Association Internationale des Economistes du Sport
AIES IASE
International Association of Sports Economists

Working Paper Series, Paper No. 06-19

Tilting the Playing Field (Why a sports league planner would choose less, not more, competitive balance): The case of English Football

Stefan Szymanski[1] and Stephanie Leach[2]

October 2006

Abstract

It has traditionally been argued that the organizer of a sports league would prefer more competitive balance to the level that emerges in a noncooperative equilibrium. This argument has been used to justify restraints on competition between teams, which also tend to raise profits at the expense of players and consumers. This paper shows that in theory a planner would prefer less, not more, competitive balance. The paper uses data from the second tier of professional English league football to show just how unbalanced a league planner would choose.

JEL Classification Codes: D71, L25, L41, L83

Keywords: competitive balance, sports leagues, football, soccer

* Original version completed January 2005. Draft: Comments welcome.

---

[1]Stefan Szymanski, Tanaka Business School, Imperial College London, South Kensington campus, SW7 2AZ, UK. Tel : (44) 20 7594 9107, Fax: (44) 20 7823 7685, e-mail: szy@imperial.ac.uk.

[2]Stephanie Leach, Tanaka Business School, Imperial College London, South Kensington campus, SW7 2AZ, UK. Tel : (44) 20 7594 9171, Fax: (44) 20 7823 7685, e-mail: stephanie.leach@imperial.ac.uk

1. Introduction

This paper concerns a simple proposition. It is widely held that (a) a degree of competitive balance between teams is required to make an attractive sporting contest and that (b) in the absence of competitive restraints the outcome of a league competition will become too unbalanced. This paper argues, contrary to the conventional wisdom, that the degree of competitive balance that obtains at the Nash equilibrium in a competitive sports league will be greater than that which would be chosen by a league planner seeking to maximise league attendance. In other words, competition produces *too much* rather than *too little* competitive balance.

We demonstrate this proposition both theoretically and empirically. Theoretically, the explanation is simple. In a sports league the teams compete to be successful, and successful teams attract more fans and more revenue. Typically, however, the marginal return on success differs between teams (usually, though not always, this is because the marginal revenue from success is greater for big city teams than for small city teams). League competition is a zero sum game: the supply of success is fixed. As in any competitive situation, competitors impose externalities on each other, since they fail to take into account the losses imposed on their rivals by their own success. Teams with a low marginal return to success create a larger externality than teams with a high marginal return to success. At a competitive equilibrium all teams over-invest relative to the level of investment that would be chosen by an attendance maximising league planner, but the teams with a low marginal return on success over-invest more than the teams with a higher marginal return. At equilibrium, a marginal reduction in investment by teams with a low marginal return will increase total support (or revenues) for the league as a whole.

Empirically, we estimate the relationship between success and attendance in the second tier of English football[3]. During the period we study, 1978 to 2003, estimating the relationship between wins and attendance in the first tier is problematic because of capacity constraints. In the second tier there were almost no capacity constraints, and

[3] Between 1892 and 1992 this was called the Second Division of the Football League; from 1992 until 2004, following the breakaway of the old First Division to form the FA Premier League, it was known as the First Division of the Football League; in 2004 it was renamed the Football League Championship. To avoid confusion we will refer to it simply as the second tier.

hence estimating the relationship between success and attendance is relatively straightforward. We measure success by the percentage of matches won during a season, treating a draw as a half a win. We find, unsurprisingly, that the sensitivity of attendance to wins varies substantially between teams. Using the estimated sensitivities we are able to simulate the attendance maximising distribution of wins, and we show that this distribution is far more uneven than the empirical distribution of wins.

The paper is set out as follows. In the next section we briefly discuss the conventional wisdom on the relationship between competitive balance and attendance. In section 3 we discuss the theoretical result and its relation to the literature on competitive balance and in section 4 we discuss both the empirical methodology and the results. Section 5 concludes.

2. The relationship between the distribution of results and the attractiveness of a sports league

It is widely agreed that a degree of "outcome uncertainty" is necessary to maintain interest in a sporting competition. The proposition was stated succinctly in one of the first mathematical treatments of a sports league:

"As the probability of either team winning approaches one, gate receipts fall substantially, consequently, every team has an economic motive for not becoming too superior" El-Hodiri and Quirk, 1971, p1306.

Or in the words of Walter Neale, the prayer of every team must be "Oh Lord, make us good, but not that good" (Neale, 1964, p21). It has been the repeated claim of sports leagues across the world that unrestrained competition between teams will result in a situation where the stronger teams become "too good" and the weaker teams "not good enough", and that this will lead to a loss of interest in the sport. This argument has been used to justify the adoption of competitive restraints. Most notably, Major League Baseball in the USA repeatedly justified its reserve clause, a system introduced in 1879 which effectively tied a player to his club in perpetuity, on the

grounds that it was necessary to maintain competitive balance. As early as 1889 the League issued a statement claiming that the Reserve Clause was instituted to protect the weaker teams and that the need to equalize playing strengths of teams was widely recognized.[4]

The need to maintain competitive balance was also cited in the notorious Federal Baseball case which handed organized baseball an antitrust exemption.[5] According to the submission of the owners, "If the reserve clause did not exist, the highly skillful players would be absorbed by the more wealthy clubs, and thus some clubs in the league would so far outstrip others in playing ability that the contests between the superior and inferior clubs would be uninteresting, and the public would refuse to patronize them." (Court of Appeals, 1920).

A similar restraint, called the "retain and transfer system" was adopted by the English Football League, in 1889, one year after its foundation. The rule was justified in 1906 thus by one former player: "I think it will be admitted that it is not for the good of football that the cream of players shall be in half a dozen teams. No matter how strong a club may be, it cannot play matches with itself. It must have opponents, and it is well that it should have plenty of able opponents… Without the restriction which the transfer system imposed no poor club could have held its players long".[6]

The system lasted until 1963 when it was struck down by Justice Wilberforce, largely on the grounds that, while maintaining a balance might be a legitimate goal of the League, the system in fact failed to do so: "Under the existing system the richer clubs- which are to be found in the larger centres of population- already tend to secure the better players; this is simply because, both from bigger gate-money and from the

[4] "As a check on competition, the weaker clubs demanded the privilege of reserving five players…", The statement was issued in connection with a dispute over the control of player contracts with the rival Players' League created in the same year. The statement went on to say "the necessity for such power of preserving the circuit of a league, by approximately equalizing its playing strength, is recognized by the League". Spalding (1911). The early development of what might be called the "competitive balance defense" is discussed in Eckard (2001). He points out that the competitive balance defense started to be used only in the late 1880s by which time it had become apparent that Congress would pass an antitrust law; prior to that the reserve clause was justified simply in terms of reducing player wages.

