is a strong predictor of second period penalties for the home team (RR 0.91, 95% CI 0.87 – 0.96) and a moderate predictor for the away team (RR 0.94, 95% CI 0.88 – 1.00). For third period regular season penalties, a larger penalty differential at the beginning of the period results in a significantly fewer number of calls on the home team (RR 0.95, 95% CI 0.91-0.99) but not for the away team (RR 0.99, 95% CI 0.95-1.04).

In all models, a fixed effect for whether or not the team is playing with the lead entering the period is not a significant predictor, and regular season penalty counts are not noticeably affected by whether or not the game reached the 95% attendance cutoff.

Control Checks

Most penalties are called as a result of a player being put in a defensive position; therefore, different proxies for aggressive behavior were attempted to control for changes in player behavior. One proxy we considered was shots on goal. Aggressive teams may register more offensive possessions and draw more penalties. However, these same teams may receive more penalties in the act of acquiring the puck to begin with, thus negating the likelihood of gaining and advantage through more time on offense. Inclusion of shots per period failed to improve our models or modify the effect of penalty differential, suggesting that an attacking offense does not correlate strongly with penalty outcomes.

While player behaviors do change throughout the game, the most likely changes are tactical in nature. Losing teams may adjust which lines receive more ice time and allow more ice time for strong offensive players. However, this is likely counterbalanced by the leading team's tactical shift of issuing more ice time to strong defenders who are less likely to commit a penalty. This change in tactics, seen through player behavior, reduces the risk that the results are due to player, rather than referee, behaviors. These findings are consistent with prior research on aggression and NHL referee actions (Heckelman & Yates, 2003).

One potential issue with our design is that players and teams may use a less aggressive style of play if they have already received a larger number of penalty calls. In soccer and basketball, for example, league rules enforce additional penalties to repeat in-game offenders and their teams that might force competitors to adjust their style of play. However, other than the penalty itself, hockey rules do not place additional punishment on players or teams for higher frequencies of penalty calls, except in the rare case that a player receives multiple game misconduct penalties for severe infractions. As a result, because they are allowed to continue aggressive behavior throughout the contest, it seems plausible that hockey players are less likely to adjust their aggressiveness compared to athletes in other sports. In our data, the weak correlation between one team's first and second period penalties (-0.03 for postseason, -0.05 for regular season), and a team's cumulative first and second period penalties with their third period penalties (0.00 for postseason, 0.03 for regular season) suggest that team's do not noticeably change their aggressiveness relative to previous infractions.

While Buraimo et al. (2009) used relative goal differential as a predictor of soccer infractions using in-game data, absolute goal differential was used in our models. As shown in Table 1, there does not appear to be a consistent growth of penalty frequency by relative goal differential. In fact, penalty counts appear to vary only with respect to how far the score deviates from a tie game at the beginning of the period. To check this claim, we estimated all of our models using relative goal differential and compared these fits to the ones provided in Tables 2-5 using the Akaike Information Criterion statistic. Inclusion of absolute goal differential, instead of relative goal differential, yielded roughly equivalent or much stronger fit statistics in our models.

Prior research has also suggested that large markets or popular teams may receive preferential treatment by officials (Price et al., 2012). To account for this, we examined the difference in penalty distribution based on presence in hockey's "Original Six," which includes the teams from Boston, Toronto, Chicago, Montreal, Detroit, and the New York Rangers. Inclusion of this variable did not result in a significant improvement to the model.

Finally, to assess a possible association between regular and postseason penalty behavior, we examined the correlation between regular and postseason penalty rank, judged by the ranking of postseason qualifiers in penalties per game. No strong positive correlations between the rankings were found, and in some years, negative correlations were found, suggesting style of play is dictated by mostly by opponents and game strategies that change for the playoffs. In total, these control checks provide some assurance that other factors, such as player behaviors, have not unduly influenced the results.

