

Are football referees really biased and inconsistent?

Evidence on the incidence of disciplinary sanction in the English Premier League

Peter Dawson¹, Stephen Dobson², John Goddard³⁺, John Wilson⁴

Abstract

This paper presents a statistical analysis of patterns in the incidence of disciplinary sanction (yellow and red cards) taken against players in the English Premier League over the period 1996-2003. Several questions concerning sources of inconsistency and bias in refereeing standards are examined. Evidence is found to support a time consistency hypothesis, that the average incidence of disciplinary sanction is predominantly stable over time. However, a refereeing consistency hypothesis, that the incidence of disciplinary sanction does not vary between referees, is rejected. The tendency for away teams to incur more disciplinary points than home teams cannot be attributed to the home advantage effect on match results, and appears to be due to a refereeing bias favouring the home team.

Keywords

refereeing bias and inconsistency, English Premier League football, bivariate Poisson regression, bivariate negative binomial regression

1. Department of Economics and International Development, University of Bath,
Bath BA2 7AY, UK
tel: +44 (0)1225 383074
email: p.m.dawson@bath.ac.uk
2. Department of Economics, University of Otago, PO Box 56, Dunedin, New Zealand
tel: +64 3 479 5296
email: s.dobson@business.otago.ac.nz
3. School of Business and Regional Development, University of Wales, Bangor, Gwynedd,
LL57 2DG, UK
tel: +44 (0)1248 383221
email: j.goddard@bangor.ac.uk
4. School of Management, University of St Andrews, The Gateway, North Haugh, St Andrews,
Fife, KY16 9SS, UK
tel: +44 (0)1334 462803
email: jsw7@st-and.ac.uk

+ Corresponding author.

The authors are grateful to an associate editor and two referees for many helpful comments and insights. The usual disclaimer applies. We are grateful to participants at a Sports Economics Workshop held at the University of Groningen in March 2005 for comments on an early draft of the paper. We are grateful to Rob Simmons and Tunde Buraimo for kindly providing access to their list of televised Premier League matches.

Are football referees really biased and inconsistent?

Evidence on the incidence of disciplinary sanction in the English Premier League

Abstract

This paper presents a statistical analysis of patterns in the incidence of disciplinary sanction (yellow and red cards) taken against players in the English Premier League over the period 1996-2003.

Several questions concerning sources of inconsistency and bias in refereeing standards are examined.

Evidence is found to support a time consistency hypothesis, that the average incidence of disciplinary sanction is predominantly stable over time. However, a refereeing consistency hypothesis, that the incidence of disciplinary sanction does not vary between referees, is rejected. The tendency for away teams to incur more disciplinary points than home teams cannot be attributed to the home advantage effect on match results, and appears to be due to a refereeing bias favouring the home team.

Keywords

refereeing bias and inconsistency, English Premier League football, bivariate Poisson regression, bivariate negative binomial regression

Are football referees really biased and inconsistent?

Evidence on the incidence of disciplinary sanction in the English Premier League

1. Introduction

In professional team sports with a high public profile, including association football (soccer), disciplinary transgressions by players and sanctions taken by referees provide a rich source of subject material for debate among pundits, journalists and the general public. Although newspaper and television pundits routinely and piously deplore incidents involving foul play or physical confrontation, there is no doubt that a violent incident, immediately followed by the referee's theatrical action of brandishing a yellow or red card in the direction of the miscreant, makes an important contribution to the popular appeal of the football match as spectacle or drama.

Due to the ever-increasing scope of television coverage of football especially at the highest level, together with improvements in video technology, the actions of players and referees have never been more keenly and intensely scrutinised than they are in the modern-day game. In sporting terms, the margins separating success from failure can be slender, and often depend ultimately on split-second decisions taken by referees and players in the heat of battle. Yet the financial implications of success or failure for individual football clubs and their players can be huge. The football authorities are under intense pressure from all sides to take steps to ensure that refereeing decisions are as fair, consistent and accurate as is humanly possible. Bearing all of these considerations in mind, it is perhaps surprising that academic research on the incidence of disciplinary sanction in professional sports is relatively sparse. This paper seeks to fill this gap, by presenting a statistical analysis of patterns in the incidence of disciplinary sanction taken against players in English professional football's highest division, the Premier League, over a seven-year period from 1996 to 2003.

The empirical analysis addresses several questions concerning possible sources of home team bias or inconsistency in refereeing standards. The hypotheses investigated include: a home advantage hypothesis, that the tendency for away teams to incur more disciplinary points than home teams is solely a corollary of home advantage, or the tendency for home teams to win more often than away

teams; a refereeing consistency hypothesis, that the propensity to take disciplinary action does not vary between referees; and a time consistency hypothesis, that the incidence of disciplinary sanction is stable over time and unaffected by changes to the content or interpretation of the rules. We examine the extent to which the rate of disciplinary sanction against each team depends upon relative team quality. To what extent does it depend upon whether the match itself is competitive (between two evenly balanced teams) or uncompetitive? Does it depend upon whether end-of-season outcomes are at stake for either team? Is it affected by the stadium audience, and does it depend upon whether the match is broadcast live on TV? We aim to provide the football authorities and other parties with a firmer factual basis than has previously been available for debate and decisions concerning the interpretation and implementation of the rules governing disciplinary sanction in football.

The structure of the paper is as follows. Section 2 reviews the previous academic literature on the topic of disciplinary sanction in professional team sports. Section 3 identifies statistical distributions that may be regarded as candidates for modelling the incidence of disciplinary sanction in football: specifically, the univariate and bivariate Poisson and negative binomial distributions. Section 4 develops a theoretical analysis of the relationship between team quality and the incidence of disciplinary sanction. Section 5 reports estimations for the expectations of the incidence of disciplinary sanction conditional on a number of covariates, and reports a series of hypothesis tests concerning the sources of refereeing bias and inconsistency. Section 6 summarises and concludes.

2. Literature review

Previous academic scrutiny of the topic of disciplinary sanction in professional team sports has focused mainly on the impact of dismissals on match results; and on issues of incentives, monitoring and detection, which arise in the economics literature on crime and punishment.

Ridder et al. (1994) model the effect of a dismissal on football match results, using Dutch professional football data from the period 1989-92. Probabilities are estimated for the match result conditional on the stage of the match at which a dismissal occurs. A method is developed for estimating the earliest stage of the match at which it is optimal for a defender to resort to foul play

punishable by dismissal in order to deny an opposing forward a goal scoring opportunity (conditional on the probability that the opportunity would be converted), assuming the defender's objective is to minimise the probability of his team losing the match. In a multivariate analysis of the determinants of match results from the 2002 FIFA World Cup, Torgler (2004) also finds a significant association between player dismissals and match results.

In the literature on the economics of crime and punishment, rule changes in professional sports have occasionally created opportunities for empirical scrutiny of the question whether increasing the resources assigned to monitoring or policing leads to an increase or a decrease in the incidence of crimes being detected. This incidence increases if the monitoring effect (more monitoring increases detection rates) exceeds the deterrent effect (the tendency for criminals to be deterred from offending because monitoring has increased).

In North American college basketball's Atlantic Coast Conference, an increase in the number of referees from two to three per match was implemented in 1979. McCormick and Tollison (1984) find that the number of fouls called per game fell sharply. If refereeing competence improved with the increase in the number of officials (with fewer fouls being missed), the actual crime rate must have decreased by even more than is suggested by the fall in the number of fouls called.

In the North American National Hockey League (NHL), an increase from one to two referees per match was phased in during the 1998-9 and 1999-2000 seasons. Heckelman and Yates (2002) note that fouls detected are observed but fouls committed are unobserved. The difference between the two enters the error term of a regression for fouls detected when the latter is used as a proxy for fouls committed. Because this difference is likely to be correlated with the number of referees, instrumental variables are used to model the latter in the regression for fouls detected. Although more fouls were detected in NHL matches with two referees than in the matches with one, this appears to have been due solely to a monitoring effect. The incidence of fouls being committed was the same under both refereeing regimes. Distinguishing between violent and non-violent offences, Allen (2002) finds detection of the latter was significantly higher with two referees than with one. Again, this suggests the monitoring effect outweighs the deterrent effect.

