Kevin Snyder* and Michael Lopez

Consistency, accuracy, and fairness: a study of discretionary penalties in the NFL

DOI 10.1515/jqas-2015-0039

Abstract: Prior studies of referee behavior focus on identifying a bias in when certain calls are made [Kovash, Kenneth, & Levitt, Steven (2009). "Professionals do not play minimax: evidence from Major League Baseball and the National Football League (No. w15347)." National Bureau of Economic Research; Rosen, Peter A. and Rick L. Wilson. 2007. "An Analysis of the Defense First Strategy in College Football Overtime Games." *Journal of Quantitative Analysis in Sports* 3(2):1–17; Alamar, Benjamin. 2010. "Measuring Risk in NFL Playcalling." *Journal of Quantitative Analysis in Sports* 6:11.]. We extend this research by evaluating the consistency of specific discretionary penalties in professional football. In doing so, all NFL plays from 2002 to 2012 are considered, isolating the occurrence of holding and pass interference calls. Even after accounting for game and play specific variables, including team characteristics, type of play, and the game's score, we find the likelihood of both penalty types follows a quadratic trend, low at the beginning and ends of the game, but high in the middle. We suggest that these penalties are uniquely called with higher levels of discretion, in an attempt by referees to imply fairness in the flow of the game.

Keywords: football; penalties; referees.

1 Introduction

The central tenets of football refereeing include consistency, accuracy and fairness (Simmons 2011). To prevent dangerous play and enforce the rules of the game, referees are tasked with identifying violations along a variety of player actions. However, the speed of the game and the variability of game action (kicking, running, passing, etc.) create a high number of possible penalties and a challenge to observe all action. While this action occurs, external pressures from fans of the home team (Moskowitz and Wertheim 2011) and a subconscious desire to equal penalty calls (Lopez and Snyder 2013) may lead to perceptions of bias. For a referee, this must be done while accounting for the safety of the players and their own interests of reputation and future opportunities.

Officiating football games requires a high level of discretion. Although a single play may take only a few seconds, a team of referees must track 22 players, each with different abilities and roles in the action. Players are taught to block, tackle, and hit their opponent as physical force serves as a means to achieve the goals of the sport. The difficulty of seeing all players simultaneously creates formal and informal strategies that rely upon straddling the boundary between legal and illegal activities. Further, the Seattle Seahawks employ this tactic as a formal means to challenge the threshold of legal activities and thus, dare the referee to invoke a penalty (Clark and Clegg 2014). Since decisions can change the outcome of the game, referees must quickly and carefully decide which actions cross the threshold of legal play.

The purpose of this paper is to investigate the consistency of referee penalty calls in the National Football League (NFL). While numerous academic studies exist on how referees award penalties, little research distinguishes between penalty type or tracks calls over the length of a contest. In ignoring this aspect of a sport, the current literature lacks explanation for how referees use discretion through specific penalty types. We seek to determine if penalty rates vary by type, time, field position and score differential. By isolating specific types of penalties, we also seek to build a model that provides insight into how referees' judgment changes throughout a game. We attempt to better discern where inefficiencies may exist in football strategy by analyzing differences in referee actions. In doing so, we pose two research questions. First, is the distribution of penalties consistent throughout the course of a football game? Second, are certain types of penalties more consistently called than others?

The paper continues with a review of the NFL and prior research on referees and penalty calls. Continuing further, we explore academic theories of referee consistency and players incentives. Next, we describe our sample and the methods used to conduct our analysis. Finally, the paper concludes with a discussion of the results, limitations, and implications for future research.

***Corresponding author: Kevin Snyder,** Southern New Hampshire University – Sport Management, 2500 North River Road, Manchester, NH 03106, USA, Tel.: +4102182238, e-mail: k.snyder@snhu.edu
Michael Lopez: Skidmore College, Saratoga Springs, NY, USA

2 Background

2.1 The National Football League

As measured through total revenues and television ratings, the National Football League is the most popular sporting league in the United States. The NFL consists of 32 teams, generating a total of approximately $10B in revenue (Schrotenboer 2014). Television ratings for football games are often among the highest rated programs of the year (Deitsch 2014). Despite this popularity, research on referee behaviors is limited compared to other sports with more independent measures of performance and identifiable outcomes. Quantifying referee decisions in football is difficult due to the high number of officials and the differing roles.

Football games are broken down into four quarters with each half beginning with a kickoff from one team to the other. Points are awarded for advancing the ball down the field in a series when a team has four plays to advance the ball a total of ten yards. Once this intermediate goal is achieved, four additional "downs" are given to repeat the process. Defenses are designed to prevent the offense from accomplishing this goal by tackling the ball carrier short of the ten yard milestone.

Officials are assigned to games as groups of seven. Each crew consists of a referee, an umpire, a head linesman, field judge, line judge, side judge, and back judge. Each referee is strategically positioned with a different responsibility for watching areas of the field and the development of the play. Penalties can be assessed by any crew member. Violations are punished by advancing the ball in the opponent's direction based on the severity of the infraction. Minor penalties begin with a 5-yard shift, and build to a 10 and 15-yard penalty disadvantage. Finally, certain fouls are penalized by placing the ball at the spot of the foul, thus providing the possibility of greater amounts of yardage being awarded. In an average NFL game, a team will incur approximately six to seven penalties, totaling about 50 yards (www.nflpenalties.com).

2.2 Penalty types

Football penalties can be categorized based on when they occur in the course of a play. First, penalties can occur before a play begins. Play is stopped when these fouls occur, thus negating a potential advantage in the course of action. These fouls are easily identified and are likely to vary little from one referee to another. Examples of pre-snap penalties include false starts, defensive offside, and substitution infractions. Second, actions that result in a common penalty are seen through fouls that occur during the context of play. This type of penalty may not be inadvertent but is committed through standard contact involved in the sport. For example, offensive or defensive holding, illegal contact and pass interference are all judgment calls based on the degree to which the offending player gains an advantage. Referees use significant amounts of discretion when declaring penalties that occur during a play. The final category of penalty occurs after the play ends and is typically deemed unsportsmanlike. These fouls can be the result of fighting, unnecessary roughness or taunting of the opponent. While some discretion is used in assessing these fouls, they remain relatively rare.

