Economic Inquiry

PERSUADED UNDER PRESSURE: EVIDENCE FROM THE NATIONAL FOOTBALL LEAGUE

MICHAEL J. LOPEZ*

We exploit a natural experiment within each National Football League game, finding the first evidence in professional sports that referees succumb to the pressures of satisfying team personnel in the vicinity of possible violations. Using generalized additive models for binomial outcomes, we show that these sideline-based differences in penalty rates, which are observed on common but influential penalties including pass interference and holding, peak near the centralized location of players and coaches on the sideline. With sizable interests in referee decisions, coaches and players often try to manipulate referee behavior with verbal and nonverbal communications; such actions appear to be persuasive. (*JEL* ZO, H3)

I. INTRODUCTION

Sports has proven to be a fertile ground for testing models of corruption, discrimination, crime, incentives, supervision, and performance (Garicano, Palacios-Huerta, and Prendergast 2005). In particular, the study of sport referees (Dohmen and Sauermann 2015) has provided a glimpse into how certain competitive and social factors affect human behavior, theories that otherwise are difficult to empirically test.

As examples, referee decision making has been shown to be associated with player characteristics such as race, size, and stature (Gift and Rodenberg 2014; Mills 2014; Pope and Pope 2015; Price and Wolfers 2007), changes to the number of referees (Heckelman and Yates 2003), and their positioning during play (Kitchens 2014), as well as their own previous judgmental decisions (Abrevaya and McCulloch 2014; Gift 2015; Lopez and Snyder 2013). In addition, a referee pressure to support the home team has also been studied extensively (Boyko, Boyko, and Boyko 2007; Buraimo, Forrest, and Simmons 2010; Dohmen 2008; Garicano, Palacios-Huerta, and Prendergast 2005; Moskowitz and Wertheim 2011; Pettersson-Lidbom and Priks 2010; Sutter and Kocher 2004), as it has been identified that crowd noise is a cue that informs referee decision making (Buraimo, Forrest, and Simmons 2010; Nevill, Balmer, and Williams 2002). Other social pressures on referees, including whether or not the contest is on television (Lane et al. 2006), and the choice to make fewer calls at game's end so as to avoid being part of a game's narrative (Moskowitz and Wertheim 2011; Snyder and Lopez 2015), have also been suggested. And although their behavior is susceptible to outside factors, the monitoring of referees has been shown to reduce bias (Parsons et al. 2011; Pope, Price, and Wolfers 2013).

One aspect missing in the literature, but a part of the flow of any athletic event, is accounting for the pressure and monitoring applied on referees by team employees. Formally, coaches and players admit that "working the referees," which includes acts of kindness and reverence as well as screaming during times of frustration, is part of a game plan (Abrams 2008). Although there is some evidence that suggests referees can be tricked into favoring either team (Petchesky

ABBREVIATIONS

AA: Armchair Analysis
AIC: Akaike Information Criterion
CI: Confidence Interval
FO: Football Outsiders
GAM: Generalized Additive Model
IIA: Independence of Irrelevant Alternatives
LOS: Line of Scrimmage
NFL: National Football League

*Many thanks to Skidmore College's Cody Couture, panelists in the 2015 Joint Statistical Meetings session on referee behavior, as well as the anonymous referees and journal co-editor for their contributions and insights.

Lopez: Assistant Professor of Statistics, Department of Mathematics, Skidmore College, Saratoga Springs, NY 12866, Phone 518-580-5297, Fax 518-580-5295, E-mail mlopez1@skidmore.edu

Economic Inquiry
(ISSN 0095-2583)
doi:10.1111/ecin.12341

2014), and that certain positions use their stature to gain leverage (Mills 2014), it is unknown to date if the immense and constant pressure placed on referees by players and coaches in all sports alters or impairs referee judgment.

If the pressure to satisfy personnel on the nearest sideline exists, it would link closely to established supervisor-subordinate theory. First, there's a rent-seeking nature to the interaction between referees and team personnel, which manifests itself when coaches and players beg for favorable decisions without reciprocation. Second, knowing that they are constantly monitored, referees could act as risk adverse agents (Prendergast and Topel 1993b). The high stakes nature of professional sport creates potentially distorted incentives for referees, who, although tasked with the role of making impartial decisions, may instead adhere to more socially acceptable standards of rule enforcement. As one theory, providing favorable decisions to satisfy nearby coaches and players would be an understandable response to the fear of a coaches' retribution to media after the game, which could jeopardize referee promotion and reputation (Boeri and Severgnini 2011). Finally, it is reasonable to view referees as supervisors who are also judges, in which case proximity effects come into play in a managerial context. While previous work has linked both the size and relative location of the crowd to referee choices (Buraimo, Forrest, and Simmons 2010), fans have low levels of one-to-one interaction with referees. Meanwhile, the constant and often aggressive correspondence between coaches and referees happens on a personal level, in which case the close proximity could be responsible for an increase in favorable biases (Judge and Ferris 1993).

This article presents the first evidence of a successful sideline pressure in sports. Using data from the National Football League (NFL), we identify four sets of penalties in which referees are forced to make difficult decisions in the presence of one team's sideline. With both standard and advanced modeling strategies, we find significant evidence that in the presence or surrounding of a particular team's coaches and players, referees are more likely to call one of several penalties on that team's opponent. Furthermore, rates of offensive holding, defensive pass interference, and aggressive defensive infractions, including personal fouls and unnecessary roughness, peak near midfield, but only on one team's sideline. Given that this is the most likely spot for team personnel to influence referees one way or the other, we posit that referee decision making is strongly impacted by a pressure to appease the nearest stake-holders. Finally, using the observed differences in penalty rates, we explore the practical significance of our findings on play and game outcomes, while comparing our effect sizes to other known referee biases.