As part of a wide-ranging investigation of the impact of changes in reward structures on effort using Spanish football data, Garciano and Palacios-Huerta (2000) draw comparisons between the numbers of yellow and red cards incurred before and after the introduction (in the 1995-6 season) of the award of three league points for a win and one for a draw. Previously two points had been awarded for a win and one for a draw. More yellow cards were awarded after the reward differential between winning and drawing was increased. This finding is consistent with theoretical models of tournaments, in which players can engage in sabotage activity. Following a rule change implemented at the start of the 1998-9 season requiring an automatic red card punishment for the tackle from behind, Witt (2005) finds evidence of an increase in the incidence of yellow cards (awarded for lesser offences), but no increase in the incidence of red cards. This finding suggests a deterrent effect was operative: football teams modified their behaviour in response to the rule change.

There are some limitations to the statistical models and methods that have been employed in this literature. While McCormick and Tollison (1984) and Garciano and Palacios-Huerta (2000) estimate separate equations for the winning and losing teams, the other papers cited above report equations for the total number of offences called against both teams combined, factoring out many of the team-specific determinants of the incidence of disciplinary sanction. Despite the discrete structure of a 'fouls' or 'cards' dependent variable, McCormick and Tollison (1984) and Heckelman and Yates (2002) report ordinary least squares (OLS) regressions. Garciano and Palacios-Huerta (2000) discard univariate Poisson regressions in favour of OLS, because the former could not be estimated using fixed effects for teams. Witt (2005) reports both OLS and univariate Poisson regressions, while Allen (2002) uses a univariate negative binomial regression.

Alleged refereeing bias in favour of the home team is a frequently-aired grievance on the part of managers, players and spectators, which has also received some attention in the academic literature. Garciano et al. (2001) and Sutter and Kocher (2004) find a tendency for referees to add more time at the end of matches when the home team is trailing by one goal than when the home team is leading. Nevill et al. (2002) played videotapes of tackles to referees who, having been told the identities of the home and away teams, were asked to classify the tackles as legal or illegal. One group of referees viewed the tape with the soundtrack (including crowd's reaction) switched on, while a second group

viewed silently. The first group were more likely to rule in favour of the home team, and the first group's rulings were more in line with those of the original match referee. Using German Bundesliga data, Sutter and Kocher (2004) analyse reports on the referee's performance, which comment on the legitimacy of penalties awarded and on cases of failure to award a legitimate penalty. There is evidence of home team bias in such decisions.

3. Modelling the incidence of disciplinary sanction in English Premier League football

Tables 1 and 2 show the frequency distributions for the numbers of yellow cards and red cards incurred by the home and away teams in the $N=2,660$ Premier League matches played during the seven English football seasons from 1996-7 to 2002-3 inclusive. The data reported in Tables 1 and 2 were originally compiled from match reports posted on the Football Association website, which have since been deleted. These data are available on request from the corresponding author. A yellow card, also known as a booking or caution, is awarded for less serious transgressions. There is no further punishment within the match, unless the player commits a second similar offence, in which case a red card is awarded and the player is expelled for the rest of the match (with no replacement permitted, so the team completes the match one player short). A red card, also known as a sending-off or dismissal, is awarded for more serious offences, and results in immediate expulsion (again, with no replacement permitted). After the match, a red card leads to a suspension, preventing the player from appearing in either one, two or three of his team's next scheduled matches. A player who accumulates five yellow cards in different matches within the same season also receives a suspension.

The dependent variables in the estimations reported in this paper are the total numbers of disciplinary 'points' incurred by the home ($i=1$) and away ($i=2$) teams in match j for $j=1\dots N$, denoted $\{Z_{1,j}, Z_{2,j}\}$ and calculated by awarding one point for a yellow card and two for a red card. Only two points (not three) are awarded when a player is dismissed having committed two cautionable offences in the same match. This metric accurately reflects the popular notion that a red card is in some sense equivalent to two yellow cards. In fact, this notion was literally true of just under one-half of the red cards awarded during the observation period (227 out of 462 dismissals in total), which resulted from

two cautionable offences having been committed in the same match. Attempts to estimate versions of the model using separate yellow cards and red cards dependent variables were successful for the former but unsuccessful for the latter, presumably because the incidence of red cards is too sparse for reliable estimation. The results for estimations based on alternative metrics for the definition of the dependent variable, with red cards contributing either one point or three points, are similar to those reported below, and are available from the corresponding author on request. Table 3 reports the sample frequency distribution for $\{Z_{1,j}, Z_{2,j}\}$, with the rows and columns for $Z_{i,j} \geq 5$ consolidated into a single row and a single column.

In the applied statistics literature, several methods have been used to model professional team sports bivariate count data, where each match yields two values of a discrete dependent variable (one for each team: commonly the number of goals or points scored, but the disciplinary points dependent variable in the present case has the same structure). Maher (1982), Dixon and Coles (1997), Dixon and Pope (2004) and Goddard (2005) use the bivariate Poisson distribution to model English football goal scoring data, while Cain et al. (2000) use the univariate negative binomial distribution. Lee (1999) models Australian rugby league scores data using a bivariate negative binomial distribution.

A description follows of the probability models that are considered as candidates for the disciplinary points dependent variables $\{Z_{1,j}, Z_{2,j}\}$. Let $f_i(z_i) = P(Z_{i,j} = z_i)$ for $z_i = 0, 1, 2, \dots$ denote the marginal probability function for $Z_{i,j}$ for $i=1, 2$ and $j=1 \dots N$. The two candidate distributions for $f_i(z_i)$ are the Poisson, where $f_i(z_i) = \exp(-\lambda_{i,j}) \lambda_{i,j}^{z_i} / z_i!$; and the negative binomial, where $f_i(z_i) = [\Gamma(\rho_i + z_i) / \{z_i! \Gamma(\rho_i)\}] \{\rho_i / (\lambda_{i,j} + \rho_i)\}^{\rho_i} \{\lambda_{i,j} / (\lambda_{i,j} + \rho_i)\}^{z_i}$, and Γ denotes the gamma function. In both cases, $E(Z_{i,j}) = \lambda_{i,j}$. For the Poisson, $\text{var}(Z_{i,j}) = \lambda_{i,j}$. For the negative binomial, the ancillary parameter $\rho_i > 0$ allows for overdispersion, such that $\text{var}(Z_{i,j}) = \lambda_{i,j}(1 + \kappa_i \lambda_{i,j})$, where $\kappa_i = 1/\rho_i$. In the sample data, the degree of overdispersion is relatively small, but non-zero. The sample mean values of $Z_{1,j}$ and $Z_{2,j}$ are 1.4650 and 2.0451, and the sample variances are 1.7216 and 2.2657.

For each of the Poisson and negative binomial specifications of $f_i(z_i)$, three formulations of the bivariate probability distribution are considered, denoted P1-P3 and N1-N3. In P1 and N1, the joint probability function is the product of the two univariate probability functions, $P(Z_{1,j} = z_1,$

$Z_{2,j}=z_2)=f_1(z_1)f_2(z_2)$. No allowance is made for correlation between $Z_{1,j}$ and $Z_{2,j}$. Below, P1 and N1 are referred to as the double Poisson and the double negative binomial, respectively.

In P2 and N2, the joint distribution function is constructed by substituting the two univariate distribution functions into the Frank copula, in accordance with Lee's (1999) model for points-scoring in rugby union. Let $F_i(z_i)$ denote the univariate distribution functions for $Z_{i,j}$ corresponding to $f_i(z_i)$. The bivariate joint distribution function is:

$$G[F_1(z_1), F_2(z_2)] = \frac{1}{\phi} \ln \left(1 + \frac{\{\exp[\phi F_1(z_1)] - 1\} \{\exp[\phi F_2(z_2)] - 1\}}{\exp(\phi) - 1} \right)$$

The ancillary parameter ϕ determines the nature of any correlation between $Z_{1,j}$ and $Z_{2,j}$. For $\phi < 0$ the correlation between $Z_{1,j}$ and $Z_{2,j}$ is positive, and for $\phi > 0$ the correlation is negative.