[5] While much weakened, the exemption remains in place to the present day.

[6] William Bassett, "Big Transfers and the Transfer System" in The Book of Football, p161, republished 1997, Desert Island Books.

contributions of local supporters of affluence, the clubs in the larger centres inevitably enjoy greater resources, and because the best, if not the only way to use these resources is to buy players with them. It is not established to my satisfaction that any substantial change would be brought about if the retention system were abolished".[7]

In 1995 The European Court of Justice struck down the transfer system for football players then operating within the European Union since it was ruled incompatible with Article 48 of the Treaty of Rome (which guarantees free movement of labour). The Opinion of Judge Advocate Lenz in this case stated "it should be noted that I share the opinion…that a professional league can flourish only if there is no too glaring imbalance between the clubs taking part. If the league is clearly dominated by one team the necessary tension is absent and the interest of the spectators will thus probably lapse within a foreseeable period…However, I am of the opinion that the transfer rules in their current form cannot be justified by that consideration".[8]

No one, however, advocates perfect balance. It is widely accepted that big city teams will tend to be stronger than small city teams, and that a degree of imbalance is not unhealthy. From the standpoint of economic theory the central question that this discussion raises is this: granted a competition may become less attractive if it becomes too unbalanced, is it then the case that the distribution of outcomes at a noncooperative Nash equilibrium will be less balanced than the distribution which would be selected by a central planner?

3. Theory

The most consistently voiced opinion in the sports economics literature is that the planner's equilibrium will coincide with the noncooperative equilibrium- this is the so-called Invariance Principle. This was first articulated by Quirk and El-Hodiri (1974) in relation to the sharing of gate revenues. They claimed that in a league "with franchise owners motivated solely by profit…the distribution of playing strengths is independent of the gate-sharing arrangements" and "is the same as it would be if the

[7] Quoted in Dabscheck (1985), p9.
[8] Quoted in Gardiner et al. (1998), pp365-7.

league were operated as a syndicate with central control over the allocation of players among teams" (pp36-7). This implies that noncooperative profit maximization by competing firms in a sports league generates the same distribution of resources as a central planner (or joint profit maximization). This position was restated by, inter alia, Fort and Quirk (1995) and Vrooman (1995).

Szymanski and Kesenne (2004) prove that, contrary to the claims of Quirk and El-Hodiri, increased gate revenue sharing will in fact lead to a *reduction* in competitive balance. The reason for this is the same as the rationale advanced in this paper. As gate sharing increases teams get closer to adopting the planner's solution, and the planner's solution entails less competitive balance than the noncooperative Nash equilibrium. In a noncooperative model, teams impose externalities on each other through their choices. Both teams over-invest, but relatively speaking it is the weaker drawing teams that over-invest most[9]; revenue sharing eliminates the externality, and this therefore produces a more uneven talent distribution. Full revenue sharing is equivalent to joint profit maximization, so it follows that a league planner will choose a *less* balanced distribution of results than at the noncooperative equilibrium.

This point is easily illustrated using a simple model. The structure of the model is close to that of the earlier sports literature, e.g. Atkinson et al (1988), and Fort and Quirk (1995). We make the following assumptions:

A1. Attendance generation: Each team generates attendance according to the number of wins, represented by a concave function $Q_i(w_i)$ with $Q_i' > 0$ and $Q_i'' \leq 0$; beyond some critical value it is possible that $Q_i' < 0$.

We here follow the US literature in modeling win percentage rather than league position or points, largely since this is more convenient both theoretically and empirically.[10]

[9] it is worth noting that, in practice, weaker teams tend to be the ones that find themselves in financial difficulties

[10] The league system in England awarded two points for a win and one point for a draw until 1982, giving an identical weighting as our win percentage. Since 1982 three points have been awarded for a win. The correlation between win percentage and league rank in our sample, which is the second tier between 1978 and 2003, is 0.94.

A2. Win production: Each team purchases talent (t) in a competitive market. Talent is assumed to be measured in perfectly divisible units and sold at a constant marginal cost. Increasing investment in talent increases the probability of winning and therefore expected win percentage. The win production function is strictly concave,

with $w_i(0) = 0$, $w_i(\infty) = 1$, $\frac{\partial w_i}{\partial t_i} > 0$ and $\frac{\partial^2 w_i}{\partial t_i^2} < 0$.

We assume a conventional league where every team plays every other team twice, home and away. Since each team can win between 0 and 100% of its games, and aggregate winning percentage for the league is n/2, where n is the number of teams.[11] It is standard in the US literature to assume that there is a fixed pool of talent, but in the case of second tier English football, it is more sensible to assume a perfectly elastic supply of talent, since teams can easily trade in either the domestic or the international market. This assumption affects only the price of talent in the market.

A3. Teams maximize profits, and the league planner maximizes attendance. To compare the planner's choice with the teams' choices in a competitive market we assume that ticket prices are identical, implying that profit maximization and attendance maximization are also identical.[12]

Proposition

[11] In this paper win percentage for a team refers to its performance over the entire season against all teams, rather than the sum of bilateral win percentages from which the season's win percentage must be derived. i.e. $w_i = \frac{1}{M_i}\sum_{j \neq i} m_{ij} w_{ij}$, where M_i is the total number of games played by a team in the season, m_{ij} is the number of games played between teams i and j in a season and w_{ij} is the percentage of these games won by team i. As long as every team plays the same number of games then the sum of win percentages must add to n/2. There is an interesting scheduling problem when, as in MLB, each team is not required to play every other team or to play an equal number of games against every other team. The scheduler can arrange any number of match-ups between any pair of teams except for 3, whose schedule must be determined in order to meet the constraint that each team plays the same number of games.