II. DATA

For the first and third quarters of an NFL game, coin tosses are used to assign teams end zones to defend. At the end of each of these 15 minutes of play, teams flip-flop sides for the second and fourth quarters. Such a set-up ensures that each team plays 30 minutes in each game moving toward and defending both end zones. Assuming that teams call plays independently of sideline location, the first and third quarter side changes make for a natural experiment that happens twice within every game. Because team benches remain on the same sideline for the duration of each contest, the flip-flopping of directions ensures that, on average, each team will run about the same number and type of play toward each team's sideline by game's end. As a result, we test if referees are influenced by pressure to appease the nearest coaches and players by contrasting penalty rates based on finishing sideline.

Armchair Analysis (AA, www.armchairanalysis.com), a website that matches the NFL's official play-by-play data to game-, team-, and play-specific traits, provided play- and game-specific characteristics for each regular season play between the 2010 and 2014 seasons. This included the offensive and defensive units, line of scrimmage, down, distance, score, outcome, directions for both runs and passes, as well as each game's stadium.[1] Directions for runs and passes are coded as left, middle, or right.[2] We exclude special teams plays, including kickoff and punt returns, because NFL game reports, and thus the AA data, do not label special teams plays with a direction.

End zone directional information—that is, the direction that the offensive and defensive units are facing—is missing from both AA's data and traditional NFL play-by-play reports. We used

1. AA claims a 99.8% accuracy rate with respect to game, team, and player statistics tracked by the league itself.

2. Pass plays are also categorized as either "deep" or "short." We characterize any run play listed as "right end" or "left end" as an outside rush.

coin toss data courtesy of Football Outsiders (FO, www.footballoutsiders.com), a website that provides advanced statistical analysis of American football, which extracted this information from postgame "Game Books" that are put out by the NFL. These Game Books, and thus the FO data, are missing 97 halves of coin tosses, roughly 4% of the overall data set. Given that the missing games are scattered throughout the 5 years, we inferred that this information is missing completely at random (Little 1988) and performed analysis on the remaining data only.

Sideline information for each of the home and visiting teams was collected manually using team websites and seating guides.[3,4] The stadium, coin toss, and play-by-play data were merged to calibrate our covariate of interest, $Sideline_i$, where for n_r rush plays and n_p pass plays,

$$\begin{aligned} Sideline_i &= \text{Direction of play i}, i = 1, \ldots, n_r, \\ &\qquad n_{r+1}, \ldots, n_r + n_p \\ &= \{\text{'Offense' if direction of play } i \\ &\qquad \text{is the offensive teams sideline,} \\ &\qquad \text{'Defense' if defense's sideline}\} \end{aligned}$$

We drop all middle rushes (those between the tackles) and middle passes from our data for two reasons. First, using plays over the middle as a control group for sideline plays requires extrapolation, given that plays over the middle may be unique for other reasons. For example, many defensive penalties in the middle of the field are late hits to the quarterback, which occur less frequently on sideline plays. Second, the statistical model described in Section II.A will use the logit link, one that makes the Independence of Irrelevant Alternatives assumption (IIA) (McFadden 1974). Under IIA, the relative odds of a penalty from one sideline versus the other act independently of plays over the middle of the field. As a result, contrasts between plays at each sideline will be identical whether or not penalties over the middle are excluded.

We consider the four outcomes Y_i, where $Y_i \in \{OHR_i, DPI_i, OPI_i, \text{ and } DAP_i\}$, most likely to vary based on pressure to appease team personnel, such that

- $OHR_i = \{1$ if there was an offensive holding penalty on play $i, 0$ o/w$\}, i = 1, \ldots, n_r$[5]
- $DAP_i = \{1$ if there was a defensive aggressive penalty on play $i, 0$ o/w$\}, i = 1, \ldots, n_r + n_s$
- $DPI_i = \{1$ if there was a defensive pass interference penalty on play $i, 0$ o/w$\}, i = n_r + 1, \ldots, n_r + n_p$[6]
- $OPI_i = \{1$ if there was an offensive pass interference penalty on play $i, 0$ o/w$\}, i = n_r + 1, \ldots, n_r + n_p$

Four violations are included in *DAP*: unnecessary roughness, personal foul, unsportsmanlike conduct, and horse collar tackle, which each penalize the defensive team 15 yards. While unnecessary roughness, personal foul, and unsportsmanlike conduct violations are also called on offensive units, they occur with a much smaller frequency (2.2 penalties per 1,000 plays, compared to 6.5 for the defensive unit) and are not considered.

A. Methods

We first examine if *Sideline* operates independently of other play and game characteristics by using χ^2 tests with each play's down, distance, and score. Furthermore, we examine the rates at which each offensive team calls plays toward each sideline to determine if there are any systematic differences in *Sideline* by team. Note that we are not able to easily identify referees that are more prone to exhibit a bias; referee crews are randomly assigned to games, and, without looking at game film, it is unknown which officials stand on which sideline during a specific contest.

We next analyze the relationship between Y and *Sideline*, again using χ^2 tests. Under an existing sideline pressure, we hypothesize that there will be higher *OHP* and *OPI* rates on plays run in the direction of the defensive team's sideline, relative to the offensive team's sideline. This can be attributed to either a pressure from the defensive team and coaches to throw a flag on the offensive unit on plays near the defensive sideline, or to a fear of throwing a penalty on the offensive team when in the neighborhood of the offensive sideline. Using a similar logic, we hypothesize that

3. Teams using multiple stadiums include Buffalo (also playing home games in Toronto), Minnesota (also at University of Minnesota), and San Francisco (new stadium for 2014). We dropped games played at London's Wembley Stadium given that there is no clear distinction for the game's home and away sidelines

4. See, for example, http://images.patriots.com/2014-individualgameprice.jpg

5. For offensive holding penalties, we look only at running plays given that on passing plays, most holding violations occur in the center of the field, away from either sideline.