Discussion

Our findings make several key contributions within the literature on referee behaviors. First, by removing the assumption that referees are continually biased towards home teams, we find that the penalty frequency is determined by penalty differential and score of the game. Second, this model recognizes the psychological desire for fairness inherent in the role of a referee. The model incorporates how bias is implemented and situational factors where one team is more likely to receive a power play. Third, this model is built upon the recognition that referees play a significant role in preserving the integrity of a sporting league, along with the associated financial returns. Finally, this research builds on the work of Buraimo et al. (2009) and Price et al. (2012) by illustrating how officials institute bias in a different sport. Our model demonstrates how and when preferences for fairness as equality are implemented into the sport of hockey. In total, these contributions better our understanding of the human element involved in refereeing.

The hypothesis of make-up calls among NHL referees is enhanced by looking at penalty calls in postseason overtime. In professional hockey, games ending regulation in a tie move to the sudden death format, in which the game ends after the first goal is scored. In 25 of the 134 postseason overtime games since 2006, at least two total penalties were called. In 19 of the contests with multiple penalties (76%), the second power play was awarded to the team that was called for the first penalty. We would expect the second power play to be awarded to the opposite team 50% of the time. However, this sample proportion is significantly different from 0.5, as judged using a 1-sample z-test for proportions ($p < 0.01$). Thus, in the most intense moments of the most important games, the tendency for a reversal of penalties appears to be the strongest.

In our data set, for both regular and postseason and in both second and third periods, teams with more penalties entering the period were called for fewer infractions. The effect of penalty differential was relatively similar comparing regular and postseason play, with stronger effects evident for the home team. Our findings complement the recent work of Mongeon and Mittelhammer (2011), which used continuous flow to suggest that referees use make-up calls to keep games close in the regular season. While play-by-play data may reveal within-period changes in behavior, discretizing by period offers the advantage in that at the end of each period, team personnel is mostly reset. As a result, a reversal in a penalty call is less likely to be effected by the changes in on-ice personnel that occur with each result of a power play.

To consider the possibility that the effect of penalty differential is modified by one of our other fixed effects, we also tested the significance of two-way interaction terms between each fixed effect and penalty differential entering the period. The only two marginally significant interaction terms linked Game 7 status with penalty differential for second period postseason data (Table 6), and games reaching our attendance cutoff with penalty differential in our second period regular season model (Table 7). Combined with Figure 1, the estimates in Table 7 provide moderate evidence that the effect of prior penalty differential is largest when games are played in front of larger

Table 6: Estimates (SE) from Generalized Linear Mixed Model (GLMM) of log(2nd Period Penalties) from Postseason Games, Including Interaction Term

	Model		
	Overall (n = 1198)	Home teams (n = 599)	Away teams (n =599)
Intercept	0.979 (0.100)**	0.879 (0.142)**	0.953 (0.131)**
Penalty differential after P1	-0.070 (0.015)**	-0.062 (0.022)**	-0.097 (0.020)**
Ahead (yes vs. no)	0.009 (0.051)	0.014 (0.078)	0.010 (0.071)
Score differential (1 vs. tied) after P1	0.058 (0.054)	0.057 (0.084)	0.044 (0.072)
Score differential (2 vs. tied) after P1	0.062 (0.070)	-0.009 (0.107)	0.130 (0.093)
Score differential (3+ vs. tied) after P1	0.361 (0.092)**	0.247 (0.143)*	0.451 (0.120)**
Cup Finals (yes vs. no)	-0.169 (0.092)*	-0.225 (0.133)*	-0.099 (0.116)
Game 7 (yes vs. no)	-0.210 (0.112)*	-0.331 (0.181)*	-0.113 (0.145)
Home team (yes vs. no)	-0.120 (0.040)**	NA	NA
Penalty differential after P1*game7	-0.139 (0.082)*	0.007 (0.140)	-0.241 (0.106)*

Note: All models include random effects for referee pairing and adjustments for team and opponent. Score differential represents the absolute value of the difference in the score between the two teams at the beginning of the period, with a tie game used as the reference category.
** $p < 0.01$
* $p < 0.10$

crowds. While we used a binary cutoff of 95% attendance to indicate games that were more highly attended, the p-value for our interaction term between penalty differential and our attendance variable was at least moderately significant ($p < 0.10$) using all integer cutoffs between 90 and 99%.