[12] This assumption essentially rules out any conflict between a league planner and a social planner's objective, assuming that a social planner was indifferent as to the identity of successful and unsuccessful teams. When we allow prices to vary, it is possible that a social planner interested in maximising attendance might prefer a different distribution of wins to the league planner.

(a) If the marginal revenue functions are identical, then the noncooperative Nash equilibrium for the league will be perfectly balanced. In this case the planner's equilibrium coincides with the Nash equilibrium.

(b) With asymmetric marginal revenue functions, the planner's equilibrium is *less* balanced than the noncooperative Nash equilibrium, in the sense that the difference in win percentage between any pair of teams will be larger at the planner's equilibrium. The only possible exception is when there exist two teams, one of which has a win percentage of 0 and the other a win percentage of 1 at the planner's equilibrium, in which case the planner's allocation of wins to these two teams will be identical to the Nash equilibrium. Note that if such a pair exists, it is unique.

Proof

Part (a) is obvious and well known in the literature (e.g. El-Hodiri and Quirk, 1971, proposition 3, p1312).

To prove part (b), first note that the profit function of team i is:

(1) $$\pi_i = p_i Q_i(w_i) - ct_i$$

where p is the price of a ticket and c is the marginal cost per unit of talent. We compare two cases, one where each team maximizes profit independently, and the other where a league planner maximizes joint profits. Suppose that both the competitive and planner's equilibrium involves an interior solution. Given the technology of winning, the noncooperative Nash equilibrium is characterized by the set of first order conditions for the choice of talent[13]

(2) $$\frac{Q_i'}{Q_j'} = \frac{\frac{\partial w_j}{\partial t_j}}{\frac{\partial w_i}{\partial t_i}} \quad \text{for all i and j.}$$

[13] Strictly speaking, if the supply of talent is fixed, teams should not choose talent but a budget which then determines the share of talent of each team. However, if we assume that the allocation of talent as a function of budgets is the same as the allocation of wins as a function of talent, then the budget and talent choice problems are identical.

It is easy to show that the second order conditions are satisfied given our assumptions.[14]

At the Nash equilibrium defined by (2), *if* $t_i > t_j$ *then* $Q'_i > Q'_j$. Thus at equilibrium the marginal revenue of a win for a dominant team is greater than the marginal revenue of a win for a weak team. For the league planner, however, the equilibrium condition is simply

(3) $$\frac{Q_i'}{Q_j'} = 1 \text{ for all i and j}$$

To meet the planner's objective (4) requires that the marginal revenue of team i to fall relative to the equilibrium described in (3), and the marginal revenue of team j to rise. Given concavity, this can only happen if wins increase for team i and fall for team j, implying increased dominance of team i at the planner's equilibrium.

Suppose, alternatively, that the Nash equilibrium involves a corner solution. Comparing any two teams in the league, there are three possibilities (i) one has a win percentage of 1 and the other zero, (ii) one has a win percentage of 0 and other has a win percentage between 0 and 1, and (iii) one has a win percentage of 1 and other has a win percentage between 0 and 1. Note that at most one team in a league can have a win percentage of 0 and one team a win percentage of 1. [15] In each case the planner wants to move the marginal revenues of the teams toward equality, and therefore *both* increase the wins of the stronger team *and* reduce the wins of the weaker team. In case (i) no change is possible, and so the balance between these two teams is unchanged. In cases (ii) and (iii), however, it is possible to increase win percentage of the strong team and reduce the win percentage of the weak team respectively, and hence it is possible for the imbalance between these two teams to increase. QED

[14] An additional condition, ignored here, is that marginal revenues must equal the marginal cost of talent. When supply is fixed the price of talent must be bid up to satisfy this condition. If marginal cost is not bid up to meet the equilibrium condition there will be pressure either to expand the league or for a new league to enter the market.

[15] In practice no team in the second tier of English football has ever had a perfect winning or losing record.

Given the modesty of the assumptions made the result is rather striking. One of the most important advantages associated with research in sports economics is the widespread availability of data, and so we now set out to test the theory using data on English football.

4. Empirical evidence

There is a large literature on the demand for team sports in general and for English soccer in particular, and much of it involves an attempt to test for the uncertainty of outcome hypothesis. There are a number of recent surveys of this literature, including Dobson and Goddard (2001), Szymanski (2003) and Borland and McDonald (2004).

The most widely used technique is to examine attendance at individual matches where the difference in the quality of the teams can be measured and to ask whether smaller gaps in team quality make for larger attendance. Despite the universal approval that the uncertainty of outcome hypothesis commands, the empirical support is surprisingly weak. For example, McDonald and Borland reviewed 18 studies of match uncertainty and concluded "only about three provide strong evidence of an effect on attendance…Other studies provide mixed evidence that suggests a negative effect on attendance of increasing home-team win probability only when the win probability is above two thirds. The majority of studies find that there is either no significant relationship between difference in team performance and attendance…, or more directly contradictory, that attendance is monotonically increasing in the probability of a home team win" (p486). In fact, this should not be all that surprising. Most of the fans at games support the home team, and most of these fans just want their team to win. Even excessive dominance does not seem to dent their enthusiasm (anecdotally, Arsenal did not lose a single game in the 2003/04 season with a win percentage of 84%, and every match was a sell-out).

In this paper we focus on seasonal average attendance and the competitive balance of an entire championship. Fewer studies have focused on this issue, although arguably the balance of an entire championship is more important for thinking about

competitive balance and competitive restraints imposed by leagues than the balance of an individual match. The first issue is to decide how to measure competitive balance. The within-season standard deviation of team win percentages is the most popular measure used in the literature. This is a convenient summary statistic for the degree of balance of league taken as a whole, and facilitates between season comparisons.[16] A variety of other metrics have been used- e.g. the Gini coefficient (Schmidt and Berri (2001)), The Hirschman-Herfindahl index (Depken (1999)), entropy (Horowitz (1997), the average number of games behind the winner (Knowles et al (1992)). No doubt there are many others that could be constructed.

