Revenue Divergence and Competitive Balance in a Divisional Sports League

Stephen Dobson*

and

John Goddard^

Abstract

The North American model of resource allocation in professional sports leagues is adapted for English (association) football. The theoretical relationship between revenue and competitive balance is shown to be robust with respect to changes in teams' objectives and labour market conditions. Empirical revenue functions are reported for 1926-1999. These indicate a shift in the composition of demand favouring big-city teams and an increase in the sensitivity of revenue to performance. An analysis of match results in the FA Cup competition suggests an increase in competitive imbalance between teams at different levels of the league's divisional hierarchy, as the theory suggests.

JEL Code: L83

* Department of Economics, University of Otago, P.O.Box 56, Dunedin, New Zealand. Phone: +64 3 479 5296. Fax: +64 3 479 8174. Email: sdobson@business.otago.ac.nz

^ Department of Economics, University of Wales Swansea, Singleton Park, Swansea, SA2 8PP, UK. Tel: +44 1792 295168. Fax: +44 1792 295872. Email: j.a.goddard@swan.ac.uk

Revenue Divergence and Competitive Balance in a Divisional Sports League

I Introduction

The implications for competitive balance of various institutional and structural characteristics of professional team sports markets are a key feature of the economics of sports literature. A major concern has been the impact on competitive balance of free agency in the players' labour market. Some twenty years before the introduction of free agency in North American (NA) baseball, Rottenberg (1956) argued that provided there was unrestricted trade between teams in players, free agency should not affect competitive balance. Profit maximising teams employ playing talent to the point that marginal revenue products (the revenue generated from the last unit of talent employed) are equal across teams, so there are no unexploited opportunities for mutually profitable trade in players. According to the invariance proposition, the location of this point does not depend on players' contractual arrangements, and is therefore unaffected by the introduction of free agency. Contrary to the received wisdom of many sports policymakers and commentators, a 'free-for-all' in the labour market should neither cause the best players to gravitate to the economically most powerful and highest-paying teams, nor cause competitive balance to decline.

Rottenberg's theoretical insights were formalised by El Hodiri and Quirk (1971) (EHQ), and have since been tested empirically by, among others, Scully (1989), Fort and Quirk (1995), Vrooman (1995) and Eckard (1998). Recently this literature has been reviewed by Eckard (2001), who finds that "(t)aken together, the results are consistent with the invariance proposition, namely, balance did not change" (p430). Eckard's own interpretation, supported both theoretically and empirically, is that competitive balance in baseball may actually have improved somewhat during the free agency era.

In the EHQ model, relative market sizes are the final arbiter of the allocation of playing talent between teams, and therefore of competitive balance. The model is therefore demand-driven. While the empirical NA literature tends to concentrate on the link between institutional characteristics of the supply side and competitive balance, the present article focuses primarily on demand-side determinants of competitive balance. Specifically, it seeks to demonstrate and draw connections between two empirical propositions concerning professional (association) football (or soccer) in England and Wales (EW): first, between the 1920s and 1990s there was increasing divergence in the base levels of spectator demand enjoyed by teams with distinct identifying characteristics; and second, over the same period there was increasing competitive imbalance within the league as a whole. According to the EHQ model, the second proposition should follow directly from the first. Since long-term change in the composition of demand has been a major factor affecting football's historical development in EW, the latter should provide a vehicle for empirical scrutiny of the theory, from a perspective that differs from that of most of the NA literature.

The article is structured as follows. Section II discusses the adaptation of the EHQ model into a form suitable for the analysis of English football. Section III presents estimated revenue functions for football teams in EW, and draws inferences about the phenomenon of revenue divergence. Section IV discusses the measurement of competitive balance between teams operating at different levels of a hierarchical, divisional league structure. A statistical analysis of win probabilities in cup ties (in which teams from different divisions meet head-on) is used to identify trends in competitive balance. Section V summarises and concludes.

II Revenue and competitive balance: theoretical considerations

Section II discusses the adaptation of the EHQ model to English football. For present purposes, the stylised two-team version of EHQ's n-team model employed by Fort and Quirk (1995) and Vrooman (1995) will suffice. The revenue functions for team i (i=1,2) are:

$$R_i = kP_i^{\delta}W_i^{\beta} \qquad [1]$$

In [1] R_i = revenue; P_i = home-town population ($P_1>P_2$ is assumed); W_i = win ratio = $T_i/(T_1+T_2)$ where T_i is the quantity of playing talent employed by team i; δ and β are the elasticities of revenue with respect to home-town population and win ratio respectively; and k is constant.