Although players want to avoid being called for a penalty, there may be opportunistic reasons for engaging in actions that specifically violate a rule. Despite the immediate negative consequences associated with a penalty, opportunistic fouls can set the tone for the game or physically intimidate the opponent. This form of risk taking typically falls within the in-game category of penalty. For example, in ice hockey, committing early penalties has been linked to an increase of the penalized team's win probability (Widmeyer and Birch 1984). In this manner, players may commit fouls of opportunism, hoping that the infraction goes unnoticed, but accepting a penalty if caught. Opportunities for these actions may arise from players attempting to determine an official's threshold for calling a penalty.

Using Australian football, other researchers have linked committing fouls to anger and a loss of control (Grange and Kerr 2010). However, the penalties studied here tend to be in the third category of post-play penalties linked to unsportsmanlike conduct. This viewpoint is also supported by scholars who find that incurring excessive penalties significantly impairs a team's opportunity to win (McCaw and Walker 1999). This suggests that opportunistic players seeking an advantage are unlikely to commit these types of infractions.

In assessing penalties committed during the course of play, judgment may be impaired by a number of inherent biases. As a popular topic for research, numerous scholars have found that referees exhibit biases towards the home team in response to crowd behaviors (Boyko, Boyko, and Boyko 2007; Buraimo, Forrest, and Simmons 2010; Pettersson-Lidbom and Priks 2010). Additionally, assessing penalties of red or yellow cards in soccer may be influenced by the home crowd and therefore, favor the home team (Sutter and Kocher 2004; Lane et al. 2006).

Working on behalf of the NFL, referees are expected to call games evenly, but as agency theory suggests, referees may subconsciously take their own desires into account when calling games (Sutter and Kocher 2004). This is consistent with all individuals doing work on behalf of others. In addition to identifying a bias for the home team, there may be a bias for when calls are likely to be partial. Numerous studies on soccer have identified the end of games as an opportunity for officials to extend the match if a home team is losing, thereby allowing them more time to change the outcome (Sutter and Kocher 2004; Boyko, Boyko, and Boyko 2007). Other studies have noted a preference to balance out calls when differences emerge in the awarding of penalties (Lopez and Snyder 2013; Abrevaya and McCulloch 2014). When taken together, these studies suggest that referees may be more likely to demonstrate bias depending on the time and status of the game. We extend these ideas to American football and narrow the search for bias to specific types of penalty calls. We build on this idea of time-based bias and continue by describing the models used to assess how players and officials modify their behavior throughout the game.

3 Methods

3.1 Penalty outcomes

Two factors that make analyzing NFL penalties complex are the varying frequencies with which different penalties occur and the different underlying causes behind each infraction. For example, more than half of the NFL's 50 penalty types (27) were called fewer than 10 times during the 2012 season (nflpenalties.com). Further, with some penalties – for example, "illegal use of the hands" – the NFL's play-by-play data fail to easily identify if the offending team was on offense or defense.

The three most common types of NFL penalties in 2012 were "offensive holding," "false start" and "defensive pass interference," (www.nflpenalties.com.) It is within this subset of violations that we focus our analysis, defining four outcomes, *OHR*, *OHP*, *DPI*, and *PS*, as follows:

OHR: An offensive holding penalty when the offense attempted a running play

OHP: An offensive holding penalty when the offense attempted a passing play

DPI: Defensive pass interference

PSV: Pre-snap violations

Violations included in the *PSV* group include false starts, illegal substitutions, delay of games, offsides, and encroachment.

Taken together, our chosen penalty outcomes all (i) occur often enough, (ii) could conceivably be caused by opportunism or anxiousness on behalf of either the offensive or defensive units (or both), and (iii), involve either high degrees (*OHR, OHP, DPI*) or low degrees (*PSV*) of judgment on behalf of the officials. In this setting, including *PSVs*, which involve little to no discretion on behalf of the officials, acts as a control group for which to compare the frequencies of other more judgment-based violations.

The decision to separate holding by running and passing plays is worth noting. Running plays called earlier in the contest may be highly similar to ones called later in the game, whereas pass plays appear more likely to be dictated by down, distance, and score. An offensive team is substantially less likely to call a riskier pass play when they are up two touchdowns than when they are down two touchdowns, for example. Running plays, meanwhile, appear less likely to be driven by in-game factors. Finally, holding penalties on passing plays, by and large, can only be whistled on the offensive players in charge of protecting the quarterback, which is generally between five and seven players. On running plays, there is a larger sample of players, including tight ends and wide receivers, which can be assessed a penalty. Perhaps as a result of these factors, previous research found a higher rate of offensive holding penalties on running plays (Kitchens 2014).

3.2 Data collection

We extracted data from the website "Advanced NFL Analytics," which aggregated play-by-play information from each NFL game between 2002 and 2012 in a comma-separated value (.csv) file, with a separate file for each season (website: http://archive.advancedfootballanalytics.com/2010/04/play-by-play-data.html).

Our first step in parsing the data was to exclude the types of plays on which *OHR*, *OHP*, *DPI*, or *PSVs* were either unable or unlikely to have occurred. These included all special teams plays and any quarterback kneel downs or spikes, which are inherently different than traditional offensive or defensive snaps. Further, the referees in charge of identifying infractions on special teams plays may be different than the ones doing so on traditional running or passing plays.

Next, we divided all non-*PSV* plays into either designed runs or passes. We included plays listed as "quarterback scrambles" as passes, due to the fact that

the majority of such plays were initially designed with the quarterback wishing to throw the ball. Play and game specific factors such as the score, time remaining, line of scrimmage, down and distance, the offensive and defensive units, and the date of the game were also extracted.