In this study we concentrate on the second tier of English football. Since 1921 there have been four principal professional football leagues in England. Until 1992 these were all controlled by the Football League, when the top tier, now known as the FA Premier League seceded to form their own organization.[17] Unlike closed American leagues, where each league has a fixed number of teams (apart from the occasional addition of "expansion franchises"), English football uses the promotion and relegation system, whereby there are a fixed number of places in each league, but teams move up and down between leagues (or league divisions as they are traditionally known in England) based on performance.[18] Throughout most of the period covered by our data the three worst performing teams each season, based on points,[19] were relegated to the third tier, the two best performing teams were automatically promoted to the top tier, while the four next best performing teams contested play-offs for an additional promotion slot.

We chose not to look at the top tier largely because of the difficulties induced by capacity constraints. Over the last fifteen years the top tier has enjoyed rapid growth in attendance, averaging 4% per year between the nadir in 1988/89 and 2002/03. In

[16] Where the number of games played in the season varies, this can be normalized against the "idealized" standard deviation, which is expected standard deviation over M games when each team's expected winning percentage is 0.5 (see e.g. Fort and Quirk (1995)).
[17] See footnote 3 on nomenclature.
[18] There has been some recent academic interest in comparing the US closed system and the promotion and relegation system which is used in football worldwide. For an American perspective on how the system operates in England see Noll (2002), for an antitrust analysis see Ross and Szymanski (2002), and for an analysis of the economic incentives see Szymanski and Valletti (2003).
[19] Since 1982 three points have been awarded for a win and one for a draw- prior to this only two points were awarded for a win.

the mid 1990s an increasing fraction of games were sell outs and by the end of the 1990s almost every single Premier League game was a sell-out. We looked at average attendance data for top tier clubs between 1977 and 2003, and found that over the period there were 117 cases out of a sample of 549 (21%) where teams were capacity constrained (defined by annual attendance exceeding 90% of stated capacity), and all but five of these cases arose in the last ten years. We considered using a Tobit specification, but there are well known difficulties concerning normality and heteroscedasticity (see e.g. Greene (2000)). An alternative strategy would be to restrict our sample to the pre-1995 period, but we preferred to concentrate on the second tier, where capacity constraints are almost unknown.[20]

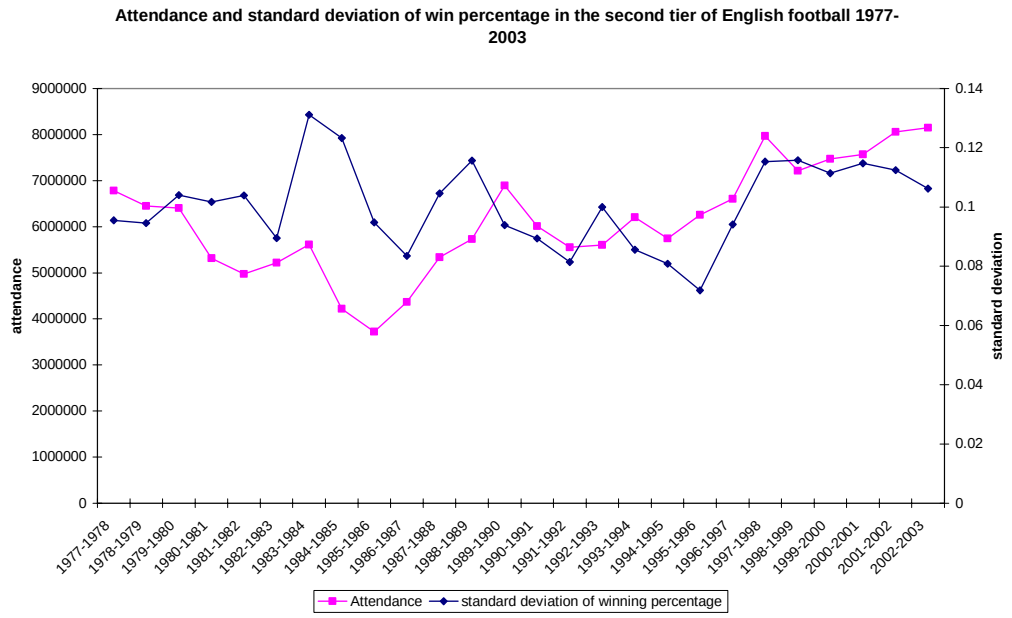

Figure 1

Figure 1 illustrates total annual attendance and the within season standard deviation of win percentages in the second tier over the sample period.[21] Attendance has varied quite substantially, from a low of around 4 million in 1985/86 to a high of over 8 million in 2002/03. The within season standard deviation of win percentage has varied between 0.072 and 0.131. There is quite a small degree of variation relative to the

[20] In fact, we did examine this restricted sample and found results similar to those presented below. The regressions are available from the authors on request.
[21] The attendance data was downloaded from http://www.european-football-statistics.co.uk/ and the league tables were downloaded from http://www.rsssf.com/.

potential variation- the theoretical minimum being 0 (when each team win 50% of its games) and the maximum 0.307 (given a league of 24 teams where each plays home and away against every other team, one team could win 46 games, another 44, another 42, and so on). Thus the actual difference between the highest and lowest standard deviation is only about 20% of the maximum possible variation. The data suggests little relationship between competitive balance as measured by the standard deviation of win percentage and attendance, which are in fact positively correlated (correlation coefficient 0.22), meaning that higher annual attendance is associated with less, not more, competitive balance.

This may help to explain why, despite the fact that there are a number of studies that examine seasonal attendance at English league clubs, none of them have reported a measure of competitive balance as an explanatory variable. Recent studies include Dobson and Goddard (1995), Simmons (1996) and Szymanski and Smith (1997), who consider mainly long term trends and the influence of factors such as prices. All include league position as an explanatory variable which is, not surprisingly, highly significant. Dobson and Goddard (1995, 2001) go further and include a club specific league position variable, and show that there is a large variation between clubs in the sensitivity of attendance to league position. The reported short term sensitivities from Dobson and Goddard (2001) are highly correlated with average league position over their sample period, which is in turn highly correlated with their estimate of a team's "base attendance", or what might be called the drawing potential of the club.[22] In other words, some teams tend to have greater drawing power than others, these teams tend to be more successful, and also have a greater short-run sensitivity of attendance to demand.