It is straightforward to establish a relationship between the parameters δ and β, and the level of competitive balance, measured by the ratio of win ratios W_1/W_2, also equivalent to the ratio of talent employed T_1/T_2. It is assumed that teams are profit maximisers. The total quantity of talent is assumed fixed (and scaled so that $T_1+T_2=1$ and $W_i=T_i$) and c, the wage cost per unit of talent, is assumed to be endogenous. The teams compete to attract talent, forcing c to a level at which the profit maximising conditions for both teams are satisfied. Individually, however, each team treats c as given, and selects T_i to maximise its own profit, $\pi_i = R_i - cT_i$. The first-order conditions are $\beta P_i^{\delta}T_i^{\beta-1} = c$ for i=1,2. This leads to the equilibrium condition:

$$T_1/T_2 = W_1/W_2 = (P_1/P_2)^{\delta/(1-\beta)} \qquad [2]$$

Given the elasticities δ and β, competitive balance is determined by the ratio of the home-town populations P_1 and P_2.

The application of this theoretical literature to professional team sports in Europe is still in its infancy. Recently, however, models originally developed in NA have been adapted to investigate a number of specific policy issues. Hoehn and Szymanski (1999) consider the influence of the involvement of the most successful teams in European-level tournaments for competitive balance in their domestic leagues. The incentive to recruit sufficient talent to succeed at European level causes domestic competitive balance to deteriorate, perhaps to the point that the small-market teams become economically non-viable. Késenne (2000) explores the implications for competitive balance of the hypothetical introduction of a NA-style salary cap in European football. More generally, the institutional, historical and economic characteristics of professional football in EW differ from those of the leading NA professional sports in respect of (at least) five major characteristics (see also Fort, 2000), which need to be addressed in order to develop a suitable adaptation of the EHQ model.

(i) Membership of the top professional sports leagues in NA is determined by the award of franchises. In contrast English football's Premier League (PL) and Football League (FL) comprise a hierarchical divisional structure. At the end of each season, several teams are promoted and relegated between divisions, so divisional membership is determined by competitive prowess.[1] From casual inspection of PL and FL revenues data, it is clear that revenues are primarily dependent on a team's position within the league rather than within its own division. The inclusion of the win ratio, W_i, in [1] is therefore inappropriate, but it is appropriate to replace W_i with $T_i/(T_1+T_2)$ because league position is monotonic in T_i. For empirical purposes if n, the number of league member teams, is large, $T_i/(T_1+T_2)$ can be

approximated on a numerical scale from 1 to 1/n by $L_i = (n+1-pos_i)/n$, where pos_i is team i's league position (1=top of PL; n=bottom of FL).[2]

(ii) In each of the major NA sports, the introduction of free agency was a once-and-for-all event. In English football, progress toward free agency has been incremental. Before 1978 a player could only move from one team to another with his current employer's consent. From 1978 a player whose contract had expired became entitled to move regardless of his employer's wishes. His new employer, however, would be liable to pay compensation in the form of a transfer fee. Following the 1995 European Court of Justice (ECJ) ruling in the Bosman case, out-of-contract players over the age of 24 became entitled to move without liability for compensation. Full free agency rights were thereby established. Before 1978 teams motivated primarily by winning not profit (see below) probably had little incentive, and could not be forced, to allow their best players to move. Since 1978 and especially since 1995, mobility (at the players' behest) has increased considerably, driven in part by differences in teams' financial positions and ability-to-pay. A closer relationship between demand shifts and competitive balance might therefore be expected post-1978 than pre-1978.

(iii) Because few (if any) other countries operate professional leagues in the same sports on a comparable scale, product and labour markets in NA (especially at the top level) are effectively closed. A regulation restricting to three the number of non-British players PL and FL teams could field at any one time remained in force until 1995, when the ECJ adjudged the application of this provision to EU nationals to be counter to European employment law.[3] The EHQ model's closed labour market assumption is therefore a reasonable approximation to the true situation in football in EW for most of the period under scrutiny. Post-1995, however, many foreign players have been employed, especially by PL teams. In the model an

open labour market dictates the removal of the restriction that T_1+T_2 is fixed. One team can increase its stock of talent without depleting that of the other team, by recruiting additional talent from abroad.