$G[F_1(z_1), F_2(z_2)]$ is undefined for $\phi = 0$, but it is conventional to write $G[F_1(z_1), F_2(z_2)] = F_1(z_1)F_2(z_2)$ in this case. The bivariate joint probability function corresponding to $G[F_1(z_1), F_2(z_2)]$ is obtained iteratively, as follows:

$$P(Z_{1,j}=0, Z_{2,j}=0) = G[F_1(0), F_2(0)]$$

$$P(Z_{1,j}=z_1, Z_{2,j}=0) = G[F_1(z_1), F_2(0)] - G[F_1(z_1-1), F_2(0)] \quad \text{for } z_1=1, 2, \dots; j=1 \dots N$$

$$P(Z_{1,j}=0, Z_{2,j}=z_2) = G[F_1(0), F_2(z_2)] - G[F_1(0), F_2(z_2-1)] \quad \text{for } z_2=1, 2, \dots; j=1 \dots N$$

$$P(Z_{1,j}=z_1, Z_{2,j}=z_2) = G[F_1(z_1), F_2(z_2)] - G[F_1(z_1-1), F_2(z_2)] \\ - G[F_1(z_1), F_2(z_2-1)] + G[F_1(z_1-1), F_2(z_2-1)] \quad \text{for } z_1, z_2=1, 2, \dots; j=1 \dots N$$

The construction of the bivariate Poisson or negative binomial distributions in P2 and N2 using the Frank copula requires some comment. As noted above, this method allows for unrestricted (positive or negative) correlation between $Z_{1,j}$ and $Z_{2,j}$, depending on the ancillary parameter ϕ . In contrast, the standard bivariate Poisson distribution constructed by combining three random variables with univariate Poisson distributions, and several alternative formulations of the bivariate negative binomial distribution described by Kocherlakota and Kocherlakota (1992), are capable of

accommodating positive correlation only. In this case, the sample correlation between Z_{1j} and Z_{2j} is +0.2780. A positive correlation might reflect a tendency for teams to retaliate in kind if the opposing team is guilty of a particularly high level of foul play. Alternatively, a common opinion among pundits and supporters is that some referees, having penalised a player from one team, often look for an opportunity to penalise an opposing player soon afterwards, in an effort to pre-empt the formation on the part of managers, players or spectators of any perception of refereeing bias.

Such explanations notwithstanding, there appears to be no compelling case for defining the bivariate distributions in a way that would exclude the possibility of obtaining a negative correlation. In Section 5, we comment briefly on the comparison between Model P2 fitted using the Frank copula, and an equivalent model fitted using the standard bivariate Poisson distribution. A further advantage gained by using the Frank copula to construct both of the bivariate (Poisson and negative binomial) distributions is that direct comparisons can be drawn between these two specifications in tests for the significance of the overdispersion parameters κ_1 and κ_2 .

In the third and final formulation of the bivariate distribution, P3 and N3 are obtained by applying a zero-inflated adjustment to the joint probabilities of P2 and N2, respectively. The zero-inflated joint probabilities are $\tilde{P}(Z_{1j}=z_1, Z_{2j}=z_2) = (1-\pi)P(Z_{1j}=z_1, Z_{2j}=z_2) + \pi D(z_1, z_2)$, where $D(0,0)=1$ and $D(z_1, z_2)=0$ for $(z_1, z_2) \neq (0,0)$, and π is an additional ancillary parameter.

Below, the probability models P1-P3 and N1-N3 are used in estimations of the unconditional and conditional expectations of the incidence of disciplinary sanction against the home and away teams. In the unconditional models, we assume $\lambda_{ij}=\lambda_i$ for $i=1,2$ and $j=1 \dots N$. The probability models P1-P3 and N1-N3 are fitted directly to the sample data for $\{Z_{1j}, Z_{2j}\}$. In the conditional models, $\ln(\lambda_{ij})$ is specified as a linear function of a set of covariates. Although in principle it is possible to specify a conditional equation for the ancillary parameter ϕ as well, problems of non-convergence in the estimation were encountered if anything more than a very small number of covariates was used. In those estimations that did converge, the estimated coefficients in the equation for ϕ were insignificant.

Two statistical procedures are used to assess the quality of the fitted models. First, a goodness-of-fit test is used to compare the observed and values of $\{Z_{1j}, Z_{2j}\}$ in Table 3 with the

expected values obtained from each fitted model. For the unconditional estimations, a standard chi-square goodness-of-fit test is employed. For the conditional estimations, the adaptation of this test described by Heckman (1982) is employed. In both cases, the distribution of the test statistic is $\chi^2(25)$.

The second procedure involves hypothesis tests of zero restrictions on the ancillary parameters of P2-P3 and N1-N3. Failure to reject any such restrictions suggests that the more complex specification can be discarded in favour of the simpler specification. The restriction $\phi=0$ implies P2 or N2 can be discarded in favour of P1 or N1. This restriction should not be tested using the z-statistic on $\hat{\phi}$ or a standard likelihood ratio (LR) test, because the bivariate distribution (defined using the Frank copula) is inapplicable when $\phi=0$. However, the sample correlation between $Z_{1,j}$ and $Z_{2,j}$ reported above, together with indicators of the quality of the fitted models reported in Section 5 such as the maximised values of the log-likelihood function, $\ln(L)$, and the goodness-of-fit tests, suggest the contribution of the parameter ϕ to the quality of the fitted versions of P2 and N2 is not in any doubt. The restriction $\kappa_1=\kappa_2=0$, which implies N3, N2 or N1 can be discarded in favour of P3, P2 or P1, raises an issue for hypothesis testing discussed by Self and Liang (1987) and Andrews (2001). The values of κ_1 and κ_2 under the null hypothesis lie on the boundary of the relevant parameter space, so the standard regularity conditions fail to hold, and the standard LR statistic follows a non-standard distribution. The same issue arises in respect of the restriction $\pi=0$, which implies P3 or N3 can be discarded in favour of P2 or N2. In each case, this issue is addressed by generating p-values based on Monte Carlo simulations of the sampling distribution of the LR statistic under the null hypothesis. For each test, the simulated data are generated from the probability model that is applicable under the null hypothesis, with all parameters set to the values that are obtained when this model is fitted to the sample data.

4. Team quality and the incidence of disciplinary sanction

In Sections 4 and 5, we develop an empirical model for the determinants of $\lambda_{i,j}$ interpreted as the conditional expectations of the disciplinary points incurred by the home ($i=1$) and away ($i=2$)

teams in match j . In Section 4, we investigate the theoretical relationship between team quality and the incidence of disciplinary sanction. The aim of the theoretical analysis that follows is to derive an inverse relationship between the degree of uncertainty of match outcome, and the equilibrium levels of aggression contributed by both teams. Uncertainty of match outcome is a function of the variable q_j , defined as a weighted sum of the home team's win and draw probabilities for match j , $q_j = P(\text{home win in match } j) + 0.5P(\text{draw})$. A weighting of 0.5 is attached to the draw probability in order to ensure that q_j and the equivalent weighted sum for the away team add up to one. Analysis of bookmakers' odds and other statistical evidence suggests that the probability of a draw does not vary greatly from one match to another, and is not very sensitive to variations in the quality of the two teams (see, for example, Dobson and Goddard, 2001; Forrest et al., 2005). Therefore a convenient measure of uncertainty of match outcome is provided by $q_j(1 - q_j)$: the product of q_j and the equivalent weighted sum for the away team. This uncertainty of match outcome measure is maximised when $q_j=0.5$.

In the theoretical analysis, it is assumed that q_j depends on: the talent differential between the two teams; a home advantage effect; and a tactical decision variable representing the level of 'aggression' contributed by each team. Let $t_{i,j}$ and $a_{i,j}$ denote the playing talent and aggression level of team i in match j , respectively. $t_{i,j}$ and $a_{i,j}$ are scaled such that we can write $q_j = \Phi[t_{1,j} - t_{2,j} + \theta(a_{1,j}) - \theta(a_{2,j}) + h]$, where Φ is the standard Normal distribution function, h is a scalar that allows for home advantage, and θ is a continuous and twice-differentiable function, with $\theta'(a_{i,j}) > 0$, $\theta''(a_{i,j}) < 0$ for all $a_{i,j}$. At low levels, more aggression enhances a team's win probability. However, this relationship is subject to diminishing returns: beyond a certain point further aggression becomes counterproductive. For convenience, we shall write $q_j = \Phi[x_j + \theta(a_{1,j}) - \theta(a_{2,j})]$, where $x_j = t_{1,j} - t_{2,j} + h$.