For purposes of assessing the relationships between our penalty outcomes and game conditions, we categorized several variables as follows. First, given that Kitchens (2014) found a higher number of offensive holding penalties assessed after a 2009 rule changed the pre-snap alignment of each crew's umpire, we noted the season in which the game occurred. Next, we wanted to use the difference in each team's score as a proxy for how aggressive an offensive unit would be with its play-calling. Given that a team can score no more than eight points on a given possession, leading by nine points, as far as team strategy, is substantially different than leading by eight. As a result, we created a binary variable for if the game was close or not, based on point differential and time remaining. A play was considered to be part of a close game if it occurred with a score differential within two possessions (16 points or less) in the game's first 45 min of play or one possession (8 points or less) in the final 15 min of regulation. Overtime plays were discarded to keep the sample consistent.

Next, because play calls are likely different based on the line-of-scrimmage, we split the field into three sections based on red-zone regions, the 20-yard region on either side of the goal-line. Lastly, we categorized down and distance as follows. All first down plays were considered equivalent, while second down plays were split into "2nd and long" (7+ yards needed for a first down) or "2nd and short." Given the importance of earning a first down on both 3rd and 4th down plays, we combined them into three categories, "3rd/4th and long," (7+ yards needed for a first down), "3rd/4th and medium," (3–6 yards needed), or "3rd/4th and short."

3.3 Descriptive statistics

Table 1 shows the penalty rate (per 1000 plays) for our four infractions. There are slightly fewer holding and defensive pass interference infractions (between 0.5 and 1.3 penalties per 1000 pass plays) when the offensive team is leading. There is a 50% increase in the likelihood of *OHP* in between the 20-yard lines (11.9 penalties per 1000 plays), relative to plays called with the offensive team within 20-yards of the end-zone (7.9 per 1000). The likelihood of *DPI* is nearly twice as high (15.6 penalties per 1000 plays, compared to 8.5) when the line-of-scrimmage is closer to the end-zone, relative to the opposite end of the field. There are very little, if any, associations between our predictor variables and *PSVs*.

All three discretion-based penalty types are associated with down and distance. *OHR* is called with the highest frequency on "3rd/4th and long" and "3rd/4th and medium," at roughly 24 violations per 1000 plays, as

Table 1: Penalties per 1000 plays.

		Offensive holding	Offensive holding	Defensive pass interference	Pre-snap infractions
		Rush plays	Pass plays	Pass plays	
Offensive team	Away	17.8	11.2	9.9	34.0
	Home	19.2	11.3	11.8	35.4
Season	2002–2009	18.0	11.0	10.6	36.0
	2010–2012	20.1	12.0	11.6	31.2
Offensive team leading?	No	18.0	11.1	10.5	34.6
	Yes	19.3	11.6	11.7	34.8
Game Close?*	No	19.3	12.4	10.7	34.6
	Yes	18.3	11.0	10.8	34.8
Yards from opposing end zone	81–100	18.2	10.6	8.5	34.2
	21–80	19.5	11.9	10.4	34.6
	0–20	14.4	7.9	15.6	35.4
Down/Distance	1st down	19.0	10.8	10.2	35.2
	2nd – long (7+ yds)	20.9	12.1	9.1	34.3
	2nd – short (<7 yds)	17.0	8.9	11.3	34.4
	3rd/4th – long (7+ yds)	22.6	14.7	10.9	34.5
	3rd/4th – medium (3–6 yds)	24.4	10.0	14.3	34.8
	3rd/4th – short (<3 yds)	10.8	7.2	16.2	33.3

*Defined as a 3 possession game (through Q3), a 2 possession game (through 5 min mark of Q4), or a 1 possession game.

opposed to 10.8 infractions per 1000 "3rd/4th and short" plays. *OHP* is also most common on "3rd/4th and long" plays (14.7 per 1000 plays), while *DPI* occurs most often on "3rd/4th and short" passes.

There are slightly more *OHP*, *OHR*, and *DPI* penalties called after the 2009 rule change regarding referee positioning. *OHPs* occurred at a similar rate for the home and away units, but *OHRs* and *DPIs* infractions both occurred more often when the home team was on offense. Overall, the rate of offensive holding on running plays (18.5 per 1000 plays) is about 60% higher than the holding rate on passing plays (11.2 per 1000).

All three discretion-based penalty outcomes appear to be strongly related to game minute, while *PSVs* occur with roughly the same frequency throughout the game. Figure 1 shows penalty rates over the course of the 60-min of regulation for *OHR*, *OHP*, *DPI*, and *PSP* (Figure 1), along with a locally weighted smoothing line and its 95% confidence limits.

Both *OHRs* and *DPIs* appear to follow a negative quartic trend by game minute. This involves relatively lower infraction rates at the beginning and ends of the game, higher peaks midway through the second and third quarters (minutes 24 and 36), along with a dip around halftime (minute 30). *OHPs* appear to follow a negative quadratic trend, with lower rates at the beginning and ends of the game.

Offensive holding calls on running plays appear to show the strongest relationship with game minute. There are roughly twice as many holding penalties called between the 5th and 55th minutes of the game (20.0 per 1000 plays) as there are during the first 5 min of action (10.5). Over the final 5 min of play, the rate of *OHR* violations is 14.4 per 1000 plays. The holding rate on running plays in the second minute of the game and the final minute of the game are both less than 7 per 1000 plays, while the infraction rate is higher than 25 during several game minutes occurring during the game's second and third quarters.

The rates of *OHPs* and *DPIs* are also lower in the game's first 5 min (7.5 and 9.2 per 1000 plays, respectively) when compared to pass plays called between the 5th and 55th minute (11.8, 11.1). The rate of *DPIs* also takes a large dip around halftime, with minutes 29 and 30 averaging 8.5 infractions per 1000 pass plays.

There is no obvious association between game minute and the likelihood of a *PSV*.