(iv) The EHQ model assumes that each team's spectator demand is proportional to its home-town population: a reasonable assumption in NA where each team's catchment area is predominantly local (due to large distances between towns), and each town has only one team. In football in EW neither of these conditions prevails. Geographical segmentation between spectator catchment areas is less than in NA because distances are small. Wherever they reside, regular travel to home and away fixtures is a practical proposition for many sports fans. Home-town population is therefore a highly imperfect indicator of each team's popularity. In [1] P_i^{δ} is therefore replaced by A_i, a set of fixed-effects capturing all relevant demographic, socio-economic and historical influences on team i's base level of spectator demand.

(v) Finally, due partly to the restrictions the franchise system imposes on supply, and partly to the restraining influence on wage expenditure of salary caps in football and basketball, most NA teams are highly profitable (Quirk and Fort, 1999). Profit maximising assumptions have been adopted in much of the academic literature, with little dissent. In EW in contrast, the intense competitive pressure generated by the league's divisional structure has certainly contributed to the chronic loss-making propensities of most PL and FL teams. Following Sloane (1971), some researchers have argued that win, revenue or utility maximisation subject to a financial constraint may be a more suitable objective function than profit maximisation (Késenne, 1996, 1999). On the other hand, the extensive commercialisation of football during the 1990s (of which the stock market flotation of several teams is one

manifestation) suggests that profit maximisation may yet be appropriate in some cases. The theoretical analysis that follows remains agnostic on this issue, and derives equilibria under both profit and win-maximising assumptions.

Taking account of (i) to (v) above, it is possible to re-work the two-team model. The adapted revenue functions for team i (i=1,2) are:

$$R_i = A_i\{T_i/(T_1+T_2)\}^{\beta} \qquad [3]$$

$A_1>A_2$ is assumed. Under profit maximising assumptions with a closed labour market ($T_1+T_2=1$) and c determined endogenously, the model is essentially the same as in the NA case. The competitive balance equilibrium is:

$$T_1/T_2 = (A_1/A_2)^{1/(1-\beta)} \qquad [4]$$

Under profit maximising assumptions with an open labour market (no restrictions on T_1 and T_2) it is reasonable to assume c is exogenous. For team 1 the first-order condition is:

$$A_1\beta\{T_1/(T_1+T_2)\}^{\beta-1}\{T_2/(T_1+T_2)^2\}=c \qquad [5]$$

Using [5] and the equivalent expression for team 2 to solve for T_1 and T_2 yields the competitive balance equilibrium:

$$T_1/T_2 = (A_1/A_2)^{1/(2-\beta)} \qquad [6]$$

Comparing [6] with [4], competition is more balanced if the labour market is open. Because team 1's employment of additional talent does not simultaneously deplete team 2's talent, it is not in team 1's interest to hire talent and dominate competition to the same extent as in the closed model.

Under win maximisation with a zero profit constraint, the teams employ talent to the break-even point, $\pi_i = R_i - cT_i = 0$. In the closed model the break-even conditions yield:

$$A_1T_1^{\beta} = cT_1 \; ; \qquad A_2(1 - T_1)^{\beta} = c(1 - T_1) \qquad [7]$$

Eliminating c and solving for T_1 and T_2 yields:

$$T_1 = (A_1/A_2)^{1/(1-\beta)}/\{1 + (A_1/A_2)^{1/(1-\beta)}\} \; ; \qquad T_2 = 1/\{1 + (A_1/A_2)^{1/(1-\beta)}\} \qquad [8]$$

[8] in turn gives rise to a competitive balance equilibrium identical to [4].

In the open model the break-even condition for team 1 is:

$$A_1\{T_1/(T_1+T_2)\}^{\beta} - cT_1 = 0 \qquad [9]$$

Using [9] and the equivalent expression for team 2 to solve for T_1 and T_2 yields:

$$T_1 = (A_1/c)\{(A_1/A_2)^{1/(1-\beta)}/[1+(A_1/A_2)^{1/(1-\beta)}]\}^{\beta} \; ; \; T_2 = (A_2/c)\{1/[1+(A_1/A_2)^{1/(1-\beta)}]\}^{\beta} \qquad [10]$$

In common with [8], [10] gives rise to a competitive balance equilibrium identical to [4]. [2] or [4] therefore turns out to be a highly robust result, valid under profit maximisation with a closed labour market, or under win maximisation with a closed or open labour market. Even under profit maximisation with an open labour market, the functional form of [6] is the same as [4], and the direction of the effect of changes in A_i and β is the same. Competition should become more unbalanced: (i) if the difference between A_1 and A_2 increases; or (ii) if β increases. (i) follows from the property that each team's competitive strength is proportional to its relative market size. The intuition for (ii) is that if the sensitivity of revenue to winning increases, it is efficient for the big-market team to employ more talent than before and the small-market team less, because the former's revenue gain exceeds the latter's loss.