It is also assumed that aggressive play by either team imposes a cost, represented by a continuous and twice-differentiable function $v(a_{i,j})$, with $v'(a_{i,j}) > 0$, $v''(a_{i,j}) > 0$ for all $a_{i,j}$. $v(a_{i,j})$ reflects the deleterious effect on future match results of player suspensions resulting from yellow or red cards awarded in the current match. $v(a_{i,j})$ is increasing in aggression, at an increasing rate as the level of aggression increases.

We assume both teams decide their own aggression levels independently and before the start of the match. We derive a Nash equilibrium for the aggression levels contributed by both teams. A distinction is drawn between this tactical decision concerning aggression, and any tendency to engage in foul play on a retaliatory or tit-for-tat basis once the match is underway. As discussed in Section 3, in the bivariate models this retaliatory effect is one of the factors captured by the parameter ϕ .

At the Nash equilibrium, each team selects its own aggression level, conditional on the other team's aggression level being taken as fixed at its current value. For example, consider team 1's choice of $a_{1,j}$, conditional on $a_{2,j}$ (and x_j). Team 1 selects $a_{1,j}$ to maximise the objective function $\pi_1(a_{1,j}; a_{2,j}, x_j) = \Phi[x_j + \theta(a_{1,j}) - \theta(a_{2,j})] - \Phi[x_j - \theta(a_{2,j})] - v(a_{1,j})$, representing the net benefit to team 1 of an aggression level of $a_{1,j}$, rather than zero aggression. The maximisation of this objective function yields team 1's reaction function, $a_{1,j} = r_1(a_{2,j}, x_j)$. A similar optimisation procedure yields team 2's reaction function, $a_{2,j} = r_2(a_{1,j}, x_j)$. The Nash equilibrium $\{a_{1,j}^*, a_{2,j}^*\}$ is located at the intersection of the two reaction functions, at which point $a_{1,j}^* = r_1(x_j, a_{2,j}^*)$ and $a_{2,j}^* = r_2(x_j, a_{1,j}^*)$.

Figure 1 illustrates the Nash equilibrium for the three cases $x_j=0, 0.5, 1$. The quadratic functional forms $\theta(a_{i,j})=a_{i,j}-0.2a_{i,j}^2$ and $v(a_{i,j})=0.2a_{i,j}+0.1a_{i,j}^2$ are used for illustrative purposes. The model is symmetric, so $a_{1,j}^* = a_{2,j}^*$ for any x_j , and the Nash equilibrium values for $x_j=-0.5, -1$ are the same as those for $x_j=0.5, 1$, respectively. The maximum numerical values for $\{a_{1,j}^*, a_{2,j}^*\}$ occur in the case $x_j=0$ where, taking account of talent and home advantage, the teams are equally balanced with identical win probabilities, so $q_j=0.5$. Figure 1 illustrates the following general property of the theoretical model: as the degree of competitive imbalance increases, the Nash equilibrium values $\{a_{1,j}^*, a_{2,j}^*\}$ decrease. If the match is evenly balanced, a little extra aggression by either team has a large effect on q_j , and the aggression levels of both teams are high at the Nash equilibrium. Conversely, if the match is unbalanced, a little extra aggression by either team has a small effect on q_j , and the aggression levels are low at the Nash equilibrium.

In the empirical model, a numerical value for q_j for each of the $N=2,660$ sample matches is generated from the ordered probit match results forecasting model developed by Goddard (2005).

This model generates probabilities for home win, draw and away win outcomes, based solely on historical data that is available prior to the match in question. The forecasting model's covariates are: the win ratios of both teams over the 24 months prior to the current match; both teams' recent home and away match results; dummy variables indicating the significance of the match for end-of-season outcomes (championship, European qualification and relegation); dummy variables indicating current involvement in the FA Cup; both teams' recent average home attendances; and the geographic distance separating the teams' home towns. To generate match result probabilities for each season (1996-7 to 2002-3 inclusive), seven versions of the forecasting model are estimated, using data for the preceding 15 seasons in each case. Full details of the forecasting model are reported in Goddard (2005) and are not repeated here.

The empirical model allows for two forms of relationship between q_j and the incidence of disciplinary sanction. First, a weaker team that is forced to defend for long periods can be expected to commit more fouls than a stronger team that spends more time attacking. This suggests a negative (positive) linear relationship between q_j and the disciplinary points incurred by the home (away) team. Second, the theoretical analysis developed in Section 4 suggests there is also a non-linear dimension to the relationship between q_j and the incidence of disciplinary sanction. In the empirical model, this is represented by the quadratic covariate $q_j(1 - q_j)$. A positive relationship is expected between this covariate and the incidence of disciplinary sanction against both teams.

5. Estimation results and tests for refereeing bias and inconsistency

Section 5 reports the estimation results for the unconditional and conditional expectations of the disciplinary points dependent variable. All estimations were carried out using the maximum likelihood estimation procedure in *Stata* 9.1. Summary estimation results based on probability models P1-P3 and N1-N3 are reported in Table 4.

For the unconditional estimations reported in the upper panel of Table 4, in all cases the estimated values of the parameters λ_1 and λ_2 are the sample means of $Z_{i,j}$, $\hat{\lambda}_1 = 1.4650$ and $\hat{\lambda}_2 = 2.045$.

In the unconditional estimations, the inclusion of additional ancillary parameters invariably produces large improvements in the quality of the fitted model, according to both the goodness-of-fit test and the simulated p-values for the LR statistics. The bivariate Poisson distribution defined by fitting the Frank copula, P2, is also found to offer a marginal improvement over the model fitted the standard bivariate Poisson distribution (which is not reported in Table 4): the maximised values of the log-likelihood function are $\ln(L)=-8753.4$ for the former, and $\ln(L)=-8753.9$ for the latter. In N3, the simulated p-values indicate that $H_0:\pi=0$ can be rejected (N3 is preferred to N2) and $H_0:\kappa_1=\kappa_2=0$ can be rejected (N3 is preferred to P3). In the goodness-of-fit tests based on the unconditional estimations, N3 is the only specification for which the null hypothesis is not rejected at the 0.01 level.

In the conditional estimations reported in the lower panel of Table 4, the ancillary parameter φ appears to produce a significant improvement in the quality of the fitted model (P2 and N2 dominate P1 and N1, respectively). However, the improvements produced by the other ancillary parameters κ_1 , κ_2 and π are relatively small. Using simulated p-values, we fail to reject $H_0:\kappa_1=\kappa_2=0$ in respect of N2 and N3. At the 0.01 level, we also fail (narrowly) to reject $H_0:\pi=0$ in P3. However, the goodness-of-fit test rejects the null in P2, but fails to do so in P3.

On the balance of these results, we select P3, the zero-inflated bivariate Poisson, as our chosen probability model, to be used as the basis for the conditional estimations that are reported in full below, and the hypothesis tests that follow. In the unconditional estimations, the use of the negative binomial probability model (with a zero-inflated adjustment) is required to represent the overdispersion in the sample data for $\{Z_{1,j}, Z_{2,j}\}$. In the conditional estimations, however, the covariates appear to be largely successful in identifying the sources of overdispersion, rendering the use of the more complex negative binomial probability model unnecessary.

In the rest of Section 5, we report the estimated model for the conditional expectations of the numbers of disciplinary points incurred by the home and away teams. In order to ensure the non-negativity of the fitted values of $\lambda_{i,j}$, $\ln(\lambda_{i,j})$ is used as the dependent variable. $\ln(\lambda_{i,j})$ is assumed to depend on covariates that vary from match to match. The team quality covariates q_j and $q_j(1 - q_j)$ have

been described in Section 4. The remaining covariate definitions are shown below. Table 5 reports summary descriptive statistics for $\{Z_{1,j}, Z_{2,j}\}$ and for all of the covariates.

$\text{sig}_{i,j}$ = 0-1 dummy variable, coded 1 if match j is significant for end-of-season championship, European qualification or relegation outcomes, for the home ($i=1$) or away ($i=2$) team.