3.4 Models of NFL penalties

Several of our NFL infractions appear to vary throughout the course of the game, as judged by referee responses to offensive holding and defensive pass interference

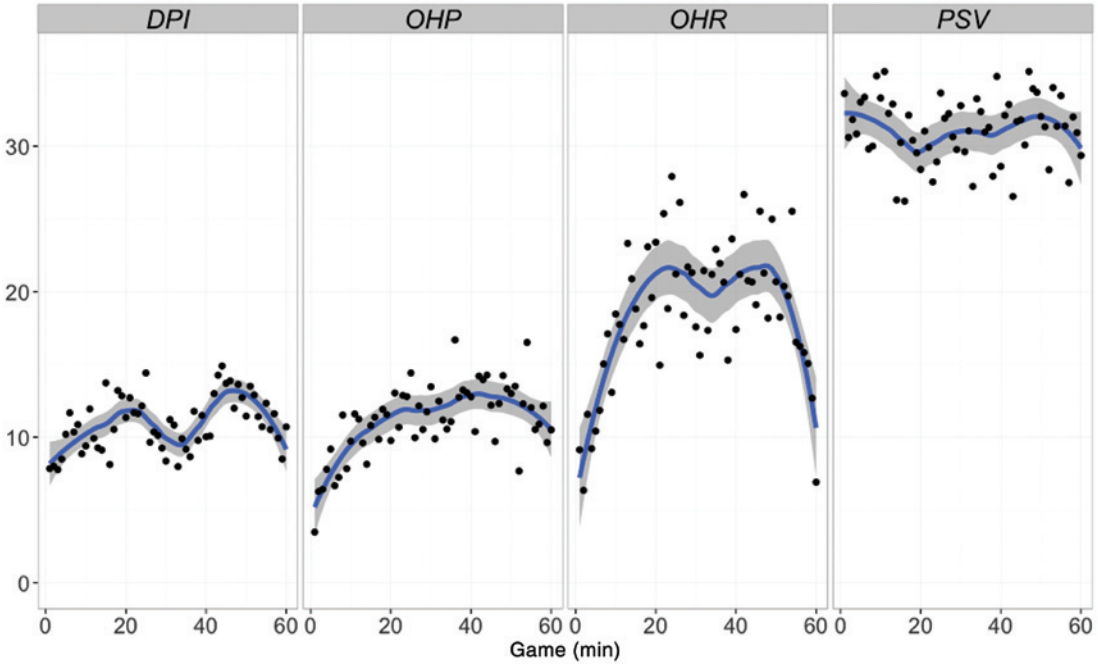


Figure 1: Mean penalty rates (per 1000 plays) for *OHR*, *OHP*, *DPI*, and *PSP* by game minute, along with smoothed fitting lines and 95% confidence intervals.

violations. However, the varying penalty rates may be attributable to changes in the types and locations of plays that also fluctuate over the course of an NFL contest. For example, *DPIs* occur less frequently on plays near a team's own end zone. Given that teams may be more likely to possess the ball in this territory at the beginning of the first and second halves, it is possible that part of the observed time trend of *DPIs* can be attributed to changes in line-of-scrimmage, and not penalty-specific changes. To account for the possibility that extraneous variables are artificially linking game minute and penalty likelihood, we propose a regression modeling framework to account for game, team, and play specific factors that are also associated with penalty outcomes. We model each of the *OHR*, *OHP* and *DPI* infractions using generalized linear mixed models for Binomial outcomes (GLMM). We do not attempt to fit models of *PSVs* given that game and play-specific characteristics, such as game minute and down and distance, do not appear to be associated with the likelihood of these infractions.

GLMM's are fit separately on each offensive play in our data set for *OHR* (using the 150,321 run plays only), *OHP* (the 214,230 pass plays only) and *DPI* (pass plays only) infractions. Let $\boldsymbol{X}_i$ be a set of play and game specific characteristics that we feel may be associated with penalty outcomes of play i, and let δ_j and $\delta_{j'}$ be seasonal-specific random effects attributed to the offensive and defensive teams, respectively. As an example, we present the GLMM of *OHR*, which is fit on the logit transform of $\Pr(OHR_{ijj'} = 1)$, where $OHR_{ijj'}$ is an indicator for whether or not holding was called on offensive team j against defensive team j' on running play i, under the assumption that

$$\begin{aligned}\text{logit}(\Pr(OHR_{ijj'} = 1)) &= \log\left(\frac{\Pr(OHR_{ijj'} = 1)}{1 - \Pr(OHR_{ijj'} = 1)}\right) \\ &= \alpha + \beta^T \boldsymbol{X}_i + \delta_j + \delta_{j'} \text{ with } \delta_j \sim N(0, \tau_1^2), \delta_{j'} \sim N(0, \tau_2^2), \\ &i = 1, \ldots, 150,321, j = 1, \ldots, 352, j' = 1, \ldots, 352\end{aligned}$$

We include unit-specific random effects for each team and season combination (i.e. Indianapolis, 2010) because penalty frequencies within a team's season are likely strongly correlated due to coaching or player specific factors.

Models of each penalty type are adjusted for identical sets of covariates (e.g. $\boldsymbol{X}_i$), including binary terms for season (2009 or before), if the home team is on offense, if the game is close, and if the team on offense is leading. Categorical terms for line of scrimmage (using 81–99 yards from the opposing end zone as a reference) and down and distance (using 1st and 10 as a reference) are also included, and are discretized as in Table 1 (see Sections 3.2, 3.3). Due to the approximately quartic trend shown in Figure 1, we use four terms for game minute, *Minute*, $Minute^2$, $Minute^3$, and $Minute^4$ to model *OHRs* and *DPIs*. In models of *OHPs*, we only use *Minute* and $Minute^2$, due to the quadratic nature of the relationship shown in Figure 1. To reduce the collinearity between our minute terms, each one is centered about minute 30 before being taken to its power, and to ensure convergence, that term is then divided by 10,000. Models were fit using the *lme4* package in R (Bates, Maechler, and Bolker 2012).

4 Results and implications

4.1 Model estimates

The parameter estimates for $Minute^4$ are significant in fits of both *OHRs* and *DPIs* at the 1% level, suggesting plausible quartic associations between those penalties with high degrees of judgment and game minute (see Tables 2 and 3 below).