These results support our contention that less balanced distribution of results will increase attendance, since the model implies that the maximum attendance is achieved by giving the highest possible win percentage to the team with the largest sensitivity of attendance to wins, the second largest win percentage to the team with the second highest sensitivity of attendance to wins, and so on. Such a distribution of wins would

[22] The estimates are reported in Dobson and Goddard (2001), pp348-351, column 2 (average league position), column 3, "base attendance", and column 4, short run sensitivity of attendance to demand. The correlation coefficient for columns 2 and 4 is -0.48 (higher league positions have lower numbers).

produce the maximum feasible standard deviation of 0.307 for the second tier, and hence a much more uneven distribution of success than has been observed in practice. However, their linear specification implies constant returns to success, whereas it is plausible to suppose that decreasing returns will set in if a team becomes extremely successful, possibly even leading to a decline in attendance if success is too predictable.

While no studies of seasonal attendance in English football has examined the effect of competitive balance, two studies of Major League Baseball have done so. Schmidt and Berri (2001) relate attendance over the past 100 years in the National League and American League to a Gini coefficient based on wins. They find over this period that demand increases significantly with increasing balance. However, when they estimate a panel regression over a shorter, more recent, period, in order to include other economic variables of interest such as price, they find that the Gini coefficient has the opposite effect- reducing competitive balance increases attendance.[23] This result flagrantly contradicts the conventional wisdom. Schmidt and Berri restrict the coefficient on win percentage to be common across all teams, a restriction which seems unlikely to be supported by the data. Humphreys (2002), proposes a measure of competitive balance which combines both the within-season standard deviation of win percentages and the between-season variation, capturing the idea that dominance between seasons is also likely to affect attendance. He also examines aggregate attendance over a 100-year period, pooling National League and American League data, and finds that attendance is significantly affected by his measure, even when conventional measures of within competitive balance are insignificant. Both these studies suggest that competitive balance affects attendance in aggregate, but they give little indication as to how the distribution of wins between individual teams will affect attendance. Indeed, these studies imply that winning impacts each team in the same way.

Our empirical strategy builds on our observation about the club specific sensitivities estimated by Dobson and Goddard. We estimate a linear relationship between attendance and win percentage for each team that appeared in the second tier between

[23] This is no longer true when they average over a three to five year period.

1977 and 2003, but we also allow for a quadratic term, on the commonly stated assumption that excessive dominance may reduce attendance.

Thanks to the promotion and relegation system, there were 70 teams that appeared in the second tier of English football over the 26 seasons. Between 1977 and 1987 there were 22 teams, in following season the number was increased to 23, and in the following year raised again to 24, since when the number has been unchanged. We thus have 603 observations in the population, but six cases, involving teams that appeared for only a single season in the sample period, were deleted. The most frequent participant was Barnsley, which appeared a total of twenty times, and the median number of appearances for a club was 8.

Our estimating equation is:

$$(4) \qquad Attendance_{it} = a_t + b_i\, wpc_{it} + c_i\, wpc_{it}^2 + \varepsilon_{it}$$

Which we estimate using ordinary least squares. While this is a very simple model, the adjusted R^2 of 0.838 suggests that it fits the data extremely well. The linear terms are all positive and significant at the 5% level or better. When we estimated the quadratic term for each team we found that this produced twelve cases with positive coefficients, implying increasing returns to wins but these were statistically insignificant. Indeed, only ten of the squared terms were significant at the 5% level and so we grouped all the insignificant terms together and estimated a single squared term for each (while retaining a separate linear term for each). We tested the restriction of equality of the quadratic coefficients and could not reject the restriction.[24] The regression output is reported in appendix 1. The yearly fixed effects closely capture the trend in total attendance: the correlation coefficient between total attendance and the year dummies is 0.94.[25]

[24] The test statistic for the restriction is distributed $\chi^2(53)$ and the value of the test statistic is 50.914, which has p-value of 0.556.

[25] Earlier studies have expressed concern about stationarity of the data. Clearly win percentage, which must always average 0.5, has no trend, but attendance in most leagues that have been studied shows a strong trend, usually upwards. Our data is trending down between 1977 and 1986 and upwards since then, as can be seen from figure 1. An obvious solution is to difference the data, but this sacrifices a large number of degrees of freedom (we lose over 100 observations, largely because promotion and relegation means that teams seldom experience a lengthy continuous spell in the division), and the

From our quadratic estimates we are able to calculate the attendance maximising distribution of win percentage. Given $b > 0$ and $c < 0$ for all i, the league planner will maximize total attendance when the marginal attendance for each team with respect to wins is equalized; in other words

(5) $b_i + 2c_i\, wpc_i = b_j + 2c_j\, wpc_j$ for all i and j.

Thus

(6) $wpc_i = (b_j - b_i)/2c_i + (c_j/c_i)\, wpc_j$

If we sum over all wpc_i not including wpc_j, then

(7) $\sum_{(i\neq j)} wpc_i = \sum_{(i\neq j)} [(b_j - b_i)/ 2c_i] + c_j\, wpc_j \sum_{(i\neq j)} (1/c_i)$

but also

(8) $\sum_{(i\neq j)} wpc_i = n/2 - wpc_j$.

This is the adding up constraint which requires that the sum of individual team win percentages equals n/2 where n is the number of teams in the league. Hence

(9) $wpc_j = \{n/2 - \sum_{(i\neq j)} [(b_j - b_i)/2c_i]\}/[1 + c_j \sum_{(i\neq j)}(1/c_i)]$

Note that (9) does not impose the constraint that win percentage lies between 0 and 1 for each team. However, when we calculated the optimal win percentage for teams over the last decade, we found that only 28 out of 240 cases lay outside the feasible range. Clearly it makes no sense to argue that attendance could be increased if teams could achieve the impossible, so we restricted the win percentages outside the feasible

---

estimates are then poorly defined. Nonetheless, even in differenced form there remains a significant positive correlation between win percentage and attendance in aggregate.

range to be either 0 (if negative) or 1 (if positive).[26] As an example, Table 1 shows the actual and simulated results for the 2002/03 season. In this case the win percentage of Sheffield Wednesday is constrained to equal 1, the win percentage of Wimbledon to be zero and the win percentage of Gillingham to be 4.3% (since two teams cannot have a zero win percentage).[27]