$\text{DM}_{i,m,j}$ = 1 if match j falls within managerial spell m for the home ($i=1$) or away ($i=2$) team; 0 otherwise ($m=1\dots 56$ represents managerial spells that contained at least 30 Premier League matches within the observation period; the matches in 24 other spells that contained fewer than 30 matches in total form the reference category).

$\text{DR}_{r,j}$ = 1 if match j is officiated by referee r ; 0 otherwise ($r=1\dots 28$ represents referees who officiated at least 30 Premier League matches within the observation period; nine other referees who officiated fewer than 30 matches each form the reference category).

$\text{DS}_{s,j}$ = 1 if match j is played in season s ; 0 otherwise (s represents seasons 1997-8 to 2002-3 inclusive; 1996-7 is the reference category).

att_j = reported attendance at match j .

sky_j = 1 if match j was televised live by BSkyB; 0 otherwise.

The model specification allows for tests of several hypotheses concerning patterns in the incidence of disciplinary sanction. The principal hypotheses of interest are as follows:

H1: The *home advantage hypothesis*. The tendency for away teams to incur more disciplinary points than home teams is solely a corollary of home advantage: the tendency for home teams to win more frequently than away teams.

H2: The *refereeing consistency hypothesis*. The average incidence of disciplinary sanction does not vary between referees.

H3: The *consistent home team bias hypothesis*. The degree to which away teams incur more disciplinary points than home teams on average (after controlling for home advantage) does not vary between referees.

H4: The *time consistency hypothesis*. The average incidence of disciplinary sanction is stable over time.

H5: The *audience neutrality hypothesis*. The incidence of disciplinary sanction is invariant to the size of the crowd inside the stadium, and is the same notwithstanding whether the match is broadcast live on TV.

The estimated conditional equations for $\ln(\hat{\lambda}_{1,j})$ based on P3 are reported below as equations (5) and (6). z-statistics for the estimated coefficients, based on robust standard errors, are shown in parentheses (intercept and dummy variable coefficients are not reported). The estimation results are interpreted and discussed below. Robust inference is used throughout, in the tests of H1-H5.

$$\begin{aligned} \ln(\hat{\lambda}_{1,j}) = & \hat{\alpha}_{1,0} - 0.6137 q_j + 5.0195 q_j(1 - q_j) - 0.0122 \text{sig}_{1,j} + 13.5467 \text{att}_j + 0.0242 \text{sky}_j \\ & (-3.16) \quad (5.37) \quad (-0.23) \quad (3.14) \quad (0.54) \\ & + \sum_{s=1}^6 \hat{\beta}_{1,s} \text{DS}_{s,j} + \sum_{m=1}^{56} \hat{\delta}_{1,m} \text{DM}_{i,m,j} + \sum_{r=1}^{28} \hat{\gamma}_{1,r} \text{DR}_{r,j} \end{aligned} \quad (5)$$

$$\begin{aligned} \ln(\hat{\lambda}_{2,j}) = & \hat{\alpha}_{2,0} + 0.8557 q_j + 3.0241 q_j(1 - q_j) + 0.1300 \text{sig}_{2,j} + 1.9502 \text{att}_j + 0.0050 \text{sky}_j \\ & (3.40) \quad (3.75) \quad (2.71) \quad (1.11) \quad (0.13) \\ & + \sum_{s=1}^6 \hat{\beta}_{2,s} \text{DS}_{s,j} + \sum_{m=1}^{56} \hat{\delta}_{2,m} \text{DM}_{i,m,j} + \sum_{r=1}^{28} \hat{\gamma}_{2,r} \text{DR}_{r,j} \end{aligned} \quad (6)$$

Relative team quality and home advantage

The specification of (5) and (6) defines a quadratic functional form for the relationship between q_j and $\ln(\hat{\lambda}_{i,j})$. This relationship is parameterised so that the coefficients reported are for q_j , a weighted sum of the home team's win and draw probabilities after allowing for home advantage, and $q_j(1 - q_j)$, a measure of competitive balance. In order to validate the assumed quadratic functional form, preliminary estimations of (5) and (6) were carried out with the terms in q_j and $q_j(1 - q_j)$ replaced by ten 0-1 dummy variables identifying observations with values of q_j in the following bands: $q_j \leq 0.4$, $0.4 < q_j \leq 0.44$, $0.44 < q_j \leq 0.48$, and so on until $q_j > 0.76$. Table 6 compares the estimated values for $\hat{\lambda}_{i,j}$ that are produced by the non-parametric (dummy variables) formulation and the parametric (quadratic) formulation, for the numerical values of q_j located at the mid-points of the bands, and with all other covariates set to their sample means. Although there is some unevenness in the expected values obtained using the non-parametric formulation, the quadratic functional form

appears to provide a good approximation to the underlying shape of this relationship. The difference between the maximised values of the log-likelihood function, $\ln(L)=-8437.9$ for the non-parametric formulation, and $\ln(L)=-8439.6$ for the parametric formulation, is small. The quadratic functional form locates the maxima of $\hat{\lambda}_{1,j}$ and $\hat{\lambda}_{2,j}$ with respect to q_j at $q_j=0.439$ and $q_j=0.641$, respectively.

The *home advantage hypothesis* H1 asserts that the propensity for away teams to collect more disciplinary points on average than home teams is solely a corollary of the home advantage effect on match results. If so, the expected incidence of disciplinary sanction for a (relatively strong) away team should be the same as that for a (relatively weak) home team, if the two teams' win probabilities (taking home advantage into account) are the same. Under H1, all coefficients in (5) should be identical to their counterparts in (6), except the coefficients on q_j which should be equal and opposite in sign (because the weighted sum of the away team's win probability and the draw probability is one *minus* this weighted sum for the home team, or $1 - q_j$). An implication of H1 is $\lambda_{1,j}-\lambda_{2,j}=0$ if $q_j=0.5$, $q_j(1 - q_j)=0.25$, and all other covariates in (5) and (6) are set to their sample means. Alternatively, $\lambda_{1,j}-\lambda_{2,j}<0$ if q_j is set to its sample mean ($\bar{q}=0.6025$) and $q_j(1 - q_j)=\bar{q}(1 - \bar{q})=0.2395$. Under H1, the observed difference between the average disciplinary points incurred by the home and away teams should be due solely to the home advantage effect (on average, the home team has a higher weighted sum of win and draw probabilities than the away team).

For $q_j=0.5$ the procedure described above yields $\hat{\lambda}_{1,j} - \hat{\lambda}_{2,j} = -0.285$. The robust standard error, calculated using the delta method (Oehlert, 1992), is 0.061. For $q_j=\bar{q}$, this procedure yields $\hat{\lambda}_{1,j} - \hat{\lambda}_{2,j} = -0.585$ (s.e.=.044). To check the robustness of these results, we also estimated (5) and (6) with all covariates and dummy variables other than q_j and $q_j(1 - q_j)$ excluded, and repeated the calculation. The results were similar: $\hat{\lambda}_{1,j} - \hat{\lambda}_{2,j} = -0.386$ (s.e.=.048) for $q_j=0.5$, and $\hat{\lambda}_{1,j} - \hat{\lambda}_{2,j} = -0.610$ (s.e.=.044) for $q_j=\bar{q}$. In both cases, $\hat{\lambda}_{1,j} - \hat{\lambda}_{2,j}$ is significantly less than zero for $q_j=0.5$. Therefore the tendency for away teams to endure a higher incidence of disciplinary sanction cannot be explained solely by the home advantage effect, although this effect does contribute towards the

observed pattern. H1 is also rejected by a Wald test of the appropriate cross-equation equality restrictions on the coefficients of (5) and (6), which yields $\chi^2(96)=171.5$ (p-value=.0000).

Other controls for team behaviour

In order to isolate the contribution of referees to the variation in the incidence of disciplinary sanction, the conditional model includes a number of additional covariates that control for the effects of team behaviour. The contribution to the model of these controls is examined in this subsection.