The estimated coefficient on the adjusted log-odds of $Minute^4$ is furthest from 0 on fits of *OHR*s ($p < 0.001$). The estimated coefficient on the adjusted log-odds of $Minute^2$ in the fit of *OHP*s is also significant ($p < 0.001$).

Due to the difficulty of interpreting coefficient estimates on quartic, cubic, and quadratic terms for game minute, Figure 2 presents the adjusted probabilities of each penalty outcome along with 95% confidence intervals, calculated by game minute. In Figure 2, we used plays called on 1st and 10, in between red zone regions, with the away team on offense, in close games prior to 2009, and the offensive team either tied or trailing, as our baseline, while assuming null offensive and defensive unit random effects. After adjusting for all the game factors in our models, it is clear the strongest time trend over the course of a football game occurs with offensive holding on running plays, among the three outcomes we modeled. The adjusted *OHR* rate jumps from 6.5 penalties per 1000 plays (minute 1), to 20.3 (minute 20) and 21.5 (minute 47), with a slight dip to 19.9 (minute 31). By game's end, the adjusted *OHR* rate drops back down to 11.4.

Offensive holding on passing plays peaks around minute 36 (adjusted penalty rate, 11.9 per 1000 plays, 95% CI 10.7–13.3), with a large dip at the beginning (6.1 per 1000 plays) and a slight drop at the end (9.5 per 1000 plays) of the game. The quartic trend of *DPI*s yields highest estimated adjusted penalty rates at minutes 13 and 50, with

Table 2: Log-odds ratios (95% CI) from generalized linear mixed models of offensive holding penalties.

		Rush plays (n = 148,225)	Pass plays (n = 210,151)
Intercept		−4.03 (−4.24, −3.82)**	−4.55 (−4.74, −4.36)**
Offensive team	Home	0.09 (0.01, 0.16)*	0.01 (−0.07, 0.10)
Season	2010–2012	0.12 (0.02, 0.22)**	0.10 (−0.03, 0.21)
Offensive team leading?	Yes	−0.04 (−0.12, 0.05)	−0.03 (−0.13, 0.06)
Game Close	Yes	0.00 (−0.10, 0.10)	−0.05 (−0.16, 0.06)
Yds from end zone (reference is 81–100)	21–80	0.13 (0.00, 0.26)*	0.15 (0.00, 0.30)*
	0–20	−0.16 (−0.32, −0.00)*	−0.16 (−0.43, −0.03)
Down/Distance (reference is 1st down)	2nd – long (7+ yds)	0.11 (0.01, 0.20)*	0.12 (0.02, 0.23)*
	2nd – short (<7 yds)	−0.10 (−0.21, 0.01)	−0.16 (−0.33, 0.01)
	3rd/4th – long (7+ yds)	0.24 (0.03, 0.46)	0.32 (0.22, 0.44)**
	3rd/4th – medium (3–6 yds)	0.36 (0.09, 0.63)**	−0.02 (−0.19, 0.14)
	3rd/4th – short (<3 yds)	−0.54 (−0.72, −0.33)**	−0.36 (−0.58, −0.14)**
Minute[a]	*Minute*^{1}	−1.04 (−27.1, 26.1)	80.83 (51.74, 109.92)**
	Minute^2	5.61 (0.24, 10.96)*	−4.84 (−6.46, −3.22)**
	Minute^3	0.14 (0.07, 0.21)**	N/A
		−0.018	
	Minute^4	(−0.025, −0.011)**	N/A

[a]Each game minute is centered (across minute 30), and divided by 1000 to ensure convergence.
*signif @ 0.05.
**signif @ 0.01.

Table 3: Log-odds ratios (95% CI) from generalized linear mixed models of defensive pass interference penalties.

		Pass plays (n = 210,151)
Intercept		−5.11 (−5.33, −4.89)**
Offensive team	Home	0.17 (0.09, 0.25)**
Season	2010–2012	0.10 (−0.01, 0.21)
Offensive team leading?	Yes	0.03 (−0.07, 0.12)
Game Close	Yes	0.10 (−0.01, 0.21)*
Yds from end zone	21–80	0.19 (0.03, 0.35)*
	0–20	0.55 (0.36, 0.74)**
Down/Distance	2nd – long (7+ yds)	−0.12 (−0.24, −0.00)*
	2nd – short (<7 yds)	0.06 (−0.09, 0.21)
	3rd/4th – long (7+ yds)	0.06 (−0.06, 0.18)
	3rd/4th – medium (3–6 yds)	0.30 (0.16, 0.44)**
	3rd/4th – short (<3 yds)	0.41 (0.25, 0.57)**
Minute[a]	*Minute*^{1}	37.51 (24.88, 49.73)**
	Minute^2	9.95 (4.65, 15.25)**
	Minute^3	0.03 (−0.02, 0.8)
	Minute^4	−0.014 (−0.021, −0.007)**

[a]Each game minute is centered (across minute 30), and divided by 1000 to ensure convergence.
*signif @ 0.05.
**signif @ 0.01.

lower infraction rates at the beginning, middle, and ends of the game.

Among our other covariates, the home team being on offense led to more *OHRs* and *DPIs*. For example, the adjusted odds of a defensive pass interference penalty are 20% higher when the home team is on offense (OR 1.19, 95% CI, 1.09–1.28). While there was not a significant change in the odds of *OHPs* or *DPIs* after the 2009 rule change in umpire positioning, there is evidence that holding penalty rates increased (OR 1.13, 95% CI 1.02–1.25) on run plays. Whether or not the offensive team is leading, and whether or not the game is close, for the most part, do not appear to be strong indicators of discretionary penalty rates.