The results illustrate our fundamental proposition. The actual standard deviation of win percentages in the season was 0.106 and the total of average attendances per team was 370,240.[28] The distribution of results under the optimal unconstrained allocation produces a standard deviation of win percentage of 0.34, beyond the theoretical maximum (because the unconstrained optimum includes win percentages in the unfeasible range). Once the extreme win percentages are constrained to fall within the bounds of possibility, their standard deviation is 0.277, still close to the theoretical maximum of 0.307, and far in excess of the actual figure for that season. This distribution of win percentages would have produced a total average attendance per team of 408,052 which is 10.2% larger than the actual attendance figure. Thus, as expected, a more uneven distribution of wins produces an increase in attendance.[29]

[26] Note that if two teams had an unconstrained win percentage greater than one then only one of them could have a win percentage of one, and the other could have at most a win percentage of 0.957. For the decade 1994-2003 there were nine cases where only one team had an unconstrained win percentage greater than one and one case where two teams had an unconstrained win percentage greater than one.
[27] Although the constrained model described here ensures that impossible win percentages for an individual team are eliminated, it does not ensure that win percentages sum to n/2. In 5 out of our ten simulations the constrained sum of win percentages are less than n/2, and hence the optimum attendance would be even larger, in one year all the optimal win percentages lay in the feasible range (0,1), while in four years the sum exceeded n/2, but at most by one percent of n/2, implying a negligible overstatement for the optimal attendance.
[28] Total attendance for the season was therefore 24 times this figure.
[29] It is perhaps obvious that this greater inequality of results would also produce a larger standard deviation of attendances among the teams.

Table 1: Actual and maximum attendance for the second tier, 2002/03

Club	Actual win percentage	Actual average attendance	Attendance maximising win percentage	constrained attendance maximising win percentage	maximum average attendance
Bradford City	0.413	12501	0.650	0.650	18944
Brighton & Hove Albion	0.370	6651	0.781	0.781	23305
Burnley	0.435	13977	0.597	0.597	17276
Coventry City	0.413	14813	0.733	0.733	21695
Crystal Palace	0.489	16867	0.749	0.749	22234
Derby County	0.402	25470	0.393	0.393	18524
Gillingham	0.500	8078	-0.104	0.043	2761
Grimsby Town	0.326	5700	0.143	0.143	5049
Ipswich Town	0.554	25455	0.360	0.360	19298
Leicester City	0.717	29231	0.893	0.893	27293
Millwall	0.511	8512	0.263	0.263	7927
Norwich City	0.543	20353	0.843	0.843	25490
Nottingham Forest	0.587	24437	0.495	0.495	21477
Portsmouth	0.750	18906	0.660	0.660	19280
Preston North End	0.489	13853	0.379	0.379	10942
Reading	0.587	16011	0.426	0.426	12247
Rotherham United	0.478	7522	0.167	0.167	5607
Sheffield United	0.620	18073	0.403	0.403	16401
Sheffield Wednesday	0.391	20327	1.362	1.000	37883
Stoke City	0.413	14588	0.767	0.767	22843
Walsall	0.424	6978	0.182	0.182	11136
Watford	0.467	13405	0.528	0.528	15173
Wimbledon	0.511	2787	-0.205	0.000	1983
Wolverhampton Wanderers	0.609	25745	0.535	0.535	23285
Total	12	370240	12	11.991	408052
Standard deviation	0.106	7327	0.340	0.277	8583

Table 2 summarises the results for the previous decade. It shows both the actual and maximum feasible attendance each season, assuming an attendance maximising distribution of win percentage, and associated standard deviations of win percentages. The table shows that total attendance could have been increased in every season, by amounts varying between 3.6% and 9.9%, if the standard deviation of win percentages had been much larger than that observed. In other words, a less balanced competition could have produced greater attendance.

Table 2 : Actual attendance and maximum feasible attendance, 1994-2003

Season	Sum of average attendance per club (actual)	Sum of average attendance per club (constrained optimal)	difference	standard deviation of win percentages (actual)	standard deviation of win percentages (constrained)	sum of win percentages (constrained)
1993/94	281938	296615	5.2%	0.086	0.274	11.571
1994/95	261222	270633	3.6%	0.081	0.238	11.503
1995/96	284512	296466	4.2%	0.072	0.266	11.543
1996/97	300336	314789	4.8%	0.094	0.231	12.000
1997/98	362128	397913	9.9%	0.115	0.289	12.115
1998/99	327961	354169	8.0%	0.116	0.287	12.060
1999/00	339712	368816	8.6%	0.111	0.256	12.087
2000/01	344097	376491	9.4%	0.115	0.283	12.124
2001/02	366164	402470	9.9%	0.112	0.306	11.906
2002/03	370240	408052	10.2%	0.106	0.277	11.991

One constraint that we have not addressed in our analysis is the capacity of the stadium. In the last decade there are 28 cases (just over 1 per season) where the optimal win percentage implied an attendance that exceeded the capacity of the stadium, with the median shortfall being 14%. If we assumed that each stadium could only accommodate its stated capacity, the optimal attendance would fall somewhat, but not by much, given that nearly 90% of teams would have had spare capacity under an optimal distribution of wins. It also seems reasonable to argue that if it were possible to engineer an optimal distribution of wins, it would also be feasible to increase capacity by the relatively small amounts required in most cases.

5. Conclusions

This paper makes a startling proposition: the distribution of success in a sports league produced at the noncooperative equilibrium is more balanced than the planner's (or the joint profit maximising) equilibrium, assuming that the planner wants to maximise league attendance. This contradicts the conventional wisdom that competition produces too little competitive balance and that governing bodies should take measures to increase competitive balance for the sake of preserving interest in the game.

However, it is the conventional wisdom that flies in the face of economic theory. Even if the result that a planner would choose *less*, not more, competitive balance seems counterintuitive, it is in fact just one example of a standard result in the economic literature. Consider for example, a Cournot quantity setting game in which firms have asymmetric (constant) marginal costs. At an interior equilibrium the high marginal cost firms depress industry profits not only because of the pure competition effect, but also because the industry could profitably replace any unit of output from one of these firms with a unit of output from a low marginal cost firm. A planner would shut down all the high marginal cost firms. A similar story can be found in auction theory. In a first or second price auction with private values, a low valuation bidder will impose an externality on the high valuation bidder (who wins the auction), by forcing them to increase their bid. In general, the elimination of the low valuation bidder will create a larger welfare gain (for the bidders only) than eliminating the high valuation bidder.[30]

Our results have significant implications for public policy in relation to sports leagues. Antitrust authorities have traditionally looked kindly upon restrictions which league authorities claim will improve competitive balance, even if these restrictions have, at the same time, negative implications for player wages, ticket prices, or access of consumers to football on TV. This paper, however, suggests that the competitive balance justification is perverse, given that attendances should be expected to fall if the league becomes more balanced relative to the noncooperative equilibrium.