The incidence of disciplinary sanction for either team might be affected by the importance of the match for end-of-season championship, European qualification or relegation outcomes. A team that still has end-of-season issues at stake might be expected to be more determined or aggressive than a team with nothing at stake. In the definitions of the dummy variables $sig_{i,j}$, the algorithm that determines whether a match is significant for either team assesses whether it is arithmetically possible (before the match is played) for the team to win the championship, qualify for European competition or be relegated, if all other teams currently in contention for the same outcome take one point on average from each of their remaining fixtures. Alternative algorithms, based on more optimistic or pessimistic assumptions concerning the average performance of competing teams over their remaining fixtures, alter the classification of a small proportion of matches at the margin, but the implications of such minor variations for the estimation results reported here are negligible.

The coefficient on $sig_{1,j}$ in (5) is insignificant, but the coefficient on $sig_{2,j}$ in (6) is positively signed and significant at the 0.01 level. A possible interpretation is that away teams feel able to ‘ease off’ in unimportant end-of-season matches; but home teams, perhaps conscious of their own crowd’s critical scrutiny, feel obliged to demonstrate maximum commitment at all times, even when no end-of-season issues are at stake.

Differences between football teams in playing personnel, styles of play and tactics represent a further possible source of variation in the incidence of disciplinary sanction. With 22 players (plus substitutes) participating in every match, in an empirical analysis at match level it is impossible to control for every change of playing personnel. In preliminary experiments with the model specification, we encountered a tendency for estimations including separate dummy variables for each

team in each season to fail to converge, due to the excessive number of coefficients. Therefore we have chosen to use managerial spells as a proxy for football team-related factors that might produce differences in the incidence of disciplinary sanction. This can be justified on the grounds that managers are primarily responsible for tactics and playing styles. Casual observation suggests managerial change is a good proxy for turnover of playing personnel: the removal of a manager is often followed by high player turnover, as the new incumbent seeks to reshape his squad in accordance with his own preferences. A Wald test of $H_0: \delta_{i,m} = 0$ for $i=1,2$ and $m=1 \dots 56$ in (5) and (6) yields $\chi^2(112) = 264.0$ (p-value = .0000), suggesting that choices of personnel and tactics made by managers do have a highly significant effect on the incidence of disciplinary sanction.

Individual referee effects

Inconsistency in the standards applied by different referees is among the most frequent causes of complaint from football managers, players, supporters and media pundits. Table 7 summarises the average numbers of disciplinary points per match awarded against the home and away teams and against both teams combined, by each of the 28 referees who officiated at least 30 Premier League matches during the observation period. (The data for a further nine referees who each officiated fewer than 30 matches are excluded from Table 7.) There appears to be considerable variation between the propensities for individual referees to take disciplinary action. For example, the most lenient referee (Keith Burge) averaged 2.526 disciplinary points per match over 57 matches, and the most prolific (Mike Reed) averaged 4.541 points over 85 matches.

Does this degree of variation in the incidence of disciplinary sanction per referee constitute statistical evidence of inconsistency in refereeing standards? H2, the *refereeing consistency hypothesis*, imposes zero restrictions on the coefficients on the individual referee dummy variables $DR_{r,j}$, which identify matches officiated by the 28 referees listed in Table 7. In (5) and (6), a Wald test of $H_0: \gamma_{i,r} = 0$ for $i=1,2$ and $r=1 \dots 28$ yields $\chi^2(56) = 171.3$ (p-value = .0000). Therefore H2 is rejected, suggesting there was significant variation in standards between referees. Since the conditional model includes controls for team quality and other potential influences on the incidence of disciplinary

sanction, the rejection of H2 should not be attributable to any non-randomness in the assignment of referees to matches: for example, the tendency for referees with a reputation for toughness to be assigned to matches at which disciplinary issues are anticipated by the authorities. However, it is acknowledged that the use of 0-1 dummy variables to model the individual referee effects may represent a simplification: for example, it does not allow for duration dependence in referees' performance, which might arise if referees modify their behaviour as they gain experience, or if the removal of unsatisfactory referees by the football authorities introduces a form of survivorship effect.

The rejection of H1, the *home advantage hypothesis*, suggests there is a bias favouring the home team in the incidence of disciplinary sanction, even after controlling for home advantage in match results. With H2 also having been rejected, it is relevant to examine whether there are significant differences between referees in the degree of home team bias. In other words, do variations in the degree of home team bias on the part of different officials contribute to the observed pattern of refereeing inconsistency? H3, the *consistent home team bias hypothesis*, imposes the restriction that the corresponding coefficients on the individual referee dummy variables in the home and away team equations are the same. H3 would imply that the rate at which away teams tend to incur more disciplinary points than home teams does not vary between referees. In (5) and (6), a Wald test of $H_0: \gamma_{1,r} = \gamma_{2,r}$ for $r=1 \dots 28$ yields $\chi^2(28)=52.21$ (p-value=.0036). Therefore H3 is rejected at a significance level of 0.01 (but not at a significance level of 0.001).

Season effects

The individual football season dummy variables $DS_{s,j}$ are included in the conditional model primarily as a control for changes over time in the content and interpretation of the rules relating to the award of yellow and red cards. The key changes during the observation period are detailed in Table 8. Most of the changes have increased the range of offences that are subject to disciplinary sanction, although there has occasionally been movement in the opposite direction.

Table 9 reports the average numbers of yellow and red cards awarded against the home and away teams per match by season. There appears to be little or no trend in the overall incidence of disciplinary sanction, despite the increase in the range of sanctionable offences. Two possible

explanations are as follows. First, whenever there is an addition to the list of sanctionable offences, players may modify their behaviour so that the numbers of cautions and dismissals remain approximately constant (Witt, 2005). Second, referees may tend to modify their interpretation of the boundaries separating non-sanctionable from sanctionable offences, and those separating cautionable from dismissable offences, so as to maintain an approximately constant rate of disciplinary sanction.

The directive issued at the start of the 1998-9 season making the tackle from behind punishable by automatic dismissal is the only rule change that appears to have had a discernible impact on the data summarised in Table 9. The mean incidence of disciplinary sanction is higher for 1998-9 than for any of the other six seasons in the observation period. Within the 1998-9 season as well, the process of adjustment to the new disciplinary regime is visible in the data: during the first three months of this season the average disciplinary points incurred by both teams per match was 4.336, while the average for the rest of the season was 3.883 (see also Witt, 2005). In subsequent seasons, although this directive remained in force, the incidence of disciplinary sanction returned to levels similar to those experienced before the directive came into effect.

In order to test H4, the *time consistency hypothesis* that the average incidence of disciplinary sanction is stable over time, the null hypothesis (expressed in terms of the coefficients of the conditional model) is $H_0: \beta_{i,s} = 0$ for $i=1,2$ and $s=1997-8$ to $2002-3$ (inclusive). A Wald test yields $\chi^2(12)=34.34$ (p-value=.0006), suggesting there was significant season-to-season variation in the incidence of disciplinary sanction. However, if the zero restrictions on the coefficients for 1998-9 are excluded from the null ($H_0: \beta_{i,s} = 0$ for $i=1,2$ and $s=1997-8$ and $1999-2000$ to $2002-3$ inclusive), the Wald test yields $\chi^2(10)=12.56$ (p-value=.2491). This suggests that with the (temporary) exception of the 1998-9 season, there was no other significant season-to-season variation in the incidence of disciplinary sanction. H4 receives qualified support from the estimation results.

Match attendance and live TV broadcast

Under H5, the *audience neutrality hypothesis*, the incidence of disciplinary sanction is unaffected by the crowd inside the stadium, and is also the same notwithstanding whether the match is

being broadcast live on TV. To control for crowd effects, the covariate att_j , defined as the reported attendance at match j , is included in the regressions for $\ln(\lambda_{i,j})$. If $H5$ is not supported in respect of the stadium audience, more than one prior concerning the direction of any effect is possible. A large attendance might be expected to add to the intensity or excitement of the occasion, resulting in more determined or aggressive play by either or both teams. Alternatively, a large attendance, presumably dominated by supporters of the home team, might put pressure on the referee to treat disciplinary transgressions by the home team more leniently, and those by the away team more severely.

The coefficient on att_j in (5) is positive and significant at the 0.01 level. The equivalent coefficient in (6) is also positive, but insignificant. With respect to the stadium audience $H5$ is rejected, but there is no evidence of any tendency for referees to treat the home team more leniently when the crowd size is larger; if anything, the opposite seems to apply.