While this analysis is able to illustrate situations where one team is more likely to be called for a penalty, the impact on the overall outcome of the game is mostly undetermined. In our sample, we looked at the effect of prior penalty differential on the likelihood of each team winning the game. Because teams with more penalties earlier in games are receiving more power play opportunities later in games, it is plausible that such teams are more likely to win these contests. This extends the work of Widmeyer and Birch (1984) by providing a rationale for their observation that aggressive teams early in games are more likely to win.

Among postseason games tied at the end of the first period, the team with more first period penalties was significantly more likely to win the contest (OR 1.20, 95% CI, 1.07-1.33) than the team with fewer first period penalties. In postseason games tied after the second period, we found moderate evidence that the team with higher cumulative first and second period penalties was more likely to win the game (OR 1.13, 95%

Table 7: Estimates (SE) from Generalized Linear Mixed Model (GLMM) of log(2nd Period Penalties) from Regular Season Games, Including Interaction Term

	Model		
	Overall (n = 900)	**Home teams (n = 450)**	**Away teams (n = 450)**
Intercept	0.409 (0.147)**	0.278 (0.202)	0.441 (0.210)*
Penalty differential after P1	-0.021 (0.030)	-0.025 (0.046)	0.026 (0.038)*
Ahead (yes vs. no)	-0.078 (0.052)	-0.077 (0.073)	-0.057 (0.089)
Score differential (1 vs. tied) after P1	-0.055 (0.057)	-0.070 (0.082)	-0.046 (0.083)
Score differential (2 vs. tied) after P1	0.003 (0.074)	0.018 (0.106)	-0.031 (0.113)
Score differential (3+ vs. tied) after P1	0.149 (0.106)	-0.051 (0.166)	0.327 (0.156)*
Attendance > 95%	0.036 (0.056)	0.134 (0.102)	-0.035 (0.084)
Home team (yes vs. no)	-0.062 (0.050)	NA	NA
Penalty differential after P1* attendance > 95%	-0.089 (0.037)*	-0.104 (0.056)*	-0.130 (0.052)*

Note: All models include random effects for referee pairing and adjustments for team, season, and opponent.
** $p < 0.01$
* $p < 0.10$

CI 1.00-1.28). No significant evidence of an increased likelihood of winning existed in regular season games for either period.

Conclusion

Our analysis identifies how referees may issue "even-up" calls over the course of a professional hockey game to achieve perceptions of balance and fairness. In-game analysis finds significant evidence that referees exhibit a form of "biased impartiality" when a team has a negative penalty differential, and there is support that this effect is strongest in games when the referees are under the most pressure. Results also suggest this tendency may be stronger for the home team.

However, there are a number of limitations to the model of officiating. In using only hockey in the sample, the same types of behaviors may not be found in other sports or other industries. The linkage between game situation and referee behavior may not be as prevalent in other settings. Idiosyncratic characteristics unique to hockey may also fail to carry over to other samples. Additionally, the data in this study is limited based on categorization by period, rather than as a continuous flow during the game. Further research that tracks penalty calls by time rather than period may provide further illumination into the actions of referees. Finally, all infractions (roughing, boarding, interference, etc.) were counted the same in this analysis. Referees may penalize specific behaviors with greater regularity when attempting to even up calls. Research into these details may provide additional enlightenment into how referees achieve their biased impartiality.

Beyond the additions mentioned above, future studies building on this work are numerous. Exploring other sports or professions could provide new insight into areas of biased impartiality. While this study only consists of games, a longitudinal study of the NHL could map different changes in league and referee incentives towards developing the game on and off the ice. Extensions of this study may also include how consumers respond to changes in refereeing behavior.

In conclusion, the model derived from this research illustrates how NHL referees even up penalty calls to increase perceptions of fairness. The effect of these calls is to balance scoring opportunities and ultimately, the number of games won or lost by each team. This study is unique in its approach of examining how bias may evolve through the course of a game. In contrast to studies concerned with only the final outcome, this isolation of bias effect provides greater insight into referee behaviors and motivations. Although long debated by spectators, players, and coaches, empirical evidence supports the idea that referees even up penalty calls in the spirit of equal opportunity and financial incentive.