III Empirical revenue functions for football teams in England and Wales

Section III reports empirical counterparts of the theoretical revenue functions developed in section II. The data set comprises annual gate revenues for all PL and FL football teams between the 1926-7 and 1998-9 seasons inclusive, obtained from the FL's archives and Football Trust (various issues).[4]

To describe the main changes in the distribution of revenue between PL and FL member teams with broadly similar demographic, geographic and historical characteristics, Dobson and Goddard (1998, 2001) have classified teams into five groups. G1 (Group 1) contains four major London teams that tend to attract significant levels of support from all parts of London and beyond (Arsenal, Chelsea, Tottenham and West Ham), and ten major teams from five other cities with populations larger than 500,000: Birmingham, Liverpool, Leeds, Manchester and Sheffield. G2 contains teams from northern and midlands cities with populations in the range 250,000-500,000. G3 contains teams from southern towns other than

London, and the smaller London teams not included in G1, whose support tends to be more localised. The two remaining groups contain teams from smaller northern and midlands towns, sub-divided into pre-1922 (G4) and post-1922 (G5) league entrants.[5]

Between the 1920s and 1990s there were marked contrasts in fortunes between groups, in terms of both performance and revenue indicators. Summary data are shown in panels (a) and (b) of Table 1. These show, for example, that while the performance of the big-city teams in G1 remained almost constant, G1's revenue share increased by over ten percentage points between the 1920s and 1990s. Until the end of the 1980s G3 enjoyed a sustained improvement in performance as football's geographical balance of power shifted towards the south. G3's gain in revenue share, however, was less pronounced. The improvement in G3's performance appears not to have been driven by a shift in the composition of demand; rather, it seems likely to reflect a catching-up process, as southern teams (many of which joined the FL later than their counterparts from similar-sized northern towns) gradually worked their way up the league's divisional hierarchy.

Economists, sociologists and social historians have all attempted to explain the shift in the composition of demand favouring the big-city teams at the expense of their small-town counterparts (Dunning et al., 1988; Walvin, 1994; Russell, 1997). The geographical bond between teams and their spectators certainly appears to have been stronger in the first half of the 20th century than subsequently. Prior to the late-1950s financial and logistical constraints militated against regular long-distance travel to sports fixtures for most spectators. Thereafter rising affluence, combined with improvements in public and private transport enabled big-city teams, in particular, to begin to draw support at a national rather than purely local or regional level, and to drain support from small-town teams. Demographic change also tended

to weaken links between communities and their local teams. Suburbanisation implied population shifts away from urban districts where most stadia were located. Meanwhile the expanding reach of televised football focused public attention on a handful of star players of the most glamorous teams, also favouring the latter at the expense of their less widely exposed small-town counterparts.

The effects of such developments are incorporated into the empirical revenue functions reported below by allowing for variation over time and between groups in the revenue function parameters. Adding time-subscripts and a disturbance term to [3]; using a league position variable as a proxy for each team's share of talent (see section II); normalising by expressing team revenues as a proportion of total revenue; and applying a log transformation, yields the following specification:

$$\ln(r_{i,t}) = \alpha_{i,t} + \beta_t \ln(L_{i,t}) + u_{i,t} \qquad [11]$$

In [11] $r_{i,t} = R_{i,t} / \sum_i R_{i,t}$; $R_{i,t}$ and $L_{i,t}$ are equivalent to R_i (revenue) and L_i (league position) as defined in section II; $u_{i,t}$ is a random disturbance term; and $\alpha_{i,t}$ and β_t are time-varying parameters of the revenue function. For [11] to be estimable, some restrictions on $\alpha_{i,t}$ and β_t are required. Both are assumed to be polynomial in t.[6]

The revenue function estimation results are reported in Table 1. Panel (c) reports the group mean values of $\exp(\hat{\alpha}_{i,t})$ (the empirical counterparts of the fixed effects A_i in [3] to [10]) for selected seasons. These provide an indication of shifts between groups in the composition of aggregate spectator demand for the league as a whole, after controlling for team performance. For G1, as expected there is a pronounced increase in the mean value of $\exp(\hat{\alpha}_{i,t})$, especially

between the 1950s and 1970s. Meanwhile, the downward trend for G3 reflects the failure of this group's revenue share to match its long-term improvement in performance, as discussed above.