The satellite broadcaster BSkyB held the Premier League's live TV broadcasting rights throughout the observation period. These rights permitted BSkyB to screen between 60 matches per season (at start of the observation period) and around 100 (by the end). The total number of scheduled Premier League matches per season is 380. If $H5$ is not supported, a tendency for players or referees to 'play to the camera' might be discernible in a different incidence of disciplinary sanction between televised and non-televised matches. However, both coefficients on sky_j in (5) and (6) are positive but insignificant. Therefore $H5$ is supported in respect of the live TV audience, with no evidence that the behaviour of players or referees is affected when the match is broadcast live on TV.

6. Conclusion

In this paper, we have reported estimations for the unconditional and conditional expectations of the incidence of disciplinary sanction against footballers in English Premier League matches. A comprehensive statistical analysis of patterns in the award of yellow and red cards over a seven-year period aims to provide the football authorities and other interested parties with a firmer factual basis than has been available previously for policy decisions and debate concerning the interpretation and implementation by referees of the rules governing disciplinary sanction in professional football.

In the estimations of the conditional expectations of the numbers of disciplinary points incurred by the home and away teams, it is found that relative team strengths matter: underdogs tend to incur a higher rate of disciplinary sanction than favourites. The incidence of disciplinary sanction tends to be higher in matches between evenly balanced teams, in matches with end-of-season outcomes at stake, and in matches that attract high attendances. Home teams appear to play more aggressively in front of larger crowds, but perhaps surprisingly the crowd size does not influence the incidence of disciplinary sanction against the away team. There is no evidence that the behaviour of players or referees is any different in live televised matches.

Despite an increase over time in the number of offences subject to disciplinary sanction, there was no consistent time-trend in the yellow and red cards data: players and officials appear to have adjusted to changes in the rules so that in the long run the rate of disciplinary sanction remained approximately constant. Individual referee effects make a significant contribution to the explanatory power of the conditional model, indicating that there are inconsistencies between referees in the interpretation or application of the rules. An obvious but important policy implication for the football authorities is that steps need to be taken in order to improve refereeing consistency.

The empirical analysis suggests that the tendency for away teams to incur more disciplinary points than home teams cannot be explained solely by the home advantage effect on match results. Even after controlling for team quality, a (relatively strong) away team can expect to collect more disciplinary points than a (relatively weak) home team with the same win probability. Therefore the statistical evidence seems to point in the direction of a home team bias in the incidence of disciplinary sanction. This interpretation is consistent with evidence of home team bias in several other recent studies, which find that the home team is favoured in the calling of fouls, or in the addition of stoppage time at the end of matches. Finally, evidence is found of variation between referees in the degree of home team bias; and this variation contributes to the overall pattern of refereeing inconsistency. These findings suggest that while all referees should be counselled and encouraged to avoid (presumably unintentional) home team bias in their decision-making, the extent to which corrective action is required is also likely to vary between officials.

References

- Allen, W.D. (2002). Crime, punishment and recidivism: lessons from the National Hockey League. *Journal of Sports Economics* 3, 39-60.
- Andrews, D.W.K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica* 69, 683-734.
- Cain, M., Law, D. and Peel, D. (2000). The favourite-longshot bias and market efficiency in UK football betting. *Scottish Journal of Political Economy* 47, 25-36.
- Dixon, M.J. and Coles, S.C. (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics* 46, 265-280.
- Dixon, M.J. and Pope, P.F. (2004). The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting* 20, 697-711.
- Dobson, S. and Goddard, J. (2001). *The economics of football*. Cambridge: Cambridge University Press.
- Forrest, D., Goddard, J. and Simmons, R. (2005). Odds setters as forecasters: the case of English football. *International Journal of Forecasting* 21, 551-564.
- Garicano, L. and Palacios-Huerta, I. (2000). An empirical examination of multidimensional effort in tournaments. Graduate School of Business, University of Chicago, mimeo.
- Garicano, L., Palacios-Huerta, I. and Prendergast, C. (2001). *Favoritism under social pressure*. NBER Working Paper 8376.
- Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting* 21, 331-340.
- Goddard, J. and Asimakopulos, I. (2004). Forecasting football match results and the efficiency of fixed-odds betting. *Journal of Forecasting* 23, 51-66.
- Heckelman, J.C. and Yates, A.C. (2002). And a hockey game broke out: crime and punishment in the NHL. *Economic Inquiry* 41, 705-712.
- Heckman, J.J. (1984). The χ^2 goodness of fit statistic for models with parameters estimated from microdata. *Econometrica* 52, 1543-1547.
- Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *The Statistician* 52, 381-393.
- Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate discrete distributions*. New York: Marcel Dekker.
- Lee, A. (1999). Modelling rugby league data via bivariate negative binomial regression. *Australian and New Zealand Journal of Statistics* 41, 153-171.
- Maher, M.J. (1982). Modelling association football scores. *Statistica Neerlandica* 36, 109-118.
- McCormick, R.E. and Tollison, R.D. (1984). Crime on the court. *Journal of Political Economy* 92, 223-235.

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

Oehlert, G.W. (1992). A note on the delta method. *American Statistician* 46, 27-29.

Ridder, G., Cramer, J.S. and Hopstaken, P. (1994). Down to ten: estimating the effect of a red card in soccer. *Journal of the American Statistical Association* 89, 1124-1127.

Self, S.G and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82, 605-610.

Sutter, M. and Kocher, M.G. (2004). Favoritism of agents – the case of referees' home bias. *Journal of Economic Psychology* 25, 461-469.

Torgler, B. (2004). The economics of the FIFA football World Cup. *Kyklos* 57, 287-300.

Witt, R. (2005). Do players react to anticipated sanction changes? Evidence from the English Premier League. *Scottish Journal of Political Economy* 52, 623-640.

Tables

Table 1 Observed numbers of yellow cards incurred by the home and away teams, English Premier League, seasons 1996-7 to 2002-3

Home team	Away team								Total
	0	1	2	3	4	5	6	7	
0	189	254	158	86	35	9	1	0	732
1	110	260	264	147	66	23	6	1	877
2	64	162	158	126	47	25	6	1	589
3	18	77	96	72	39	14	3	4	323
4	3	13	29	32	16	8	2	0	103
5	1	3	12	11	2	1	0	1	31
6	0	0	1	2	1	1	0	0	5
Total	385	769	718	476	206	81	18	7	2660

Source: The Football Association.

Table 2 Observed numbers of red cards incurred by the home and away teams, English Premier League, seasons 1996-7 to 2002-3

Home team	Away team			Total
	0	1	2	
0	2258	231	8	2497
1	119	34	3	156
2	2	3	0	5
3	2	0	0	2
Total	2381	268	11	2660

Source: The Football Association.

Table 3 Sample frequency distribution for the bivariate disciplinary points dependent variable, $\{Z_{1,j}, Z_{2,j}\}$

$Z_{1,j}$	$Z_{2,j}$						Total
	0	1	2	3	4	5+	
0	182	235	141	84	45	18	705
1	104	244	238	137	66	49	838
2	65	150	140	121	51	57	584
3	17	71	92	73	48	34	335
4	4	18	35	42	18	16	133
5+	2	7	17	20	10	9	65
Total	374	725	663	477	238	183	2660