References

Buraimo, B., Forrest, D., & Simmons, R. (2009). The 12th man?: Refereeing bias in English and German soccer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*, 431-449.

Boyko, R. H., Boyko, A. R., & Boyko, M. G. (2007). Referee bias contributes to home advantage in English Premiership football. *Journal of Sports Sciences*, *25*, 1185-1194.

ESPN. (2012). ESPN. Retrieved from www.espn.com

Fraser, K. (May 7, 2011). How often do make-up calls actually happen? Retrieved from http://www.tsn.ca/nhl/story/?id=364984

Fraser, K. (June 13, 2011). Refs are human, but make-up calls don't atone. Retrieved from http://www.tsn.ca/blogs/kerry_fraser/?id=368803

Garicano, L., Palacios-Huerta, I., & Prendergast, C. (2005). Favortism under social pressure. *Review of Economics and Statistics*, *87*, 208-216.

Heckelman, J., & Yates, A. (2003). And a hockey game broke out: Crime and punishment in the NHL. *Economic Inquiry, 41*, 705-712.

Jewell, R. T. (2009). Estimating demand for aggressive play: The case of English premier league football. *International Journal of Sport Finance, 4*, 192-210.

Koyama, M., & Reade, J. (2008). Playing like the home team: an economic investigation into home advantage in football. *International Journal of Sport Finance, 4*, 16-41.

Leeds, M., & von Allmen, P. (2004). *The economics of sports,* 2nd ed. Reading, MA: Addison-Wesley.

McCulloch, C. E., & Neuhaus, J. M. (2001). *Generalized linear mixed models.* Wiley Online Library.

Mongeon, K., & Mittelhammer, R. (2011). Home advantage, close games, and big crowds: economics and existence of rationally biased officiating. Working paper.

Moskowitz, T. J., & Wertheim, L. J. (2011). *Scorecasting: the hidden influences behind how sports are played and games are won*, Crown Business: New York.

Nevill, A. M., Balmer, N. J., & Williams, M. A. (2002). The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise*, *3*, 261-272.

National Hockey League. (2012). National Hockey League. Retrieved from www.nhl.com

Petterson-Lidbom, P., & Priks, M. (2010). Behavior under social pressure: empty Italian stadiums and referee bias. *Economic Letters*, *108*, 212-214.

Price, J., Remer, M., & Stone, D. F. (2012). Subperfect game: Profitable biases of NBA referees. *Journal of Economics and Management Strategy*, *21*, 271-300.

Robst, J., VanGilder, J., Berri, D. J., & Vance, C. (2011). Defense wins championships? The answer from the gridiron. *International Journal of Sport Finance, 6*, 72-84.

Ryder, A. (2004). Poisson Toolbox. *Hockey Analytics.* Retrieved from http://hockeyanalytics.com/Research_files/Poisson_Toolbox.pdf

Scoppa, V. (2008). Are subjective evaluations biased by social factors or connections: An econometric analysis of soccer referee decisions. *Empirical Economics*, *35*: 123-140.

Staples, D. (March 2, 2012). Bettman puts bright and bubbly spin on NHL revenues, *Edmonton Journal.* Retrieved from http://blogs.edmontonjournal.com/2012/03/02/bettman-puts-very-positive-spin-on-nhl-revenues/

Sutter, M. & Kochera, M. G. (2004). Favortism of agents-the case of referees' home bias. *Journal of Economic Psychology*, *25*: 461-469.

Wedderburn, R. W. M. (1974). Generalized linear models specified in terms of constraints. *Journal of the Royal Statistical Society*, *36*, 449-454.

Widmeyer, W., & Birch, J. (1984). Aggression in professional ice hockey: a strategy for success or a reaction to failure? *The Journal of Psychology, 117*, 77-84.

Widmeyer, W., & McGuire, E. (1997). Frequency of competition and aggression in professional ice hockey. *International Journal of Sport Psychology*, *28*, 57-66.