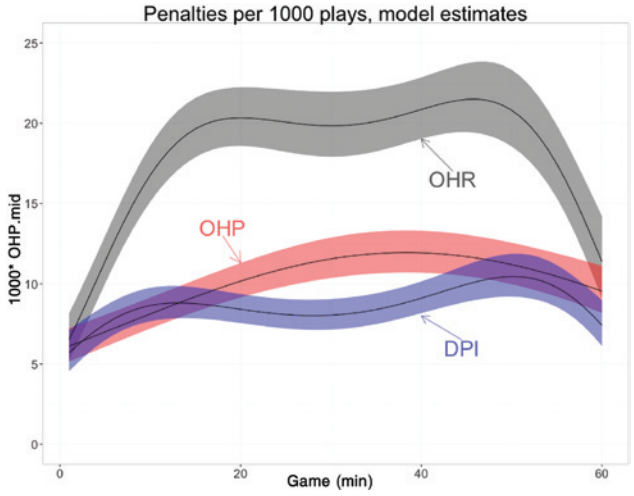


Figure 2: Model estimated penalty rates per 1000 plays.

Down and distance, along with the line of scrimmage, appear strongly associated with our penalty outcomes, with many of the relationships following intuition. For example, there is a substantially higher likelihood of a defensive pass interference penalty when the offense moves towards its' opponents end zone, as opposed to coming out of its own end zone (OR 1.73, 95% CI 1.43–2.10). Red zone passes thrown into the end zone, for example, are often "jump-balls," which likely yield a higher infraction rate.

Third and 4th down plays yield interesting associations between our penalty outcomes. Relative to "1st and 10" plays, *OHP* odds were significantly higher on "3rd/4th and long" plays (OR 1.38, 95% CI 1.25–1.55), but significantly lower on "3rd/4th and short" ones (OR 0.70, 95% CI 0.56–0.87). Interestingly, *DPI*'s occurred at higher rates on "3rd/4th and medium" and "3rd/4th and short" plays but not on "3rd/4th and long" ones, relative to "1st and 10" plays.

The random effect term with the highest estimated standard deviation, and thus the penalty outcome with the highest estimated team-by-year variability, occurred among the offensive team in the model of *OHP*s (estimated τ_1, 0.25). The lowest estimated standard deviation was among the defensive unit in our *OHR* model (estimated τ_2, 0.13). These results follow intuition; a team's offensive personnel likely has a large effect on preventing holding calls on passing plays, but defensive units, as far as drawing holding infractions on running plays, are likely less variant.

While there is no way of knowing if our models best describe the relationship between $\boldsymbol{X}$ and the penalty outcomes, we compared our results to alternative model fits using the Bayesian Information Criterion (BIC). The BIC is a useful tool for trying to identify the "best" model, where the best model is one that has both less error and is parsimonious in the number of predictors. Models with a smaller BIC are considered better.

For each penalty, we explored the interactions between *Game Close* and *Leading*, as it seemed plausible that the effects of game score on penalty outcomes would differ by whether or not the team leading was on offense or defense. In all three fits, the interaction term was not significant ($p > 0.05$), and the full model, which includes this interaction term, yielded a higher Bayesian Information Criterion (BIC).

To account for a possible lack of stationarity over the several years that our data encompass, we also fit models using fixed effects for each year. For all three outcomes, the BIC was higher in the full model than in the reduced models, where the reduced models only included a term to represent if the play occurred after the 2009 rule change. Specifically, the BIC dropped from 27,351 to 27,287 for *OHR*, from 26,078 to 25,976 for *OHP*, and from 25,292 to 25,215 for DPI when using the reduced fits, implying that the trade-off in reducing error was not worth increasing the number of parameters.

We also fit a model of *OHP* that included quartic and cubic terms for game minute, as to compare with the quadratic model shown. The quadratic model appeared to fit better, as judged by a lower BIC (25,976, compared to 25,994).

Finally, regression models can yield biased estimates when there is insufficient covariate overlap. Specific to our study, one worry could be that bad teams, which are often trailing at the end of the game, may be overrepresented as the offensive team on passing plays and the defensive team on running plays at the ends of games. Under this scenario, extending our results to represent all teams may be extrapolating. As one check, we looked at the fractions of plays that each offense spent with a lead of at least nine points, which varied between no plays (6 teams) and 39% (New England, 2007). In eliminating teams in the lower (less than 7% of offensive plays with a nine-point lead or higher) and upper (more than 20%) quartiles, we aimed to identify a group of offenses that were likely closer to the overall league average. Per-minute penalty rates using games played with these teams on offense can be found in the appendix (Figure 3); results are similar to those shown in Figure 1.

5 Discussion

We propose that the likelihood of a certain penalty calls changes significantly throughout the game based on

game and play conditions. Additionally, we find multiple penalty types that are less likely to occur based on the status of the game. These results expand on the prior research that has considered limited ways in which referee bias can be identified.

As an answer to one of our research questions, we find that judgment penalty calls vary considerably over the course of a game, particularly with regards to offensive and defensive holding, as well as defensive pass interference, while pre-snap violations occur with relatively the same frequency. Further, based on the distribution of penalty calls, we provide evidence of statistically significant reductions of penalty calls at the beginning and end of games. Using generalized linear mixed effects models, we propose a quartic model that suggests a "breaking in" process at the beginning of the game and the possibility of officials ignoring all but the most egregious penalties at the end of games. While our model implies a quartic trend by treating time as continuous, is important to note because of halftime, incentives for teams may vary between minutes 30 and 31. However, relative to the rates in other minutes between minutes 25 and 35 of a game, divergences in the penalty rates just before and after halftime are difficult to distinguish in Figure 1. An alternative possibility is that local maxima of penalty rates occur somewhere in the middle of each half.

Understanding these results in context of each other and the incentives of the game contribute to the broader literature of refereeing. While other research implies that committing penalties can have both negative and positive outcomes (Widmeyer and Birch 1984; McCaw and Walker 1999), we suggest that both teams could benefit from increasing risk taking behaviors early in games to capitalize on the decreased likelihood of being whistled for a foul. Although many observers and scholars have noted the decline in penalty calls late in a game, we are able to isolate specific categories of penalty calls that are more likely to be ignored, and the resulting relative decreases in call frequencies.