6. Pass interference penalties cannot be called on rushing penalties.

there will be higher *DAP* and *DPI* rates on plays run toward the offensive team's sideline, relative to the defensive team's sideline. All together, if penalty rates vary by *Sideline*, it would link closely to the risk adverse nature of the referee position (Boeri and Severgnini 2011).

In addition to an overall difference in penalty rates, our second goal is to identify if a referee bias varies by the play's line-of-scrimmage. Coaches, players, and other team staff are required to stand within a 36-yard zone in the center of each sideline, between endpoints of a trapezoid that extends between the field's two marked 32-yard lines (NFL 2014). By rule, the sidelines at each stadium fit the same dimensions. Under a referee bias to appease a sideline, we expect the greatest difference in penalty rates between these two 32-yard line marks (e.g., the middle of the field), where the proximity between referees and team members peaks.

Generalized additive logistic models (GAM), (Hastie and Tibshirani 1986) are used to measure the effect of coach and player proximity on Y. GAMs require fewer assumptions and will allow us to more flexibly gauge the association between line-of-scrimmage and penalty likelihood, relative to a purely parametric approach such as multiple logistic regression. For example, one issue with multiple logistic regression is that it requires the associations between line-of-scrimmage and penalty outcomes to be specified as linear, quadratic, and/or piecewise, despite the true functional form being unknown. Let LOS_i be the line-of-scrimmage for play i. Each of our penalty outcomes is modeled semiparametrically using the full model

$$(1)\quad \text{logit}\left(P\left(Y_i = 1\right)\right) = \beta_1 * I\left(Sideline_i = Offense\right) + f_{Offense}\left(LOS_i\right) + f_{Defense}\left(LOS_i\right),$$

where $I(Sideline_i = Offense)$ is an indicator for whether or not the play was run in the direction of the offensive team's sideline, $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$, and $f_{Offense}(LOS_i)$ and $f_{Defense}(LOS_i)$ are the smoothed functions of the log-odds of a penalty based on plays run at the offensive and defensive teams sidelines, respectively, by *LOS*.

Model (1) allows for plays to have different average baseline penalty rates on each sideline, as measured parametrically using β_1. But in addition to an overall difference, the nonparametric smoothing functions allow us to check for possible differences in the effect of *LOS* within each *Sideline*. Informally, the smoothing functions in Model (1) reflect the "wiggliness" of rates across *LOS*s, and in using $f_{Offense}(LOS)$ and $f_{Defense}(LOS)$, we allow for the shape of the association between *LOS* and the penalty outcomes to vary by *Sideline*.

For each Y, Model (1) is compared to three reduced fits:

$$(2)\quad \text{logit}\left(P\left(Y_i = 1\right)\right) = \beta_1 * I\left(Sideline_i = Offense\right) + f_{Overall}\left(LOS_i\right),$$

$$(3)\quad \text{logit}\left(P\left(Y_i = 1\right)\right) = f_{Overall}\left(LOS_i\right),$$

$$(4)\quad \text{logit}\left(P\left(Y_i = 1\right)\right) = \beta_1 * I\left(Sideline_i = Offense\right).$$

Models (2) and (3) include an overall surface term for line of scrimmage, $f_{Overall}(LOS_i)$, implicitly making the assumption that any effect of the play's line-of-scrimmage on penalty likelihood acts independently of play direction. If (2) or (3) provide stronger fits than (1), we would conclude that an *LOS* impact does not significantly differ by *Sideline*. Model (4) assumes no effect of the game's line-of-scrimmage on a penalty outcome, but like Model (2), still allows for differences in the baseline penalty rates by *Sideline*. Models (1) and (2) are semiparametric as they contain both parametric and nonparametric components; Model (3) and Model (4) are fully nonparametric and parametric, respectively.

Because the NFL's play-by-play data do not give precise information on where each pass was thrown to, other than to identify each throw as either "deep" (15 yards or more) or "short," one last modification is made to more accurately consider where *DPI* and *OPI* infractions occur. Using the rough midpoints of the play-by-play distance labels, we estimated the location of the throw by adding 8 yards to the line-of-scrimmage for each short pass, and 25 yards for each deep pass. This will enable us to compare interference penalties by both *LOS* and the estimated yard line to which the ball was thrown.

Models are fit in the *R* statistical software (R Core Team 2014), and are contrasted using the Akaike information criterion (AIC, Akaike (1974)), which includes a penalty for unneeded parameters to discourage overfitting. In each GAM, *LOS* surfaces are estimated using penalized thin-plate regression splines (Gu and Wahba

TABLE 1
Play Characteristics and Penalty Outcome Counts (with %'s) by Play Direction

Covariate	Level	Sideline Offense	Sideline Defense	p Value[a]
Play type	Rush	8,326 (19.6)	7,864 (19.3)	0.152
	Pass	33,996 (80.4)	32,929 (80.7)	
Score (offense)	Behind 2+ possessions	9,695 (22.9)	9,425 (23.1)	0.456
	Behind 1 possession	11,253 (26.6)	10,624 (26.0)	
	Tied	7,748 (18.3)	7,502 (18.4)	
	Up 1 possession	8,163 (19.3)	7,984 (19.6)	
	Up 2+ possessions	5,463 (12.9)	5,258 (12.9)	
Down/distance	First and 10	17,436 (41.2)	16,935 (41.5)	0.750
	Second and long	10,747 (25.4)	10,411 (25.5)	
	Second and short	3,450 (8.2)	3,273 (8.0)	
	Third/fourth and long	6,229 (14.7)	5,911 (14.5)	
	Third/fourth and short	4,460 (10.5)	4,263 (10.5)	
Penalty outcomes	*OHR*	278 (3.3)	298 (3.8)	0.133
	DAP	298 (0.7)	205 (0.5)	<0.001
	DPI	494 (1.5)	408 (1.2)	0.018
	OPI	179 (0.5)	183 (0.6)	0.643
Total plays		42,322	40,793	

Note: *OHR* taken on run plays, *DPI* and *OPI* pass plays, *DAP* on all plays.
[a]Calculated using χ^2 tests of association.