One caveat concerning our results is that they are focused entirely on "within season" competitive balance. We make no apologies for this, since almost all pro-competitive balance arguments in the literature have been couched in these terms- i.e. that the championship in a given season becomes less interesting the more it is dominated by one or a small number of teams. However, there is an argument to say that "between season" balance matters, in the sense that long run domination by the same teams is

[30] Imagine bidder 1 values the object at V_1 and bidder 2 at V_2, with $V_1 > V_2$. In an English auction bidder 1 wins at a price equal to V_2, so bidder 1's net gain is $V_1 - V_2$ and bidder 2's is zero. Eliminate bidder 1 and bidder 2 wins the object and pays 0, giving a net gain of $V_2 – (V_1 - V_2) = 2V_2 – V_1$. Eliminate bidder 2 and bidder 1 wins at a price of 0 with a net gain of $V_1 – (V_1 - V_2) = V_2$. But clearly the elimination of bidder 2 produces the larger net gain, since $V_2 > 2V_2 – V_1 \Rightarrow V_1 > V_2$, which is true by assumption. This also assumes, of course, that if the low valuation bidder wins he cannot transfer the object of the auction to the high valuation bidder.

not desirable. In principle, our model suggests that the same teams should dominate season after season, since we estimate only a single attendance/wins sensitivity parameter for all the seasons that a team appears in our data. However, this is really a restriction imposed by the data- given the small number of observations per team it would be impractical to estimate time varying parameters. Our theoretical argument is simply that attendance is maximised in each season by awarding wins according the attendance/wins sensitivity parameters, and that these imply a much less balanced competition than we observe in practice. We suspect that attendance could be increased still further if changes in the parameters over time were accounted for. Moreover, it is worth recording that over time European leagues have been dominated by a very small number of teams (Buzzacchi et al (2003) provide striking evidence of this), and yet remain hugely popular with fans. At the very least a more sophisticated argument is required on the part of those who advocate more competitive balance.

References

Atkinson, Scott, Linda Stanley and John Tschirhart. 1988. "Revenue Sharing as an incentive in an agency problem: an example from the National Football League" *Rand Journal of Economics*, 19, 1, 27-43

Buzzachi, L., S. Szymanski and T. Valletti, 2003. "Equality of opportunity and equality of outcome: open leagues, closed leagues and Competitive Balance", *Journal of Industry, Competition and Trade*, 3, 3, 167-86

Dabscheck, B. (1985) "Sporting Labour Markets and the Courts" Sporting Traditions,

Depken, Craig. 1999. "Free-Agency and the Competitiveness of Major League Baseball" *Review of Industrial Organization*, 14, 205-217.

Jeffery Borland and Robert McDonald (2003) "Demand for Sport", *Oxford Review of Economic Policy*, 19, 4, 478-502

Dobson, Stephen and John Goddard, 1995. "The demand for professional league football in England and Wales 1925-1992", *The Statistician*, 44, 259-77.

Dobson, Stephen and John Goddard. 2001. The Economics of Football. Cambridge University Press.

Eckard, Woodrow. 2001. "The Origin of the Reserve Clause," *Journal of Sports Economics*, 2, 2, 113-130.

El-Hodiri, Mohamed and James Quirk. 1971. "An Economic Model of a Professional Sports League" *Journal of Political Economy*, 79, 1302-19

Forrest, David and Robert Simmons. 2002. "Outcome uncertainty and attendance demand in sport: the case of English soccer," *Journal of the Royal Statistical Society*, *Series D* (*The Statistician*), Vol. 51, No.2, 229-241"

Fort, Rodney and James Quirk. 1995. "Cross Subsidization, Incentives and Outcomes in Professional Team Sports Leagues" *Journal of Economic Literature*, XXXIII, 3, 1265-1299

Gardiner, Simon, Alexandra Felix, John O'Leary, Mark James, and Roger Welch. 1998. Sports Law. London: Cavendish Publishing.

Greene, William, (2000), Econometric Analysis, 4th Edition. Prentice Hall.

Horowitz, Ira. 1997. "The increasing competitive balance in Major League Baseball" *Review of Industrial Organization*, 12, 373-387.

Humphreys, Brad. 2002. "Alternative measures of competitive balance in sports leagues" *Journal of Sports Economics*, 3, 2, 133-148.

Knowles Glenn, Keith Sherony and Michael Haupert. 1992. "The demand for major league baseball: a test of the uncertainty of outcome hypothesis" *The American Economist*, 36, 72-80.

Neale, Walter. 1964. "The Peculiar Economics of Professional Sport", *Quarterly Journal of Economics*, 78, 1, 1-14.

Noll, Roger. 2002. "The economics of promotion and relegation in sports leagues: the case of English football" *Journal of Sports Economics*, 3, 2, 169-203.

Quirk, James and Mohamed El Hodiri. 1974. "The economic theory of a professional sports league" in *Government and the Sports Business,* Roger Noll, ed. Washington: Brookings Institution.

Ross, Stephen F. and Stefan Szymanski. 2002. "Open competition in league sports" *Wisconsin Law Review*, vol 2002, 3, 625-656.

Schmidt, Martin and David Berri. 2001. "Competitive Balance and Attendance: the Case of Major League Baseball", *Journal of Sports Economics*, 2, 2, 145-167.

Simmons, R. (1996) "The demand for English league football: a club level analysis", *Applied Economics*, 28, 139-55

Spalding, Albert. 1911. America's National Game. New York: American Sports Publishing Company.

Szymanski, Stefan 2003. "The Economic Design of Sporting Contests" *Journal of Economic Literature*, XLI, 1137- 1187.