The cross-sectional standard deviation (across all teams) of $\hat{\alpha}_{i,t}$ in [11], denoted s($\hat{\alpha}_{i,t}$), provides a convenient summary measure of divergence between clubs in the composition of aggregate spectator demand. Panel (d) of Table 1 shows that s($\hat{\alpha}_{i,t}$) has increased over time. There was relatively little variation before the end of the 1950s, but the trend was upward in the 1960s. Following a levelling out in the 1970s and 1980s, a further sharp increase occurred in the 1990s. Panel (e) of Table 1 shows a (broadly) parallel increase in $\hat{\beta}_t$. This elasticity is estimated to have risen from below 0.25 to above 0.50 between the 1920s and 1990s. Figures 1 and 2 plot the movement in the series s($\hat{\alpha}_{i,t}$) and $\hat{\beta}_t$ over the entire estimation period.

IV Measuring competitive balance using FA Cup match results data

In the empirical NA team sports literature, it is standard practice to measure competitive balance within a league using the cross-sectional variance or standard deviation of the win ratios of member teams (Bennett and Fizel, 1995; Fort and Quirk, 1995; Eckard, 1998, 2001). This procedure is unsuitable in the case of English football, however, because of its divisional competitive structure (see footnote 1). Intra-divisional variation in win ratios may say something about competitive balance across the league as a whole: greater imbalance might be discernible within divisions to some extent. Yet such measures are unlikely to provide a powerful measure. Most of the action is inter- rather than intra-divisional, yet inter-division variation is not considered.

However, all PL and FL member teams take part in a competition that does allow direct comparisons between the playing strengths of teams from different divisions. The FA Cup (Football Association Challenge Cup) is a sudden-death knock out tournament involving both league and non-league teams. A major attraction is the cup's propensity to produce shock results, such as the elimination of a PL team by an opponent from the lower reaches of the FL or from non-league.[7]

Recently, Szymanski (2001) has used FA Cup match attendance data in an effort to identify trends in competitive balance in football in EW since the 1970s. Analysing cup attendances in matches where a corresponding league fixture (between the same teams in the same season) took place, Szymanski finds that cup attendances declined relative to league attendances. In the demand for sports literature, one version of the 'uncertainty of outcome' hypothesis is that attendances depend on the level of competitive balance between all teams in the competition concerned. Accordingly, Szymanski infers that cup attendances declined due to increasing inter-divisional competitive imbalance between league teams. League attendances are less affected, because rising competitive imbalance is mainly an inter- rather than an intra-divisional phenomenon. The present article draws similar conclusions about the trend in competitive balance. It does so, however, by measuring the latter directly, rather than indirectly via a hypothesised (and largely unsubstantiated) relationship between competitive balance and attendance.

In section IV inferences about changes in competitive balance are drawn directly from a statistical investigation of trends in win probabilities in cup matches conditional on league position. Koning (2000) and Dobson and Goddard (2001) use ordered probit regression to model league match results in football. Adapting this approach, the model describing the

result of the cup match between home team i and away team j played in season t, denoted $y_{i,j,t}$, is:

Home win $\Rightarrow$ $y_{i,j,t} = 1$ if $\mu_2 < y^*_{i,j,t} + \varepsilon_{i,j,t}$

Draw $\Rightarrow$ $y_{i,j,t} = 0.5$ if $\mu_1 < y^*_{i,j,t} + \varepsilon_{i,j,t} \leq \mu_2$

Away win $\Rightarrow$ $y_{i,j,t} = 0$ if $y^*_{i,j,t} + \varepsilon_{i,j,t} \leq \mu_1$

where $y^*_{i,j,t} = \gamma_{0,t} + \gamma_{1,t} L_{i,t} + \gamma_{2,t} L_{j,t}$ [12]

In [12], $\varepsilon_{i,j,t}$ is a NIID disturbance term. $\gamma_{1,t}<0$ and $\gamma_{2,t}>0$ are expected. $\gamma_{k,t}$ (k=0,1,2) are time-varying parameters of the ordered probit model. As before, a polynomial functional form for $\gamma_{k,t}$ is assumed.[8]

