Abstract: This paper considers the Coase theorem as applied to sports. Professional team sports are frequently cited as a natural context for examining whether the distribution of assets (player talent) that can be freely traded in the market coincides with the efficient or joint profit maximising distribution. This paper shows that this is neither true in theory for a plausible trading mechanism, nor true in practice in the second tier of English professional soccer.
1. Introduction

The Coase Theorem is both one of the simplest and most profound ideas in economics. Coase’s insight was first expressed in print as a theorem by George Stigler, following the publication of the famous article “The Problem of Social Cost” by Nobel Laureate Ronald Coase (1960). Stigler stated it thus: “with zero transactions costs, private and social costs will be equal”. The significance of this statement is that “if private cost is equal to social cost, it follows that producers will only engage in an activity if the value of the product of the factors employed is greater than the value which they would yield in their best alternative use” (Coase (1988), p158). In other words, bargaining in an unrestricted market will produce full economic efficiency (assuming zero transactions costs), obviating the need to invoke government intervention in the form of Pigouvian taxes and subsidies to correct externalities. The implications for social policy are profound- simply by establishing property rights all externalities will be internalised and private transactions will be publicly optimal. Most notably, the lesson of the Coase Theorem for environmental economics is that we need only establish property rights over the quantities of greenhouse gases in the atmosphere and toxins in the oceans, and pollution will controlled at socially optimal, sustainable levels. Many economists would echo the words of Avinash Dixit and Mancur Olson (2000) “In his article ‘The Problem of Social Cost’, Ronald Coase introduced a very powerful idea of great importance. Coase’s article has been arguably the single largest influence on thinking about economic policy for the last three decades. It is one of the most – if not the most – widely cited economics article in recent times”.

Big theorems are notoriously difficult to test. Darwinism remains resolutely untestable, no one is holding out much hope of testing the theories of Freud or Marx and even in physics developments such as string theory remain testable only in principle. One problem with big theories is that they require big experiments-experimental frameworks that are capable of capturing a substantial degree of the complexity that a big theory addresses. In the world of economics big theorems such as the fundamental theorems of welfare economics or the law of demand are generally approached through specific examples. In this paper the Coase Theorem is approached through the medium of a sports league. While Coase’s article dates from
1960, a colleague at Chicago University published a discussion of the market for baseball players in 1956 which almost completely anticipates the more famous paper (Rottenberg (1956)). As in any team sport, the players are the principal asset, and teams historically have traded these assets, frequently for cash. In baseball a rule enforced by the owners, known as the Reserve Clause, prohibited players from moving teams without the permission of their current employer, effectively endowing the employer a monopsony right over the income stream of the player. As this restraint come under pressure from the players and their union, the owners sought to defend their rule by arguing that if players were free to move they would quickly migrate to the wealthiest teams, disturbing the essential element of “competitive balance” allegedly fostered by the Reserve Clause. Rottenberg argued, in the manner of Coase, that ownership rules would make no difference to the distribution of talent in a league. If owners controlled the movement of players, trade between club would cause each player to move to the location where his (marginal revenue) product is greatest. If players were free to move, bidding by the clubs to hire players would produce the same distribution (the only difference being that any economic rents would now accrue to the player, not the owner).

This paper re-examines the application of the Coase Theorem to the market for players in a sports league. It is shows that plausible trading mechanisms will not achieve Coasian efficiency. Some empirical evidence from English professional soccer illustrates the problem.

2. The Coase Theorem and its discontents

The Coase Theorem has subject to significant scrutiny in the economics literature and has been widely challenged (see e.g. Ellickson (1991) and Samuelson (1995)). Three examples of academic critiques are discussed here:

(i) Practicality (e.g. Canterbery and Marvasti (1992)). Even if it is true that costless bargaining with full property rights produces efficiency, many economists have argued that this is of little practical value, since most market failures which the Coase Theorem addresses refer to situations
where property rights are very difficult to define precisely or in a way that that is legally enforceable (e.g. rights over the ocean fisheries- even if territorial waters are assigned, the fish often fail to respect the boundaries so that enforcing rights over fish that temporarily stray into another’s jurisdiction is likely to be difficult). Additionally, the relevance of the costless bargaining paradigm is questionable since most of the difficult and important problems arise where bargaining costs are very high (e.g pollution rights).