Szymanski, Stefan and Stefan Késenne, 2004. "Competitive balance and gate revenue sharing in team sports" *Journal of Industrial Economics*, LII, 1, 165-177.

Szymanski, Stefan and Ron Smith. 1997. "The English Football Industry, Profit, Performance and Industrial Structure" *International Review of Applied Economics*, 11, 1, 135-153

Szymanski, Stefan and Tommaso Valletti (2003) "Promotion and relegation in sporting contest", mimeo, Tanaka Business School, Imperial College London.

Vrooman, John. 1995. "A General Theory of Professional Sports Leagues" *Southern Economic Journal*, 61, 4, 971-90

Appendix: (a) Regression Estimates, team attendance/win sensitivity

	Levels terms			Squared terms		
	Coefficient	*Std.Error*	*t-prob*	*Coefficient*	*Std.Error*	*t-prob*
Barnsley	28351	2656	0	-9114	3079	0.003
Birmingham City	37751	2775	0	-9114	3079	0.003
Blackburn Rovers	28783	2680	0	-9114	3079	0.003
Bolton Wanderers	31360	2822	0	-9114	3079	0.003
Bournemouth	29820	4020	0	-9114	3079	0.003
Bradford City	32034	2914	0	-9114	3079	0.003
Brighton & Hove Albion	34425	2769	0	-9114	3079	0.003
Bristol City	33361	3173	0	-9114	3079	0.003
Bristol Rovers	22967	3077	0	-9114	3079	0.003
Burnley	31070	2923	0	-9114	3079	0.003
Bury	15711	4647	0.001	-9114	3079	0.003
Cambridge United	20964	3038	0	-9114	3079	0.003
Cardiff City	25054	3032	0	-9114	3079	0.003
Carlisle United	23558	3559	0	-9114	3079	0.003
Charlton Athletic	26666	2687	0	-9114	3079	0.003
Chelsea	52412	8769	0	-30521	13070	0.02
Coventry City	33562	4406	0	-9114	3079	0.003
Crewe Alexandra	15765	3500	0	-9114	3079	0.003
Crystal Palace	33854	2658	0	-9114	3079	0.003
Derby County	64093	8375	0	-55921	14550	0
Fulham	26524	2864	0	-9114	3079	0.003
Gillingham	18298	3720	0	-9114	3079	0.003
Grimsby Town	22795	2693	0	-9114	3079	0.003
Huddersfield Town	31225	2866	0	-9114	3079	0.003
Hull City	27085	3051	0	-9114	3079	0.003
Ipswich Town	75927	12180	0	-77344	20230	0
Leeds United	44919	2916	0	-9114	3079	0.003
Leicester City	36476	2788	0	-9114	3079	0.003
Leyton Orient	21386	3214	0	-9114	3079	0.003
Luton Town	26234	2871	0	-9114	3079	0.003
Manchester City	84161	9697	0	-59727	15190	0
Middlesbrough	37356	2863	0	-9114	3079	0.003
Millwall	24990	2719	0	-9114	3079	0.003
Newcastle United	47882	2865	0	-9114	3079	0.003
Norwich City	35564	2859	0	-9114	3079	0.003
Nottingham Forest	58603	10860	0	-38813	17700	0.029
Notts County	39272	9085	0	-36051	16970	0.034
Oldham Athletic	24555	2650	0	-9114	3079	0.003
Oxford United	24180	2935	0	-9114	3079	0.003
Peterborough United	28283	4728	0	-9114	3079	0.003
Plymouth Argyle	31654	3231	0	-9114	3079	0.003
Port Vale	24815	3016	0	-9114	3079	0.003
Portsmouth	32224	2722	0	-9114	3079	0.003
Preston North End	27100	3072	0	-9114	3079	0.003
Queens Park Rangers	30268	2922	0	-9114	3079	0.003
Reading	27964	3095	0	-9114	3079	0.003
Rotherham United	23237	3491	0	-9114	3079	0.003
Sheffield United	51271	8462	0	-38508	15360	0.013
Sheffield Wednesday	45014	2977	0	-9114	3079	0.003
Shrewsbury Town	22373	2880	0	-9114	3079	0.003
Southend United	21904	3249	0	-9114	3079	0.003
Stockport County	18934	3504	0	-9114	3079	0.003
Stoke City	34180	2718	0	-9114	3079	0.003
Sunderland	47010	2769	0	-9114	3079	0.003
Swansea City	32545	3814	0	-9114	3079	0.003
Swindon Town	28654	2878	0	-9114	3079	0.003
Tranmere Rovers	22842	2875	0	-9114	3079	0.003
Walsall	80356	27050	0.003	-165217	68550	0.016
Watford	29811	2737	0	-9114	3079	0.003
West Bromwich Albion	50580	7970	0	-36102	15070	0.017
West Ham United	43131	3059	0	-9114	3079	0.003
Wimbledon	16459	3139	0	-9114	3079	0.003
Wolverhampton Wanderers	59370	8767	0	-36580	15160	0.016
Wrexham	25371	3606	0	-9114	3079	0.003

(b) Regression Estimates, year dummies

	Coefficient	*Std.Error*	*t-prob*
YR79	52	739	0.943
YR80	-52	753	0.945
YR81	-1988	762	0.009
YR82	-2806	753	0
YR83	-3299	759	0
YR84	-2884	750	0
YR85	-4658	746	0
YR86	-5122	748	0
YR87	-4808	758	0
YR88	-4106	755	0
YR89	-3698	751	0
YR90	-1934	747	0.01
YR91	-2644	745	0
YR92	-2566	748	0.001
YR93	-2980	757	0
YR94	-2010	746	0.007
YR95	-2279	742	0.002
YR96	-1937	745	0.01
YR97	-1212	744	0.104
YR98	1163	757	0.125
YR99	436	745	0.559
YR00	758	745	0.309
YR01	1609	738	0.03
YR02	2079	760	0.006
YR03	1983	749	0.008

(c) Diagnostics

R^2	0.865
Adjusted-R^2	0.838
Heteroscedasticity Test	F(164,332) = 1.0251 (0.4210)
Normality test [31]	78.710 (0.00)

[31] We performed tests for heteroskedasticity and normality. Residual outliers appear for teams such as Ipswich Town and Newcastle United.