Specifically, we find the biggest discrepancy in offensive holding calls initiated during the beginning and end of games. The rate of offensive holding penalties occurring in the first 5 min of a game is approximately half that of the rate called during minutes 5 through 55, and, even after adjusting for the play's line of scrimmage and down and distance, these violations happen about four times as often on plays midway through the second quarter when compared to the game's first 2 min. This is likely due to factors from both referees and players. Referees may issue warnings early in games rather than calling penalties. However, an equally likely explanation is that players conserve energy while feeling out opponents early in games, thus foregoing an opportunity to take a risk on borderline infractions. Interestingly, the number of pre-snap penalties remains relatively stable throughout the entire game.

Despite committing fewer penalties, further analysis reveals that teams have consistent success in running or stopping the run during the game's opening 5 min. A team's yards per carry, for example, is identical between the first 5 min of the game (4.30) and between minute 6 and 55 (4.30). The consistency of these metrics implies that the types of running plays called early in the game are similar to ones called later in the game, further isolating referee behavior as the cause of the lack of consistency in penalty rates.

Similar patterns emerge for defensive pass interference calls, as these penalties are called significantly less often in the beginning and ends of the game. Perhaps given this additional freedom, these scenarios may provide players with an opportunity to opportunistically engage in pushing the boundaries of the rules. While prior research has focused on the impact of referee bias on outcomes of games, we extend this line of inquiry by including American football and find similar inconsistency in the beginning of games. Rather than favoring a home team, this bias rewards teams that engage in opportunistic behaviors that test the limits of what infractions will be called.

These outcomes have numerous implications for future research. Our models build on theories of referee bias and incentives. In any sport, many penalties are obvious, require little judgment and are called with near perfect consistency. By isolating specific calls where judgment is and is not needed, we illustrate how referees' judgment fluctuates throughout a game. Future research on other sports, such as hockey or basketball, may suggest new avenues for understanding how referees do their job. Further, additional research is needed to link in-game strategy and the degree to which players respond to conservative play calls. For example, how do penalty calls change when coaches shift to a more conservative strategy? Are certain players or teams more successful at exploiting the first and last 5 min of a game to increase their odds of winning? These questions can provide a more holistic view of the behavior of all participants in a sporting contest.

One unique feature of football is the ability of offenses to script, or pre-determine, which plays will be used at the beginning of a game. As teams are able to practice this sequence of plays prior to the game, greater familiarity may lead to a decreased likelihood of being out of position,

resulting in a player committing a penalty. However, the ability to script plays is limited to the offense. Defenses must adjust to the personnel and formations implemented by the offense. Similar patterns of penalties are found in defensive player behaviors, suggesting that both teams may participate in a process of familiarization with the opponent, seeking for a more favorable time to attack. Further, the negative repercussions of a severe penalty can still be overcome if enforced early rather than later in a game. Although this study does not focus on player psychology, these reasons may provide explanation for low levels of penalties early in games.

While teams may have difficulty in coaching players on the appropriate amount of contact, this research is an initial foundation into understanding how football games are officiated. Further investigation into patterns of specific referees may lead teams to optimize play calling and coaching of aggressive behaviors. Combined with film study, coaches may be able to identify specific player actions that are called differently throughout the game. This would help players and coaches optimize the level of aggressiveness. As an obvious practical contribution, highlighting time differences in how penalties are called at the beginning and end of games gives coaches a better awareness of when they can expect an increased or decreased rate of penalties. Such insight could also drive interactions between coaches and officials; lobbying for a call at the beginning and ends of a game, for example, may be more of a longshot.

Finally, referees are often criticized for calling games differently towards the end of games. However, these actions are largely accepted by both teams because participants implicitly understand how games will be called. As participants in the sport, athletes want to feel that the outcome was the result of their actions, rather than an ill-timed penalty for or against their team.

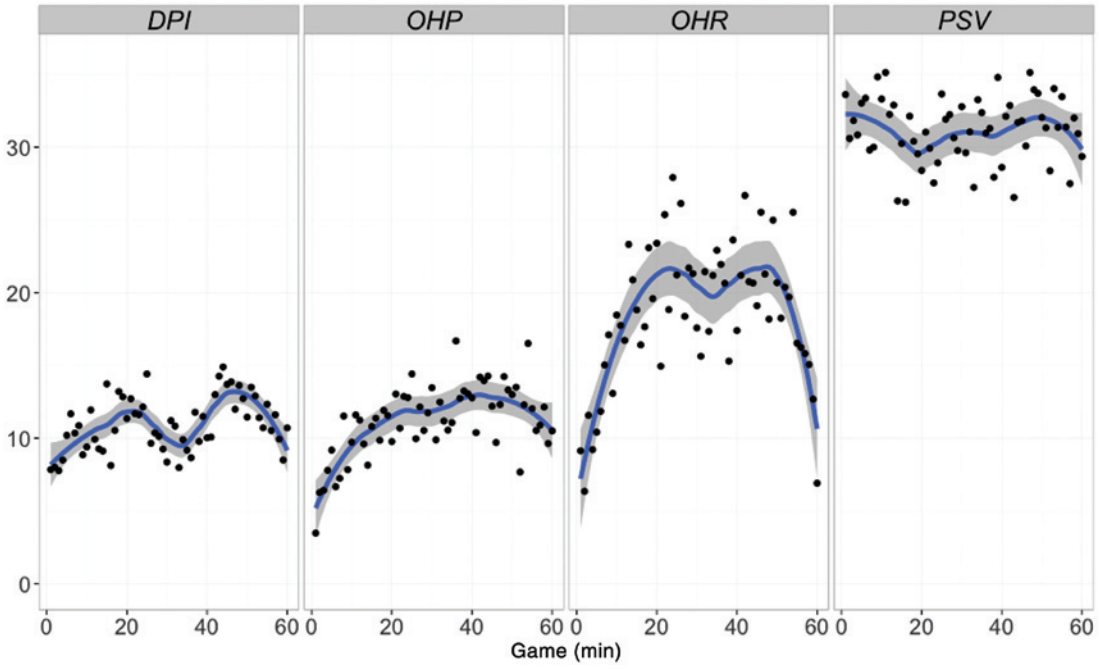
6 Conclusion

Football remains a fruitful, yet underdeveloped area of research for player and referee actions. We attempt to advance the literature by modeling penalty calls to understand changes in referee judgment and consistency. Our results suggest a quartic or quadratic model of penalties across the four quarters of a football game. Lower numbers of penalty calls early in games could be attributed to a "feeling out" process, while lower penalties late in games may be associated with referees attempting to allow players to more directly influence the outcome of close games.