Table 4 Unconditional and conditional models: summary estimation results

	ϕ	κ_1	κ_2	π	$\ln(L)$	$\chi^2(25)$	LR: $\kappa_1=\kappa_2=0$	LR: $\pi=0$
Unconditional models								
P1 double Poisson	-	-	-	-	-8866.3	367.6 <i>.0000</i>	-	-
P2 bivariate Poisson	-1.7847	-	-	-	-8753.4	94.6 <i>.0000</i>	-	-
P3 zero-inflated biv.P.	-1.5381	-	-	.0278	-8739.2	64.3 <i>.0000</i>	-	30.2 <i>.0000</i>
N1 double n.b.	-	.1231	.0523	-	-8841.2	272.1 <i>.0000</i>	50.2 <i>.0000</i>	-
N2 bivariate n.b.	-1.8915	.1259	.0548	-	-8727.9	45.0 <i>.0084</i>	52.8 <i>.0000</i>	-
N3 zero-inf. biv .n.b.	-1.7317	.1052	.0386	.0151	-8725.5	40.1 <i>.0287</i>	27.4 <i>.0000</i>	4.8 <i>.0028</i>
Conditional models								
P1 double Poisson	-	-	-	-	-8527.1	232.2 <i>.0000</i>	-	-
P2 bivariate Poisson	-1.7695	-	-	-	-8441.4	45.5 <i>.0073</i>	-	-
P3 zero-inflated biv.P.	-1.6795	-	-	.0097	-8439.6	40.1 <i>.0285</i>	-	3.6 <i>.0128</i>
N1 double n.b.	-	.0217	.0000	-	-8526.4	229.3 <i>.0000</i>	1.4 <i>.1623</i>	-
N2 bivariate n.b.	-1.7941	.0278	.0026	-	-8440.2	43.5 <i>.0123</i>	2.4 <i>.0912</i>	-
N3 zero-inf. biv .n.b.	-1.7063	.0191	.0000	.0084	-8439.1	39.7 <i>.0313</i>	1.0 <i>.2095</i>	2.2 <i>.0504</i>

Note: $\ln(L)$ is the maximised value of the log-likelihood function.

$\chi^2(25)$ is the chi-square goodness-of-fit statistic, described in Section 3.

LR: $\kappa_1=\kappa_2=0$ is the LR (likelihood ratio) statistic for a test to compare the negative binomial model (N1, N2 or N3) with the corresponding Poisson model (P1, P2 or P3).

LR: $\pi=0$ is the LR statistic for a test to compare the zero-inflated model (P3 or N3) with the corresponding non-inflated model (P2 or N2).

p-values for $\chi^2(25)$ and simulated p-values for the LR statistics are shown in italics.

Table 5 Descriptive statistics: sample data

	$\lambda_{1,j}$	$\lambda_{2,j}$	q_j	$q_j(1 - q_j)$	att_j	$sig_{1,j}$	$sig_{2,j}$	sky_j
Mean	1.4650	2.0451	.6025	.2244	31.7	.8692	.8650	.1695
Variance	1.7216	2.2657	.0151	.00088	123.4	-	-	-
St. dev.	1.3121	1.5052	.1228	.0296	11.1	-	-	-
1st quartile	0	1	.5233	.2150	23.0	-	-	-
Median	1	2	.6094	.2352	31.2	-	-	-
3rd quartile	2	3	.6814	.2458	38.0	-	-	-
Minimum	0	0	.1331	.0817	7.7	-	-	-
Maximum	10	9	.9103	.2500	67.7	-	-	-
No. of 0's	-	-	-	-	-	348	359	2209
No. of 1's	-	-	-	-	-	2312	2301	451

Table 6 Non-parametric and parametric representations of the relationship between q_j and $(\hat{\lambda}_{1,j}, \hat{\lambda}_{2,j})$

q_j	0.38	0.42	0.46	0.50	0.54	0.58	0.62	0.66	0.70	0.74	0.78
Non-parametric											
$\hat{\lambda}_{1,j}$	1.636	1.699	1.760	1.691	1.504	1.681	1.454	1.426	1.226	1.226	0.674
$\hat{\lambda}_{2,j}$	1.680	1.803	2.068	1.877	1.936	2.147	2.074	2.181	2.113	1.924	1.993
Parametric											
$\hat{\lambda}_{1,j}$	1.708	1.735	1.734	1.705	1.651	1.572	1.474	1.359	1.234	1.102	0.969
$\hat{\lambda}_{2,j}$	1.719	1.823	1.914	1.990	2.049	2.090	2.111	2.112	2.092	2.053	1.995

Note: The non-parametric representation is obtained by estimating (5) and (6) with the covariates q_j and $q_j(1 - q_j)$ replaced by ten 0-1 dummy variables for banded values of q_j . The values for q_j shown in the top row are the central values in each band (except for the bottom and top bands). The parametric representation is based on the estimated version of (5) and (6) reported in Section 5. The values reported are the fitted values of the dependent variable in each case, with all other covariates set to their sample mean values.

Table 7 Average total disciplinary points awarded per match, by referee

Referee	Matches	Disciplinary points awarded			Referee	Matches	Disciplinary points awarded		
		Home team	Away team	Total			Home team	Away team	Total
1 Reed	85	1.788	2.753	4.541	15 Bennett	68	1.603	1.853	3.456
2 Willard	60	1.900	2.350	4.250	16 Barry	117	1.385	2.060	3.444
3 Barber	147	1.728	2.463	4.190	17 Jones	112	1.411	1.991	3.402
4 Riley	131	1.626	2.511	4.137	18 Ashby	33	1.212	2.152	3.364
5 Harris	52	1.750	2.327	4.077	19 Wilkie	81	1.358	1.975	3.333
6 Knight	41	1.829	2.171	4.000	20 Dunn	136	1.368	1.956	3.324
7 Styles	56	1.929	2.018	3.946	21 Elleray	129	1.295	1.984	3.279
8 Rennie	94	1.819	2.096	3.915	22 Winter	143	1.231	1.979	3.210
9 Dean	54	1.685	2.111	3.796	23 Gallagher	122	1.262	1.918	3.180
10 Wilkes	30	1.400	2.333	3.733	24 Halsey	74	1.338	1.730	3.068
11 D'urso	85	1.624	2.094	3.718	25 Alcock	78	1.000	2.026	3.026
12 Poll	160	1.619	2.069	3.688	26 Wiley	90	1.433	1.578	3.011
13 Bodenham	44	1.455	2.045	3.500	27 Durkin	145	1.248	1.469	2.717
14 Lodge	102	1.392	2.108	3.500	28 Burge	57	0.877	1.649	2.526

Source: The Football Association.

Note: Referees who officiated at fewer than 30 Premier League matches between the 1996-7 and 2002-3 seasons (inclusive) are not shown in Table 7.

Table 8 Rule changes and changes of interpretation, by season

SEASON	RULE CHANGES/CHANGES OF INTERPRETATION
1996-7	Referees are reminded to severely punish the tackle from behind.
1997-8	Failure to retreat the required distance at free kicks and delaying the restart of play are to be interpreted as yellow card offences.
1998-9	The tackle from behind which endangers the safety of an opponent is to be interpreted as a red card offence. The red card offence of denying an opponent a goal scoring opportunity is changed to denying an opposing team a goal scoring opportunity (widening the scope of this offence).
1999-2000	Simulation (diving, feigning injury or pretending that an offence has been committed) is to be punishable with a yellow card. Referees are reminded to punish racist remarks with a red card. Swearing is also an offence warranting a red card.
2000-1	Offensive gestures are to be punishable with a red card.
2001-2	Some relaxation of the rule requiring referees to issue a yellow card if a player celebrates a goal by removing his shirt. However, celebrations that are provocative, inciting, ridiculing of opponents or spectators or time wasting remain punishable with a yellow card. Referees are reminded to punish intentional holding or pulling offences with a yellow card.
2002-3	Referees are reminded to be strict in punishing simulation and the delaying of restarts, especially if players remove shirts for any length of time celebrating a goal.

Source: *Rothmans Football Yearbook* (various editions).

Table 9 Average numbers of yellow and red cards and total disciplinary points awarded per match, by season

Season	Home team			Away team			Both teams		
	Yellow	Red	Total	Yellow	Red	Total	Yellow	Red	Total
1996-7	1.305	0.026	1.350	1.808	0.084	1.934	3.113	0.111	3.284
1997-8	1.303	0.058	1.405	2.016	0.124	2.189	3.318	0.182	3.595
1998-9	1.582	0.074	1.695	2.147	0.116	2.316	3.729	0.189	4.011
1999-2000	1.411	0.055	1.497	1.932	0.129	2.118	3.342	0.184	3.616
2000-1	1.355	0.084	1.487	1.800	0.084	1.921	3.155	0.168	3.408
2001-2	1.247	0.084	1.389	1.803	0.103	1.955	3.050	0.187	3.345
2002-3	1.326	0.071	1.432	1.703	0.124	1.882	3.029	0.195	3.313

Source: The Football Association.