1993). As in Wood (2006) and implemented by Mills (2014), the model was fitted using penalized iteratively reweighted least squares, and a generalized cross-validation procedure was used to prevent overfitting with the smoothing degrees of freedom. Finally, given the difficulty in interpreting the smoothed function estimates from GAMs, we will also plot the resulting estimated penalty rates by *LOS* and *Sideline* to identify where possible differences exist.

III. RESULTS

AAs and FOs data yielded 152,751 offensive plays, including 69,636 which were discarded as middle runs or middle passes. Of the 83,115 outside plays, $n_p = 66,925$ were passes (80.5%), with the remaining plays rushing attempts. Just over half (42,322, 50.9%) were run in the direction of the offensive team's sideline.

There do not appear to be any obvious differences between play calls (run/pass) or play situations (score, down, and distance) based on *Sideline* (Table 1). However, there are significantly higher rates of *DAP* (p value $< .001$) and *DPI* (p value $= .018$) among plays run at the offensive team's sideline. This follows our hypothesized association that suggested more defensive penalties in the presence of offensive team personnel. There are not significant differences in the rates of *OHR* or *OPI* by *Sideline*.

There do not appear to be any teams that vary their play call distribution by *Sideline*. The highest fraction of team-specific sideline calls is Jacksonville (53.1% toward its own sideline), and the lowest is Oakland (48.8%), with other team rates symmetrically distributed in between those two cutoffs.

Model estimates are shown in Table 2. After adjusting for *LOS* using GAMs, there are baseline differences in both *DAP* and *DPI* across all model specifications, as judged by significant β_1 estimates. There is an estimated 39% increase (95% confidence interval [CI], 1.16–1.66) in the odds of a *DAP* when the play is run toward the offensive team's sideline, compared to one run toward the defensive team's sideline. The overall odds of a defensive pass interference are 22% higher (95% CI, 1.06–1.40) on plays run toward the offensive team's sideline. There are no significant differences in the baseline rates of *OHR* or *OPI* violations. For each Y, estimates of β_1 are robust between the four model fits.

As implied by a noticeably lower AIC, Model (1) produces the best fit for *OHR* and *DPI* outcomes, implying that the effect of *LOS* on penalty likelihood significantly differs by *Sideline* for these penalties. There do not appear to be unique sideline effects based on *LOS* for *DAP* or *OPI*, whose lowest AIC's are achieved via Models (2) and (3), respectively.

Figures 1–4 show estimated penalty rates, along with 95% CIs, based on each play's *LOS*

TABLE 2
Model Statistics and Estimated Coefficient

	OHR		DAP		DPI		OPI	
Model	**AIC**	$\hat{\beta}_1\left(SE_{\hat{\beta}_1}\right)$	**AIC**	$\hat{\beta}_1\left(SE_{\hat{\beta}_1}\right)$	**AIC**	$\hat{\beta}_1\left(SE_{\hat{\beta}_1}\right)$	**AIC**	$\hat{\beta}_1\left(SE_{\hat{\beta}_1}\right)$
(1)	**4,969**	−0.117 (0.086)	6,145	0.331 (0.091)**	**9,519**	0.199 (0.069)**	4,498	−0.056 (0.107)
(2)	4,976	−0.130 (0.085)	**6,142**	0.334 (0.091)**	9,523	0.161 (0.067)*	4,495	−0.054 (0.105)
(3)	4,976	N/A	6,154	N/A	9,527	N/A	**4,493**	N/A
(4)	4,978	−0.130 (0.085)	6,142	0.335 (0.091)**	9,561	0.162 (0.067)*	4,505	−0.054 (0.105)

Notes: *OHR* taken on run plays, *DPI* and *OPI* pass plays, *DAP* on all plays. Model with lowest AIC in bold. N/A, not applicable.
*(**)p value < .05 (.01).

FIGURE 1
Offensive Holding Penalties by Sideline, Line of Scrimmage, Estimated per 1,000 Running Plays with 95% Confidence Intervals

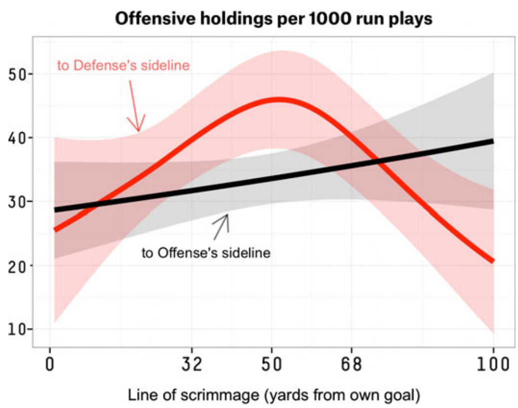
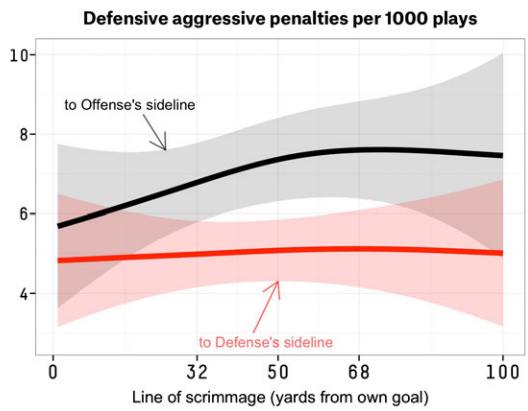


FIGURE 2
Defensive Aggressive Penalties by Sideline, Line of Scrimmage, Estimated per 1,000 Plays with 95% Confidence Intervals



and using model (1) for each penalty outcome.[7] The x-axis in each figure represents the *LOS*, with the offensive team moving out of its own end zone from left to right. The 36-yard area containing team personnel lies between yard markers representing 32 and 68 yards from the offensive team's end zone.