(ii) Tautology (e.g. Usher (1998)). He argues that in a zero transaction cost world efficiency must be guaranteed among maximising agents, regardless of whether property rights exist at all- since otherwise there will exist unrealised gains from trade. Hence while it is strictly true that any allocation of property will produce efficiency in such a world, the existence of property rights is not necessary. The Coase Theorem, stripped of the necessity of property, merely becomes the statement that in a world where agents are willing and able to bargain until all potential gains from trade are realised the outcome will be economically efficient, which, as stated, appears tautological, since economic efficiency is defined as the realisation of all potential gains from trade.

(iii) Falsity (e.g. Aivazian and Callen (2003)). These authors relate the Coase Theorem to Edgeworth’s notion of the core. The core is defined as the set of efficient equilibrium bargains among parties. The perfectly competitive equilibrium of neoclassical economics belongs the set of resource allocations that are in the core, but others may exist as well. Clearly the outcome of Coasian bargaining must be in the core in as well. Different initial allocations of rights might product different allocations of resources, but all outcomes should be in the core (e.g. the allocation of resources will be different if the polluter has the absolute right to pollute compared to a situation where citizens have an absolute right to protection from the effects of pollution, but the Coase Theorem says that the amount of pollution should be fixed at the efficient level in both cases). However, if there is an allocation of property rights exists for which the core is empty- i.e. there is no equilibrium bargain that is efficient- then the Coase
Theorem fails. Aivazian and Callen provide a simple example of just such a case.

While theoretical objections abound, it is perhaps more important to understand whether the implications of the Coase Theorem are really relevant for economic policy. In other words, we need to understand whether, in a world where property rights are well defined and bargaining is not too costly, the outcomes of bargaining are plausibly close to efficiency. The team sports literature has been widely cited as an example of a situation where the Coase Theorem is put to a practical test.

3. The Coase Theorem in the sports literature

One common characteristic of team sports as they developed on both sides of the Atlantic has been the desire of the owners of teams belonging to professional leagues to control the market for players, in particular to establish monopsony rights. Thus the Reserve Clause of baseball (see e.g. Quirk and Fort (1992) for an explanation) functioned in much the same way as the Retain and Transfer System of English soccer (see e.g. Sloane (1969))². This inevitably led to challenges in the courts by the players claiming the right to move freely between employers. Simon Rottenberg's celebrated (1956) article examined this issue and presented the team owner's rationale:

"the defense most commonly heard is that the reserve rule is necessary to assure an equal distribution of playing talent among opposing teams; that a more or less equal distribution of talent is necessary if there is to be uncertainty of outcome; and that uncertainty of outcome is necessary if the consumer is to be willing to pay admission to the game. This defense is founded on the premise that there are rich baseball clubs and poor ones and that, if the players' market were free, the rich clubs would outbid the poor for talent, taking all competent players for themselves and leaving only the incompetent for other teams." (p. 246)

² In fact, the two systems were so similar that it is hard to believe that the Football League did not copy the National League. However, no evidence to this effect has ever been produced.
Rottenberg argued that (a) the Reserve clause did nothing to prevent the migration of talent to the big city teams and so would not affect the distribution of talent and that (b) by establishing monopsony power over a player throughout his career the team owners were able to hold down wages and raise profitability. Point (a) has since been identified as an example of the Coase Theorem at work: the initial distribution of ownership rights should have no impact on the efficient (here profit maximizing) distribution of resources. El-Hodiri and Quirk (1971) and Quirk and El-Hodiri (1974) took this analysis one stage further in a formal dynamic model showing that, if teams have differing revenue generating potential, (i) profit maximizing behavior will not lead to an equal distribution of resources (playing talent) and (ii) revenue redistribution on the basis of gate sharing will have no impact on the distribution of playing talent. Points (a) and (ii) are both examples of the well-known invariance principle.

There have been two significant changes in talent allocation rules in North American sports over recent years. Firstly, in 1976 major league baseball players won the right of free agency after completing six years service, and this practice rapidly spread to the other sports. Secondly, the draft rules of the NFL, which allocated the right to hire new talent entering the league on the basis of the reverse order of finish of the previous season’s competition were adopted by the other sports (see Paul Staudohar (1996) for more details on both of these innovations). These changes can be studied to identify the impact of changes in talent allocation rules on competitive balance.