The data set comprises FA Cup matches played between the 1921-2 and 2001-2 seasons inclusive (except 1945-6 when the FA Cup was staged but the league was suspended) that were either first matches or (first) replays involving two league member teams, between Round 1 and the quarter-finals (currently Round 6) inclusive. Second and subsequent replays, semi-finals and finals are excluded because these matches are usually played at neutral venues. Match results are recorded after 90 or 120 minutes' play; results of penalty shoot-outs are not recorded.[9]

Since the rules concerning match duration differ between first matches and replays, the parameters of [12] also differ, and separate estimations are required. The estimated home win, draw and away win probabilities are $h^{(z)}(L_{i,t}, L_{j,t}) = 1-\Phi(\hat{\mu}_2 - \hat{y}^*_{i,j,t})$, $d^{(z)}(L_{i,t}, L_{j,t}) =$

$\Phi(\hat{\mu}_2 - \hat{y}^*_{i,j,t}) - \Phi(\hat{\mu}_1 - \hat{y}^*_{i,j,t})$ and $a^{(z)}(L_{i,t}, L_{j,t}) = \Phi(\hat{\mu}_1 - \hat{y}^*_{i,j,t})$ respectively, where Φ is the standard normal distribution function; z=1 denotes the first match of a cup tie; and z=2 denotes a replay.

Let $E(y_{i,j,t})$ represent the expected outcome for team i by the end of the (first) replay, calculated as a probability-weighted average of scores of 1 if team i wins the cup tie; 0.5 if the tie is still level; and 0 if team i loses:

$$E(y_{i,j,t}) = h^{(1)}(L_{i,t}, L_{j,t}) + d^{(1)}(L_{i,t}, L_{j,t})\{a^{(2)}(L_{j,t}, L_{i,t}) + 0.5d^{(2)}(L_{j,t}, L_{i,t})\}$$

If team i initially plays away to team j, the expected outcome for team i is $1 - E(y_{j,i,t})$. Team i's expected outcome, conditional on league positions but unconditional on which team initially plays at home, is $w_t(L_{i,t}, L_{j,t}) = 0.5\{1 + E(y_{i,j,t}) - E(y_{j,i,t})\}$. Variations over time in $w_t(\overline{L}_i, \overline{L}_j)$ for constant values of $\overline{L}_i$ and $\overline{L}_j$ provide an indication of trends in competitive balance. An alternative summary measure is $s(w_t)$, the cross-sectional standard deviation (across i) of $w_t(L_{i,t}, 0.50)$: the expected outcomes (as defined above) of all teams against the league's median team.

The estimation results are reported in Table 2. Panel (b) reports the estimates of $\gamma_{k,t}$ for the first matches of cup ties for selected seasons. Panel (c) reports the equivalent estimates for replays. In both cases, the decrease over time in the numerical value of $\hat{\gamma}_{0,t}$ controls for a decline in the importance of home advantage, apparent in the summary data reported in panel (a). Meanwhile $\hat{\gamma}_{1,t}$ and $\hat{\gamma}_{2,t}$ both tend to increase in absolute value.[10] Cup results have therefore become increasingly correlated with league positions. By the end of the estimation

period, any given difference in league positions counted for more than it had in earlier times. This in turn suggests an increase in competitive imbalance between teams at different levels of the league hierarchy.

Panels (d) and (e) of Table 2 report further summary measures of the trend in competitive balance. Panel (d) reports the trends in $w_t(0.75,0.25)$ and $w_t(0.90,0.10)$: the probabilities of success (after a maximum of 210 minutes) for teams positioned at the 10th and 25th percentile of the league hierarchy, against teams positioned at the 75th and 90th percentiles respectively, conditional on league positions but unconditional on home advantage. These data show that the probability of a 'shock' result declined considerably between the 1920s and the 1990s.

For selected seasons, panel (e) of Table 2 reports $s(w_t)$ as defined above. The full series is plotted in Figure 3. According to the theoretical models described in section II, the trend in $s(w_t)$ shown in Figure 3 should be linked to trends in $s(\hat{\alpha}_{i,t})$ and $\hat{\beta}_t$ shown in Figures 1 and 2. Clearly the long-term trend in all three series is upward, so to this extent the empirical findings are consistent with the theory. There are some inconsistencies, however, in the timing of the main upward shifts in $s(w_t)$ on the one hand, and $s(\hat{\alpha}_{i,t})$ and $\hat{\beta}_t$ on the other, which require some further interpretation.