While *OHR*s are relatively consistent on plays run toward the offensive team's sideline, there is a noticeable spike on plays run toward the defensive team's sideline between roughly each of the 30-yard markers (Figure 1). At the 50-yard line, the fraction of *OHR*s is significantly higher—a difference of about 12 penalties per 1,000 plays—on plays run at the defensive team's sideline. This suggests a successful pressure from defensive coaches and sideline players to draw violations on the offense, but only in locations where personnel are located during the run of play.

*DAP*s are higher on plays run at the offensive team's sideline (Figure 2). In the middle of the field, infractions are about 50% more likely to occur in front of the offensive team's sideline. The differences in *DAP*s by *Sideline* are smaller closer to each end zone.

Near an offensive team's own end zone, *DPI*s are called significantly more often on plays run in the direction of the offensive team sideline (Figure 3). In contrast to *OHR*s and *DAP*s, however, differences in *DPI* rates by *Sideline* appear to subside around the 50-yard line. While this initially seems to violate our hypothesis of a greater sideline effect near midfield, keep in mind that most *DPI*s occur well

7. Exact smoothing parameter estimates are available upon request.

FIGURE 3
Defensive Pass Interference Calls by Sideline, Line of Scrimmage, Estimated per 1,000 Pass Plays with 95% Confidence Intervals

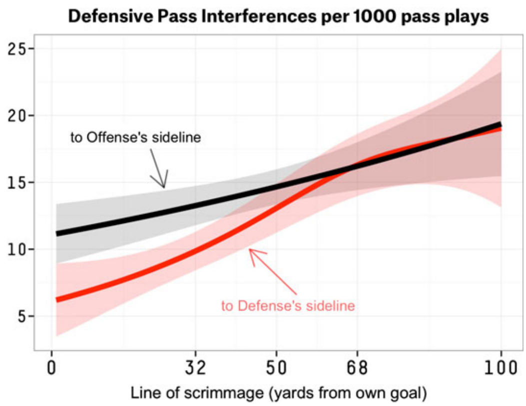
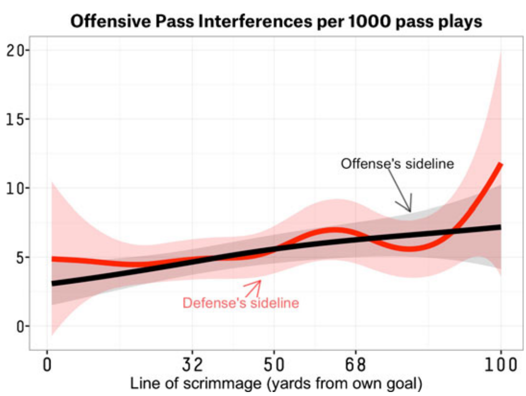


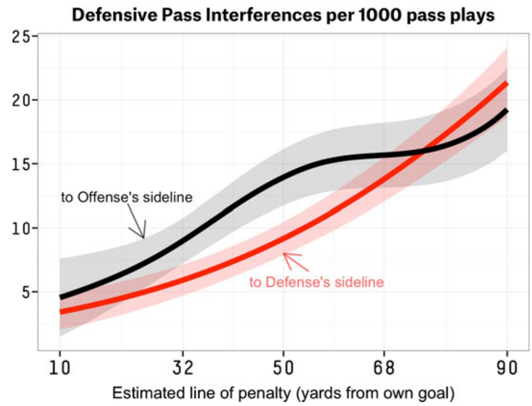
FIGURE 4
Offensive Pass Interference Calls by Sideline, Line of Scrimmage, Estimated per 1,000 Pass Plays with 95% Confidence Intervals



beyond the *LOS*, at the yard line to which the ball was thrown.

Figure 5 likewise shows the model-estimated penalty rates for *DPI*, but instead of surfaces for *LOS*, it uses an estimated yard line to which the pass was thrown. As in our other penalty outcomes, the largest differences in *DPI* rates occur on plays likely to have ended near the middle of the sideline. Of n_p, 46,225 (69.1%) were plays that more than likely were thrown in vicinity of team personnel on either sideline. The ratio of defensive pass interference calls was nearly 3:2

FIGURE 5
Defensive Pass Interference Calls by Sideline, Imputed Line of Penalty, Estimated per 1,000 Pass Plays with 95% Confidence Intervals



(raw counts, 293 and 204) in this window, favoring plays toward the offensive team's sideline. On passes thrown near one of the end zones, the ratio of flags by sideline was roughly 1:1 (counts of 201 and 204). Such a drastic difference adds context to Figure 3, which is limited in the fact that it uses only the *LOS* to contrast *DPI* frequencies.

Unlike our other outcomes, there are no noticeable differences in *OPI* rates by *Sideline* (Figure 4), a conclusion that holds when using either *LOS* or the estimated yard line to which the pass was thrown.

IV. DISCUSSION

It is generally assumed that football referees assess each possible violation on its own accord. However, we find multiple penalties whose likelihoods vary significantly based on whether or not the violation occurred in the presence of one team's sideline. Further, model estimates suggest that the sideline locations including team personnel result in the largest differences in call rates. The highest ratios of the estimated differences range from 1.3 (offensive holding) to 2 (defensive pass interference) times as many violations, comparing one sideline to the other.