(i) Free Agency

The advent of free agency in MLB in 1976 for six year veterans is a clear natural experiment\(^3\). The owners claimed that as a result of this limited free agency the best veterans would migrate to the big city teams and competitive balance would be undermined. A number of studies have attempted to use this rule change to test the invariance hypothesis, and the findings from these studies are reported in Table 3.

\(^3\) In this case the change was exogenous- i.e. not itself motivated by a desire to affect competitive balance (see Bruce Meyer (1995) for a discussion of natural experiments).
Table 1: The impact of Free Agency on Competitive Balance in MLB

<table>
<thead>
<tr>
<th>Study</th>
<th>Measure of Competitive Balance</th>
<th>Impact on Competitive Balance in NL</th>
<th>Impact on Competitive Balance in AL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daly and Moore (1981)</td>
<td>Movement of free agents to large market teams</td>
<td>(-)</td>
<td>(-)</td>
</tr>
<tr>
<td>Scully (1989)</td>
<td>Standard deviation of win percent and Gini coefficient of pennant wins</td>
<td>(+)</td>
<td>(0)</td>
</tr>
<tr>
<td>Balfour and Porter (1991)</td>
<td>Standard deviation of win percent, persistence of win percent</td>
<td>(+)</td>
<td>(+)</td>
</tr>
<tr>
<td>Fort and Quirk (1995)</td>
<td>Standard deviation of win percent and Gini coefficient of pennant wins</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>Vrooman (1995)</td>
<td>Standard deviation of win percent relative to idealized standard deviation</td>
<td>(+)</td>
<td>(+)</td>
</tr>
<tr>
<td>Vrooman (1996)</td>
<td>Persistence of win percent</td>
<td>(+)</td>
<td>(+)</td>
</tr>
<tr>
<td>Butler (1995)</td>
<td>Standard deviation of win percent and serial correlation of win percent</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>Horowitz (1997)</td>
<td>Entropy</td>
<td>(-)</td>
<td>(0)</td>
</tr>
<tr>
<td>Depken (1999)</td>
<td>Hirschman-Herfindahl index of wins relative to ideal</td>
<td>(0)</td>
<td>(-)</td>
</tr>
<tr>
<td>Eckard (2001)</td>
<td>Analysis of variance of win percent</td>
<td>(+)</td>
<td>(+)</td>
</tr>
</tbody>
</table>

Most of the studies simply look at the standard deviation of win percentages before and after 1976 (Scully (1989), Balfour and Porter (1991), Quirk and Fort (1995), Vrooman (1995), Michael Butler (1995)), while other measures include persistence in win percent (Balfour and Porter (1991), Vrooman (1996)), entropy (Horowitz (1997)), the Hirschman-Herfindahl index (Depken (1999)) and analysis of variance (Eckard (2001)). Most of these studies find either no change (seven cases) or an improvement in competitive balance (nine cases), contrary to the claim of the owners that free agency would reduce competitive balance (four cases only). However, this meta-data is hardly a ringing endorsement for the invariance principle, since “no effect” is reported in only seven out of twenty cases. Of course, it can be argued that many other factors have altered competitive balance (e.g. the increasing dispersion of local...
TV revenues), but in that case the data, without controlling for these factors, can hardly be said to represent a test at all.

Some other studies have approached the invariance principle as a direct test of the Coase Theorem and tried to establish whether the distribution of talent in the league has been affected by the introduction of free agency. George Daly (1992) observes that under the Reserve Clause top line players were seldom traded, a situation that has been affected by free agency where the top stars have a choice after six years leading to increased mobility. Timothy Hylan, Maureen Lage and Michael Treglia (1996) in a study of pitcher movements finds that these players have become less mobile since free agency, a surprising result and one that they claim does not support the Coase Theorem. However, Donald Cymrot, James Dunley and William Even (2001) examine player mobility in 1980, controlling for possible selection bias and find that, for that season at least, there was no evidence that restricted players (with less than six years service) enjoyed more or less mobility than unrestricted free agents after controlling for player characteristics.

Daniel Marburger (2002) considers a different implication of the invariance principle. If trade is possible between two independent leagues then it should be more profitable to hire a player from the same league than the rival league. Intra-league trade raises the winning probability of the buying team by more than an inter-league trade, since in the former case not only does the buyer have a larger share of talent, but the seller now has a weaker team. Under the Reserve clause this effect will be built into the seller’s price, but under free agency it will not, since the free agent is indifferent to the adverse effect on the team he is leaving. Thus with free agency the relative price of intraleague trades should fall and their share of total trades increase. Marburger found a statistically significant increase in the share of intraleague trades, from 60% to 73%, in MLB 1964 and 1992. This finding seems consistent with the invariance principle.