However, our models have numerous limitations. First, given that our data are observational, estimates from our models should not be considered causal effects. Instead, our practical conclusion is only that, after accounting for the game and play characteristics that vary in an NFL contest, penalty rates vary substantially. Second, data for this research was restricted to professional football. Results may not extrapolate to other sports with differing contexts and idiosyncrasies. Further, our measure of judgment is limited to situations where players have crossed the line of accepted actions. Other types of judgment, such as the ruling of a catch or boundary plays, may represent different patterns. Finally, the collection of our data is limited to what can be directly captured and measured. Although our analysis sought to identify penalties committed during a play, post-play judgment is often used to assess unsportsmanlike penalties. Only action occurring between the whistles is considered in this study.

This paper builds upon the work of referee analysis. Our model demonstrates two primary phenomena – a tendency of referees, through specific penalty types, to withhold penalty calls early and ignore borderline infractions late in games. Explanations for these findings are found in the incentives of referees to dictate the threshold of allowable actions and to avoid appearing to direct impact the outcome of a game. By identifying times of the game where referees incentives change, we are able to isolate instances of inconsistently called penalties. We hope that this work serves as inspiration for others to explore how coaches, players, and referees concurrently shift their behaviors to achieve a desired outcome.

Appendix



Appendix 1: Mean penalty rates (per 1000 plays) for *OHR*, *OHP*, *DPI*, and *PSP* by game minute with smoothed fitting lines and 95% confidence intervals (for league average teams).

References

Abrevaya, Jason and Robert McCulloch. 2014. "Reversal of Fortune: A Statistical Analysis of Penalty Calls in the National Hockey League." *Journal of Quantitative Analysis in Sports* 1–50.

Alamar, Benjamin. 2010. "Measuring Risk in NFL Playcalling." *Journal of Quantitative Analysis in Sports* 6:11.

Bates, Douglas, Martin Maechler, and Ben Bolker. 2012. "lme4: Linear mixed-effects models using S4 classes.".

Boyko, Ryan H., Adam R. Boyko, and Mark G. Boyko. 2007. "Referee Bias Contributes to Home Advantage in English Premiership Football." *Journal of Sports Sciences* 25(11):1185–1194.

Buraimo, Babatunde, David Forrest, and Robert Simmons. 2010. "The 12th Man?: Refereeing Bias in English and German Soccer." *Journal of the Royal Statistical Society: Series A* (*Statistics in Society*) 173(2):431–449.

Clark, K. and J. Clegg. 2015. "The Seahawks Grabby Talons." *Wall Street Journal*, January 10, 2014. Accessed March 4, 2015 (http://www.wsj.com/articles/SB10001424052702303754404579310500005285822).

Deitsch, Richard. 2014. "An NFL Ratings Bonanza." *Sports Illustrated*, January 8, 2014. Accessed July 29, 2014 (http://mmqb.si.com/2014/01/08/nfl-tv-ratings-nbc-cbs-espn-fox-playoffs/).

Grange, Pippa, and John H. Kerr. 2010. "Physical Aggression in Australian Football: A Qualitative Study of Elite Athletes." *Psychology of Sport and Exercise* 11(1):36–43.

Kitchens, Carl. 2014. "Identifying Changes In The Spatial Distribution Of Crime: Evidence From A Referee Experiment In The National Football League."*Economic Inquiry* 52(1):259–268.

Kovash, Kenneth, and Levitt, Steven. 2009. "Professionals do not play minimax: evidence from Major League Baseball and the National Football League (No. w15347)." National Bureau of Economic Research.

Lane, Andrew M., Alan M. Nevill, Nahid S. Ahmad, and Nigel Balmer. 2006. "Soccer Referee Decision-Making: 'Shall I Blow the Whistle?'." *Journal of Sports Science & Medicine* 5(2):243.

Lopez, Michael J. and Kevin Snyder. 2013. "Biased Impartiality Among National Hockey League Referees." *International Journal of Sport Finance* 8(3):208–223.

McCaw, Steven T. and John D. Walker. 1999. "Winning the Stanley Cup Final Series is Related to Incurring Fewer Penalties for Violent Behavior." *Texas Medicine* 95(4):66–69.

Moskowitz, Tobias and L. Jon Wertheim. 2011. *Scorecasting: The Hidden Influences Behind How Sports are Played and Games are Won*. New York: Random House LLC.

Pettersson-Lidbom, Per, and Mikael Priks. 2010. "Behavior Under Social Pressure: Empty Italian Stadiums and Referee Bias." *Economics Letters* 108(2):212–214.

Rosen, Peter A. and Rick L. Wilson. 2007. "An Analysis of the Defense First Strategy in College Football Overtime Games." *Journal of Quantitative Analysis in Sports* 3(2):1–17.

Schrotenboer, Brent. 2014. "NFL Takes Aim at $25 Billion, but at What Cost?" *USA Today*, February, 5. 2014.

Accessed July 29, 2014 (http://www.usatoday.com/story/sports/nfl/super/2014/01/30/super-bowl-nfl-revenue-denver-broncos-seattle-seahawks/5061197/).

Simmons, P. 2011. "Competent, Dependable and Respectful: Football Refereeing as a Model for Communicating Fairness." *Ethical Space* 8(3/4):33–42.

Sutter, Matthias and Martin G. Kocher. 2004. "Favoritism of Agents–the Case of Referees' Home Bias." *Journal of Economic Psychology* 25(4):461–469.

Widmeyer, W. Neil and Birch, Jack. (1984). "Aggression in professional ice hockey: A strategy for success or a reaction to failure?" *The Journal of Psychology* 117(1):77–84.