Models (1) through (4) do not adjust for any variables or functions of variables besides *Sideline* and *LOS*. However, our covariate of interest, *Sideline*, appears to act independently of other play- and game-specific covariates (Table 1). As a result, problems such as omitted variable bias, in which variables that affect both the treatment (*Sideline*) and the outcome (Y) can produce confounded associations, should not impact our estimated *Sideline* effects. To check this assumption, we also adjusted each model for point differential, down and distance, game minute, a season-specific factor, and an indicator for whether or not the home team was the offensive team. Each estimated β_1 in these adjusted fits was within a one hundreth of the estimates shown in Table 2.

We attempted other model specification checks. First, we fit Model (1) using an interaction between the game's minute (1–60) and *Sideline* to check the consistency of our sideline effects over the course of the game. None of the fits yielded significant interaction terms, implying that a sideline bias appears to be more instantaneous, as opposed to a pressure that grows or decays over the course of a contest. Additionally, we fit a model without β_1 but with smoothed *LOS* functions for each sideline. The AIC term from each of these fits ranked no better than the second best, among the models shown in Table 2, suggesting that our inclusion of a term to indicate the difference in average penalty rates by *Sideline* helps from a model fit perspective.

In principle, each of our penalty outcomes should result in enough punishment on the offending team as to deter players from committing the foul to begin with. One alternative explanation for a referee sideline bias, however, is that players adapt their behavior in the presence of coaches and players on the other team, perhaps acting more aggressively near opponents. This alternative justification is most reasonable for the *DAP* findings, with several such penalties likely occurring on the sideline itself and off the playing field (although this is impossible to check, as our play-by-play data does not consistently indicate if a play finished out-of-bounds). However, such an explanation would not justify the observed differences in *OHR*s, given that most players first engage in contact close to the line-of-scrimmage, away from referees and the sideline. Furthermore, *DPI*s on outside passing plays can happen anywhere within the outer third of the field, and in most of these locations, it is difficult to reason that the defending players act differently because of the sideline they are next to. As one final consideration, there is also the possibility that players were aware of a sideline bias and adapted their behavior accordingly, in which case our estimates could underestimate the true effects. Ultimately, although sideline choice appears independent of team- and play-specific factors, true penalty rates are unknown given the possible endogeneity.

While *DPI* frequencies appear to vary based on the location of the throw, we found no significant differences in *OPI* rates. However, there are several reasonable explanations for the null finding. First, while *DPI*s apply only from the time the ball is thrown until it is touched, *OPI*s can be thrown anytime beginning with the snap (Section V, NFL 2014). As a result, of the pass interference calls, only *DPI*s are guaranteed to be thrown near where the eventual pass lands according to the play-by-play data. Anecdotally, *OPI*s likely include several "pick" plays, in which one offensive player sets a pick for another offensive player to get open and receive the ball. In these and other scenarios, although the pass may be thrown toward a sideline, it is not as reasonable to assume a sideline pressure because the violation itself may have occurred on a different part of the field. As an additional justification for the null findings, *OPI* may require less discretion on behalf of the officials, in which case such decisions would be less likely to vary based on a referee's pressure to appease. Finally, there are fewer than half as many *OPI*s as *DPI*s. In addition to making it more difficult to discern a statistical difference with *OPI*, perhaps it is also less likely that coaches and players try to sway referee decisions with such a relatively rare infraction.

Although the differences in certain penalty rates meet criteria for statistical significance, a final important consideration is that of practical significance. Notably, each of our penalty outcomes carries a hefty punishment against the violating team. *DPI*s and *DAP*s each give the offensive unit a new set of downs with which to possess the ball, and *DPI*s are a spot foul, meaning that the line-of-scrimmage moves to wherever the penalty took place. As a result, while *DAP*s are usually a 15-yard infraction, many *DPI*s can be worth additional yardage. Additionally, offensive teams are penalized 10 yards for any holding penalty. Such punishments are relatively severe compared to other infraction types, and can have obvious impacts on the game.

Further identifying the effects of penalties on each game, however, can take some care. Anecdotally, Burke (2015) compared an offensive team's win probability with and without a potential *DPI* call from a 2014 game between Detroit and Dallas, finding that an accepted penalty (on one team's sideline, at roughly the 35 yard line) would have increased the offensive team's chances of winning from 67% to 79%. In this example, win probabilities were derived from a model that included the game's score, time, down, distance, and field position (Burke 2014). We tried a similar approach using our data, fitting a multiple logistic regression model of game outcome (1 if the offensive team won the game, 0 otherwise) as a function of the five fixed effects suggested by Burke (2014). Additionally, we also included each of the pairwise interactions between these covariates to account for the varying effects of each covariate on game outcome (e.g., point differential is more important later in the game). In comparing the offensive team's win probability before and after accepted *DPI*s, the median win probability added after the penalty was 3.6%, and 25% of all *DPI*s increased the offensive team's win probability by 6.1% or more. Although team-level advantages to a sideline pressure may even out across the league's 32 teams with a large number of games, because the NFL's regular season is only 16 games, changes in win probability accounted for by a sideline bias are unlikely to be a zero-sum game within a season. Relatedly, Snyder and Lopez (2015) found a significant and nonrandom variation when looking at an offensive team's ability to draw *DPI*s, implying that certain teams may be better at drawing interference penalties than others.

As the majority of offensive plays in an NFL game do not end in *OHR*s, *DPI*s, or *DAP*s, our penalty outcomes are relatively rare. However, the relative fraction of penalty outcomes that can be attributed to a sideline bias is noteworthy. For example, our data set contains 596 *DPI*s estimated to have occurred near midfield (204 at defensive sideline, 99 down the middle, 293 at offensive sideline). Using the defensive sideline *DPI* rate as a baseline, roughly 15% of all *DPI*s in this area can be explained by a sideline pressure. This adds context to Figures 1–5, which show absolute differences in rates varying by up to 12 penalties per 1,000 plays (*OHR*), or relative differences as high as 50% (*DPI*) to 100% (*DAP*), given the *LOS*.