(ii) The rookie draft

The stated intention of the rookie draft system is to provide weaker teams with opportunities to acquire talented players by awarding them first pick. Of course, an additional consequence of this system is the creation of monopsony power. The draft
system was instituted by the NFL in 1936 as a way of strengthening weak performing teams to maintain competitive balance, and has since been adopted by all the other major leagues (Fort and Quirk (1995) and Staudohar (1996) provide details).

Daly and Moore (1981) first analyzed whether the draft achieved its stated intention by examining competitive balance before and after the introduction of the MLB draft in 1965. They found a significant improvement in the balance of the National League and a smaller improvement in the balance of the American League. The Japanese Professional Baseball League adopted a draft system at exactly the same time as MLB, and a study by La Croix and Kawaura (1999) also found that competitive balance improved over time (measured by the Gini coefficient for pennants) in both the Central and Pacific Leagues. As they point out, these results are “virtually identical” to Fort and Quirk’s (1995) results for MLB. Kevin Grier and Robert Tollison (1994) examined the impact of the rookie draft in the NFL by running an autoregressive specification for win percentage together with the average draft order over the previous three to five seasons, and found that a low draft order significantly raises performance. These results seem to provide consistent evidence against the invariance principle and in support of the owners’ stated position.

4. Trading mechanisms and the allocation of talent

One difficulty with much of the preceding analysis is that the relationship between the trading mechanisms in sports leagues and the efficient distribution of talent is poorly defined. In this section we consider a formal model of choice in a sports league: the approach followed is to derive the distribution that teams will select when they maximise profits and to compare this with a plausible candidate for an efficient allocation of talent. The model shows that when talent in freely traded in the market (so that a property right over its allocation exists), then the Nash equilibrium distribution of talent is inefficient.

4 although the within season measure (standard deviation of win percent) was significant only for the Pacific League.
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.\(^5\)

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,

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

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.\(^6\)

It is standard in the US literature to assume that there is a fixed pool of talent, while in the competing soccer leagues of the world it is more plausible to imagine that any league faces an elastic supply, since teams can easily trade in either the domestic or

---

\(^5\) 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.

\(^6\) 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.
the international market. In this version of the model the 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.\(^7\)

Under these assumptions we examine the Nash equilibrium in talent budgets.

**Proposition**

(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:

\(^7\) 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.
\[ \pi_i = p_i Q_i(w_i) - ct_i \]

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 \(^8\)

\[ \frac{Q'_i}{Q'_j} = \frac{\partial w_j}{\partial t_i} \]

for all \( i \) and \( j \).

It is easy to show that the second order conditions are satisfied given our assumptions.\(^9\)

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

\[ \frac{Q'_i}{Q'_j} = 1 \]

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.

\(^8\) 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.

\(^9\) 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.
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. 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

That a Nash equilibrium can be socially inefficient is well known. The problem here is what is sometimes called the “competitive externality”. In a contest teams impose externalities on their rivals when they try to win, since they reduce their rivals’ probability of winning, and hence there is overinvestment relative to the cooperative solution. The result here seems counterintuitive. At the Nash equilibrium the weak teams win too often and the strong teams do not win often enough—there is too much competitive balance (most sports analysts typically imagine there is too little competitive balance). In an asymmetric contest, relatively speaking it is the weaker drawing teams that over-invest most, because the opportunity cost of their success is a loss of fans or income at teams that have a greater potential to attract fans or income.\textsuperscript{12}

\textsuperscript{10} In practice no team in the second tier of English football has ever had a perfect winning or losing record.

\textsuperscript{11} Probably the best known example is the case of a Cournot industry firms where have asymmetric marginal costs, where at the Nash equilibrium the solution involves oversupply by high marginal cost firms. One solution is for the low marginal cost firms to finance a reduction in output by the high marginal cost firms—either through a takeover or by paying them to shut down. In practice this option may be restricted, either by antitrust law or because of time consistency issues— if cartel members will pay to shut down an inefficient plant then it may be profitable to build one.