Perhaps most importantly, the aggregated effects of a sideline bias dominate any differences in penalties between home and away teams. For example, our data set of sideline passes contained 60 additional *DPI* calls when the home team was on offense. This difference represents about two-thirds of the difference in flags that we observed based on *Sideline*. Moreover, using the same data presented above, home teams on offense were called for *more OHR*s (we would expect fewer under a home bias) and drew *fewer DAP*s (we would expect more), relative to the away team being on offense. On average, as far *OHR* and *DAP* are concerned, there is no obvious advantage to being a home team, particularly when compared to a *Sideline* effect. Additionally, using an interaction term in Model (1), we checked if our sideline effects differed based on whether or not the home team was on offense. There was no evidence that a sideline effect differed based on the offensive team's status, as judged by the significance of the interaction term.

Finally, although there were 842 *DAP*s in our data set, this only represents about 45% of these penalty outcomes; an additional 1,029 occurred on special teams. As mentioned in Section II, directional information for special teams plays is not included in the NFL's play-by-play reports. Given that several of these missing penalties came on kick and punt returns, the importance of a sideline pressure on *DAP* calls is potentially higher than what we were able to find. Other than the 4% of halves missing coin toss data, all sideline *OPI*s, *DPI*s, and *OHR*s are accounted for.

V. CONCLUSION

A home bias due to referee calls has been extensively studied, and one purveying theory is that under duress, the crowd noise prompts referees to make decisions that support the home team (Nevill, Balmer, and Williams 2002; Sutter and Kocher 2004; Unkelbach and Memmert 2010). This confirms work in behavioral economics and psychology, where it has been shown that the most salient cues have the largest effects when people are forced to make decisions under a time pressure (Wallsten and Barton 1982).

Until now, however, it has been assumed that the primary impetus for uncertain referees has been a crowd noise in favor of the home team (Buraimo, Forrest, and Simmons 2010; Nevill, Balmer, and Williams 2002). We propose that in several settings, a sideline pressure dwarfs that of the home crowd. However, this is not to say a home bias does not exist in football; instead, we argue that if noise is a cue, a sideline noise is more salient than that of the home crowd. This follows results of Buraimo, Forrest, and Simmons (2010), who noted that effect of noise on referee behavior was proportional to how close noise was to the referees.

There are several extensions to our work. The role of referees (supervisors, judges) in sports is defined by their neutrality toward the productivity and actions of the players and coaches (workers) within their subjective judgements. Therefore, bias arising from proximity has specific lessons for managers in the workplace. Given that the most extreme sideline biases occur near the middle of the field, when coaches and players are closest to the referees, our results match those of Judge and Ferris (1993), who showed that proximity alone increases favorable rulings. But there may also be a risk adverse nature to referee behavior. Although they are assigned to behave impartially, the constant fear of upsetting nearby coaches and players, mixed with the possibility of being blamed after the game by this same group, may be driving referees to make biased decisions in favor of the nearest sideline.

Given the multimillion dollar athletes and coaches that are pressuring them, one final theory would be to consider the referees to be subordinates, and not as judges or supervisors. In this role, the less well-known referees are goaded by athletes and coaches into making favorable decisions, and the sideline banter is as much of a workplace bullying (Rayner and Hoel 1997) as it is a lobbying. Such a pressure from the sideline also has a rent-seeking nature to it, with coaches and players seeking favorable decisions without reciprocation.

One response to social pressure is to make performance less sensitive to evaluations (Prendergast and Topel 1993a). However, this is difficult in sport given the public nature in which all participants, in particular the referees, operate. Given that referees have shown a willingness to adapt their behavior when made aware of their biases (Pope, Price, and Wolfers 2013), one solution may simply be to make the league aware of this behavior. Ultimately, however, unless team personnel are removed from direct contact with the run of play, it may be impossible to remove a sideline bias from the NFL.

Several sports, for example, basketball, hockey, and soccer, also feature similar planned directional changes to the one in football. However, granular data with respect to the location of referees and participants in these sports may be difficult to obtain. If such information is available, future work is warranted to unify and extend the effects presented here.

REFERENCES

Abrams, J. "Working the Refs Is Part of a Coach's Game Plan." 2008. Accessed August 1, 2015. http://www.nytimes.com/2008/12/16/sports/basketball/16refs.html?pagewanted=all _r=0.

Abrevaya, J., and R. McCulloch. "Reversal of Fortune: A Statistical Analysis of Penalty Calls in the National Hockey League." *Journal of Quantitative Analysis in Sports*, 10(2), 2014, 207–24.

Akaike, H. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, 19(6), 1974, 716–23.

Boeri, T., and B. Severgnini. "Match Rigging and the Career Concerns of Referees." *Labour Economics*, 18(3), 2011, 349–59.

Boyko, R. H., A. R. Boyko, and M. G. Boyko. "Referee Bias Contributes to Home Advantage in English Premiership Football." *Journal of Sports Sciences*, 25(11), 2007, 1185–94.

Buraimo, B., D. Forrest, and R. Simmons. "The 12th Man?: Refereeing Bias in English and German Soccer." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2), 2010, 431–49.

Burke, B. "Win Probability and Win Probability Added." 2014. Accessed August 1, 2015. http://www.advancedfootballanalytics.com/index.php/home/stats/stats-explained/win-probability-and-wpa.

———. "How Much Did the Reversed Penalty Affect the Lions Cowboys Game?" 2015. Accessed August 1, 2015. http://www.advancedfootballanalytics.com/index.php/home/analysis/game-analysis/210-how-much-did-the-reversed-penalty-affect-the-lions-cowboys-game.