\textsuperscript{12} it is worth noting that, in practice, weaker teams tend to be the ones that find themselves in financial difficulties
This also explains why the invariance principle, as applied to gate revenue sharing, does not hold in the context of a Nash contest. The argument used in the invariance principle is that at equilibrium the marginal revenue of winning must be equalized across all teams (see e.g. Vrooman 1995). However, Szymanski and Kesenne (2004) show that noncooperative behaviour only produces a joint profit maximizing distribution of talent when there is full revenue sharing (all match income is shared 50:50 with the opposing team). When teams do not share revenues, the marginal benefit of talent investment is equalized across teams, but the competitive externality is also present. Thus they are able to show that gate sharing tends to increase the gap in performance between teams (i.e. reduce competitive balance) as teams get closer to the joint profit maximizing distribution.

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.

5. 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. 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. 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. Throughout most of the

---

13 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)).
14 See footnote 3 on nomenclature.
15 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
period covered by our data the three worst performing teams each season, based on points, 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 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. 

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).

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.

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.
Figure 1 illustrates total annual attendance and the within season standard deviation of win percentages in the second tier over the sample period. 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 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

The attendance data was downloaded from http://www.european-football-statistics.co.uk/ and the league tables were downloaded from http://www.rsssf.com/.
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. 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 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

---

19 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).
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.\textsuperscript{20} 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 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.

\textsuperscript{20} This is no longer true when they average over a three to five year period.
Our estimating equation is:

\[(4) \quad \text{Attendance}_{it} = a_t + b_i \text{wpc}_{it} + c_i \text{wpc}_{it}^2 + \epsilon_{it}\]

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.\(^{21}\) 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.\(^{22}\)

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

\[(5) \quad b_i + 2c_i \text{wpc}_i = b_j + 2c_j \text{wpc}_j \quad \text{for all } i \text{ and } j.\]

Thus

\[(6) \quad \text{wpc}_i = \frac{(b_j - b_i) + (c_j/c_i) \text{wpc}_j}{2c_i} \]

---

\(^{21}\) 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.

\(^{22}\) 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 estimates are then poorly defined. Nonetheless, even in differenced form there remains a significant positive correlation between win percentage and attendance in aggregate.
If we sum over all \( wpc_i \) not including \( wpc_j \), then

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

but also

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

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

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

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 range to be either 0 (if negative) or 1 (if positive).\(^{23}\) 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).\(^{24}\)

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

---

\(^{23}\) 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.

\(^{24}\) 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.
was 370,240. 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.

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

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

| Total                  | 12                    | 370240                     | 12                                  | 11.991                                        | 408052                    |
| Standard deviation      | 0.106                 | 7327                       | 0.340                               | 0.277                                         | 8583                      |

25 Total attendance for the season was therefore 24 times this figure.
26 It is perhaps obvious that this greater inequality of results would also produce a larger standard deviation of attendances among the teams.
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

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

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.
6. Discussion and Conclusions

This paper has shown that in the distribution of talent in a league where teams have asymmetric revenue generating potential will be more balanced at the Nash equilibrium than the social optimum. Note that this is not only a case where competition adversely affects the interests of the producers but also the welfare of the customers, since greater fan utility could be created by a more unbalanced contest. This is a surprising result, since it is usually argued that competitive equilibria are likely to produce too little competitive balance, requiring redistribution of resources from strong teams to weak teams to produce a more balanced contest.

This result depends on the particular trading mechanism considered- teams choose talent budgets independently, and hence do not account for the impact that their budgetary choices have on their rivals. There are at least four ways that this inefficiency could be overcome by the league members:

(i) Integration of teams in a single business enterprise
(ii) a collusive arrangement to involving side-payments between the teams
(iii) tacit collusion supported by the threat of punishments in a dynamic context
(iv) manipulation of the investment rules by a planner

All of these pose problems. Integration is problematic because the attractiveness of league competition typically relies on the notion that contestants are independent. For the same reason most leagues prohibit ownership of more than one team – if fans suspect the contest is not genuine, then they lose interest. A similar concern tends to rule out side payments- although player trading does suggest ways in which the externality might be overcome. If talent is scarce, then large drawing teams could pay more for players that they acquire from small drawing teams than those players are worth, encouraging those teams to sell more talent than they otherwise might. This suggests a bidding model for talent, rather than the quantity setting model explored in this paper. Dakhlia and Pecorino (2004) explore a model where strong teams may preempt the market by buying all the talent available in the market, essentially outbidding rivals. In their model this gives rise to the opposite problem- strong teams hire too much talent and the competition is too unbalanced (see the appendix for an
example that integrates the two approaches). Tacit collusion is an unlikely way to achieve collectively desirable sporting outcomes, since contestants tend to discount the future heavily in the quest for glory. For example, many athletes seem willing to take banned steroids that result in significantly shorter lifespans in order to achieve short term success. Businesses may adopt a longer term perspective, but given the significant degree of randomness in the outcome of sporting contests it would be hard to maintain effective collusive strategies. Most leagues have a either a commissioner, empowered to represent the collective interest of the owners, or a governing body that imposes rules on contestants. This is one way in which manipulation is likely to take place. In practice, however, such organizations typically give greater weight to the views of the many small teams rather than the few strong teams. Hence they are unlikely to promote measures that redistribute from the weak to the strong. In the end, it may simply be transactions costs, the old enemy of Coase, that prevents the Coasian bargain from being struck.

While mechanisms for achieving Coasian efficiency can be imagined, the empirical evidence derived from the distribution of success in the second tier of the English league demonstrates that leagues may in practice be unable to correct the externality. In a related paper we have shown that exactly the same effect can be shown empirically for Major League Baseball. The Coase theorem is an important idea in economics, and professional sport is commonly cited as an example of where the existence of property rights in player services should ensure an efficient distribution of talent in market. The evidence of this paper suggests sports markets lend scant support for the theorem.
References


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


Leach, Stephanie and Stefan Szymanski, 2005, “Tilting the Playing Field: Why a sports league planner would choose less, not more, competitive balance: The case of English football” mimeo, Tanaka Business School, Imperial College London.


Szymanski, Stefan, 2004a, “Professional team sports are only a game: The Walrasian fixed supply conjecture model, Contest-Nash equilibrium and the invariance principle” *Journal of Sports Economics*, 5, 2, 111-126

Szymanski, Stefan, 2004b, “Tilting the Playing Field: Why a sports league planner would choose less, not more, competitive balance: The case of baseball”, mimeo, Tanaka Business School, Imperial College London.


Appendix: the pre-emption result (Dakhlia and Pecorino).

Szymanski (2004a) considered a simple model of competition in a league where demand depends on (a) the success of each team and (b) the degree of competitive balance. The model assumes is a supply of talent to the market, which may be fixed or elastic, and team success, measured by the percentage of games won depends on the share of total talent hired. Teams are assumed to maximise profit. In a league of two teams this boils down a contest success function

\[ w_1 = \frac{t_1}{t_1 + t_2}, \quad w_2 = 1 - w_1 \]

Where \( w \) is win percentage and \( t \) is talent hired, which is assumed perfectly divisible, and a profit function, where is here given the simple form

\[ \pi_1 = (\sigma - w_1) w_1 - ct_1, \quad \pi_2 = (1 - w_2) w_2 - ct_2, \quad \sigma > 1 \]

where \( c \) is the (constant) marginal cost of talent and \( \sigma \) indicates that team 1 is capable of generating a larger revenue than team 2 from any given level of success. Note that the demand for competitive balance ensures that revenues are ultimately decreasing in success, but that given the adding up constraint in (1), there is no guarantee of an interior solution.

How do teams choose talent? We can imagine this as a quantity setting (Cournot-type) or price-setting (Bertrand-type) game. As a quantity game, teams allocate a budget to hiring talent and the talent they can hire is proportional to their share of total budgets. The first order conditions for talent choice are therefore

\[ \frac{\partial \pi_1}{\partial t_1} = (\sigma - 2w_1) w_2 - cT = 0, \quad \frac{\partial \pi_2}{\partial t_2} = (1 - 2w_2) w_1 - cT = 0 \]

where \( T = t_1 + t_2 \), so that at the Nash equilibrium

\[ w_1^* = \frac{\sigma}{1 + \sigma} \]