Dohmen, T. J. "The Influence of Social Forces: Evidence from the Behavior of Football Referees." *Economic Inquiry*, 46(3), 2008, 411–24.

Dohmen, T., and J. Sauermann. "Referee Bias." *Journal of Economic Surveys*, 2015. Accessed October 1, 2015. http://onlinelibrary.wiley.com/doi/10.1111/joes.12106/full

Garicano, L., I. Palacios-Huerta, and C. Prendergast. "Favoritism under Social Pressure." *Review of Economics and Statistics*, 87(2), 2005, 208–16.

Gift, P. "Sequential Judgment Effects in the Workplace: Evidence from the National Basketball Association." *Economic Inquiry*, 53(2), 2015, 1259–74.

Gift, P., and R. M. Rodenberg. "Napoleon Complex Height Bias among National Basketball Association Referees." *Journal of Sports Economics*, 15(5), 2014, 541–58.

Gu, C., and G. Wahba. "Smoothing Spline ANOVA with Şcomponent-Wise Bayesian 'Confidence Intervals'." *Journal of Computational and Graphical Statistics*, 2(1), 1993, 97–117.

Hastie, T., and R. Tibshirani. "Generalized Additive Models." *Statistical Science*, 1(3), 1986, 297–310.

Heckelman, J. C., and A. J. Yates. "And a Hockey Game Broke Out: Crime and Punishment in the NHL." *Economic Inquiry*, 41(4), 2003, 705–12.

Judge, T. A., and G. R. Ferris. "Social Context of Performance Evaluation Decisions." *Academy of Management Journal*, 36(1), 1993, 80–105.

Kitchens, C. "Identifying Changes in the Spatial Distribution of Crime: Evidence from a Referee Experiment in the National Football League." *Economic Inquiry*, 52(1), 2014, 259–68.

Lane, A. M., A. M. Nevill, N. S. Ahmad, and N. Balmer. "Soccer Referee Decision-Making:? Shall I Blow the Whistle?" *Journal of Sports Science and Medicine*, 5(2), 2006, 243–53.

Little, R. J. "A Test of Missing Completely at Random for Multivariate Data with Missing Values." *Journal of the American Statistical Association*, 83(404), 1988, 1198–202.

Lopez, M., and K. Snyder. “Biased Impartiality among National Hockey League Referees.” *International Journal of Sport Finance*, 8(3), 2013, 208–23.

McFadden, D.. “Conditional Logit Analysis of Qualitative Choice Behavior,” in *Frontiers in Econometrics*, edited by P. Zarembka. New York: Academic Press, 1974, 105–42.

Mills, B. M. “Social Pressure at the Plate: Inequality Aversion, Status, and Mere Exposure.” *Managerial and Decision Economics*, 35(6), 2014, 387–403.

Moskowitz, T., and L. J. Wertheim. *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won*. New York: Crown Archetype, 2011.

Nevill, A. M., N. J. Balmer, and A. M. Williams. “The Influence of Crowd Noise and Experience upon Refereeing Decisions in Football.” *Psychology of Sport and Exercise*, 3(4), 2002, 261–72.

NFL. “NFL Rulebook.” 2014. Accessed August 1, 2015. http://static.nfl.com/static/content/public/image/rulebook/pdfs/4_2013_Field.pdf.

Parsons, C. A., J. Sulaeman, M. C. Yates, and D. S. Hamermesh. “Strike Three: Discrimination, Incentives, and Evaluation.” *American Economic Review*, 101, 2011, 1410–35.

Petchesky, B. “Earl Thomas Says the Seahawks “Are Playing the Referees Too.” 2014. Accessed August 1, 2015. http://deadspin.com/earl-thomas-says-the-seahawks-are-playing-the-referees-1648391758.

Pettersson-Lidbom, P., and M. Priks. “Behavior under Social Pressure: Empty Italian Stadiums and Referee Bias.” *Economics Letters*, 108(2), 2010, 212–14.

Pope, B. R., and N. G. Pope. “Own-Nationality Bias: Evidence from UEFA Champions League Football Referees.” *Economic Inquiry*, 53(2), 2015, 1292–304.

Pope, D. G., J. Price, and J. Wolfers. “Awareness Reduces Racial Bias.” Technical Report, National Bureau of Economic Research, 2013.

Prendergast, C., and R. Topel. “Discretion and Bias in Performance Evaluation.” *European Economic Review*, 37(2), 1993a, 355–65.

———. “Favoritism in Organizations.” Technical Report, National Bureau of Economic Research, 1993b.

Price, J., and J. Wolfers. “Racial Discrimination among NBA Referees” Centre for Economic Policy Research, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2014,. Accessed August 1, 2015. http://www.R-project.org/.

Rayner, C., and H. Hoel. “A Summary Review of Literature Relating to Workplace Bullying.” *Journal of Community and Applied Social Psychology*, 7(3), 1997, 181–91.

Snyder, K., and M. Lopez. “Consistency, Accuracy, and Fairness: A Study of Discretionary Penalties in the NFL.” *Journal of Quantitative Analysis in Sports*, 11(4), 2015, 219–30.

Sutter, M., and M. G. Kocher. “Favoritism of Agents—The Case of Referees’ Home Bias.” *Journal of Economic Psychology*, 25(4), 2004, 461–69.

Unkelbach, C., and D. Memmert. “Crowd Noise as a Cue in Referee Decisions Contributes to the Home Advantage.” *Journal of Sport and Exercise Psychology*, 32(4), 2010, 483–98.

Wallsten, T. S., and C. Barton. “Processing Probabilistic Multidimensional Information for Decisions.” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5), 1982, 361–84.

Wood, S. *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: CRC Press, 2006.