\(^{27}\) A more general model that is used to analyse the impact of gate revenue sharing is to be found in Szymanski and Kesenne (2004).
Team 1 dominates \((w_1 > \frac{1}{2})\) in equilibrium because it has the larger drawing power \((\sigma > 1)\). However at equilibrium the marginal revenue of a win for team 1 exceeds the marginal revenue of team 2

\[
\frac{\partial R_1}{\partial w_1} = \sigma - 2w_1^* = \frac{\sigma(\sigma - 1)}{(\sigma + 1)} > \frac{\sigma - 1}{(\sigma + 1)} = 1 - 2w_2^* = \frac{\partial R_2}{\partial w_2}
\]

This implies a distribution of talent in the league that is not jointly efficient. To see this note that joint profits are

\[
\pi_1 + \pi_2 = (1 + \sigma)w_1 - 2w_1^2 - cT,
\]

which is maximized when

\[
w_1^M = (1 + \sigma)/4 > w_1^*
\]

Hence the quantity bidding mechanism entails “too much” competitive balance at the Nash equilibrium. Intuitively, this result is a consequence of asymmetry. Competition always involves an externality- each team’s actions under competition fails to account for the negative effect that actions have on rivals’ profits. The externality imposed by the team with the lower win percentage in equilibrium is bigger precisely because the big team loses more than the small team when its rival wins more.

Dakhlia and Pecorino (2004) consider a rent-seeking model where teams not only bid for a quantity of talent but also submit a bid for the wage rate per unit of talent. If each team offers the same wage rate then the Nash equilibrium distribution of talent will be the same as above. However, if one team bids higher than the other it can attract all the talent, generating a corner solution. In their model, where teams only have a demand for winning and there is no value in competitive balance, they show that the dominant team will be willing to pre-empt all of the talent by offering a bid at with its rival’s demand for talent is zero, as long as the quantity of talent is not too great. However, if the supply of talent is large enough, pre-emption is not profitable,
given that the team would have to hire all of the talent in order to pre-empt the market.  

The incentive to pre-empt can be identified by comparing the profit level at an interior equilibrium for a given marginal cost of talent with the profit made by one team raising price by $\varepsilon$ above marginal cost, hiring all the talent and winning all the time. If this deviation can be shown to be profitable then a form of Bertrand competition will ensue.

To derive the condition for a profitable deviation first consider the demand for talent at the interior equilibrium. First note that $w_1 / w_2 = t_1 / t_2 = \sigma$. Writing (3) in terms of $t_1$ and $t_2$, substituting for $t_2$ we obtain

$$t_1^* = \frac{\sigma^2 (\sigma - 1)}{(1 + \sigma)^2 c}, \quad t_2^* = \frac{\sigma (\sigma - 1)}{(1 + \sigma)^2 c}$$

This implies that team 1 makes profit equal to

$$\pi_1^* = \frac{\sigma^4 + \sigma^2}{(1 + \sigma)^3}$$

It can now be shown that team 1 would want to pre-empt by offering a wage rate “$c + \varepsilon$” if the total supply to the market $T^S = t_1^* + t_2^*$. For $\varepsilon$ small enough, the profits from pre-emption are

$$\sigma - 1 - c T^S = \frac{\sigma^3 - 1}{(1 + \sigma)^2}$$

Pre-emption can therefore be profitable if

---

28 Efficiency in their model of pure rent seeking (à la Tullock) is slightly peculiar, in that the most efficient result is for team 1 to win all the time since it values the payoff more. Moreover, even if team 1 pre-empts all the talent, it only need to employ $\varepsilon$ of it to win with certainty, since team 2 hire zero in equilibrium. The point here is that the simple rent seeking game requires more structure in order for an interior solution to be efficient. If, for example, there is a demand for competitive balance, then an interior solution can be efficient.
Thus pre-emption can be a profitable strategy if the dominant team is sufficiently large. As Dakhlia and Pecorino show, pre-emption additionally requires that the marginal profit of team 2 is negative when it hires zero units of talent, requiring that $bT^S > 1$, where $b$ is the pre-emptive bid of team 1 and $T^S$ is the total supply of talent. Additionally, however, if $T^S > t_1^* + t_2^*$ then it becomes less and less likely that pre-emption is profitable.

Thus, when there is bidding for talent, there is a possibility that we shift from an inefficient interior solution to a pre-emptive corner solution, which is also inefficient. In both cases it is assumed that any market clearing mechanism must involve identical treatment for each unit of talent. If talent were sold for different prices to different teams then there would be no equilibrium among the players, since low paid players would with identical skills would be willing to move to high paying teams.