The increase in $s(w_t)$ in the 1920s and 1930s does not seem to be directly attributable to parallel changes in the parameters of [11] (though there were increases in both $s(\hat{\alpha}_{i,t})$ and $\hat{\beta}_t$ in the late-1920s). The growth of a more professional ethos in football during the inter-war period provides a possible alternative, non-demand side interpretation of the tendency for competitive imbalance to increase during this period. The modern-day football manager's job

specification, including responsibility for all aspects of team affairs and the acquisition and disposal of players, first began to evolve between the wars.[11] As the style of management of the leading teams (in particular) moved onto a more professional footing, it is unsurprising to find the emergence of a larger competitive gulf between these teams and the rest.

The second significant increase in competitive balance, which began in the late-1970s, was preceded by a sharp rise in $s(\hat{\alpha}_{i,t})$ in the 1960s, and a more gradual increase in $\hat{\beta}_t$ that had been underway since the early-1950s. The peculiarities of pre-1978 contractual arrangements (see section II) may explain why these demand shifts did not feed through immediately into rising competitive imbalance. Before 1978 a lack of player mobility may have helped keep the lid on the damaging consequences for competitive balance of the shifting composition of spectator demand.[12] After 1978, when out-of-contract players secured the right to initiate a transfer, these restraints on player mobility were eased, with inevitable consequences for competitive balance. Further sharp increases in $s(\hat{\alpha}_{i,t})$ and $\hat{\beta}_t$ during the 1990s have added more fuel to the fire of rising competitive imbalance, and seem likely to explain the recent acceleration in the upward trend in $s(w_t)$.

V Conclusion

The theoretical model of resource allocation in professional sports leagues first developed by El Hodiri and Quirk (1971) suggests that competitive balance is demand-driven. Ultimately, the allocation of playing talent between teams is dependent on each team's relative market size. The equilibrium allocation of playing talent, at which all teams are simultaneously maximising either profit or performance subject to a profit constraint, can therefore be expressed as a function of the parameters of a set of team-specific revenue functions. This article has considered the adaptation of theoretical models developed in North America to

professional football in England and Wales. The functional form of the theoretical relationship between revenue and competitive balance has been shown to be robust with respect to variations in the teams' objective functions, and variations between closed and open players' labour market conditions.

This article has investigated demand-side determinants of competitive balance in professional football in England and Wales. Empirical football team revenue functions have been reported, estimated using annual data from the 1920s to the 1990s. A direct measure of competitive balance between teams at all levels of the league's divisional hierarchy, based on a statistical analysis of match results cup competition, has been developed.

The theoretical model predicts that competition will become more unbalanced, either if there is divergence between teams' base levels of spectator demand, or if the elasticity of team revenue with respect to league position increases. Empirically, both of these conditions appear to have been met. Teams from the larger cities and towns experienced an increase in their base levels of demand relative to their small-town counterparts, and there was divergence in revenue shares. The pace of revenue divergence was particularly fast during the 1960s and 1990s. The estimated elasticity of revenue with respect to league position has increased steadily, from below 0.25 in the 1920s to above 0.50 by the end of the 1990s.

The statistical analysis of FA Cup match results indicates that there has indeed been an increase in competitive imbalance between teams operating at different levels of the league's hierarchy, as the theoretical model predicts. The incidence of shock results in cup ties has declined over time, indicating an increase in the gradient between league position and competitive strength. Competitive imbalance appears to have increased in the late-1920s and

1930s. It has been suggested that this may be due to the emergence of a more professional ethos in football during this period, and not by demand-side shifts. A further significant increase in competitive imbalance has been underway since the late-1970s. This was preceded by changes in the composition of demand whose initial impact on competitive balance may have been blunted by a lack of mobility in the players' transfer market. Contractual changes implemented in 1978 and 1995 have since eased these restrictions on mobility. Consequently the most talented players have gravitated increasingly towards the highest-paying teams, and inter-divisional competitive imbalance has risen accordingly.

Footnotes

1. League competition in English professional football is currently organised into four divisions: the Premier League (PL) comprising 20 teams; and Football League (FL) comprising three divisions (FLD1-FLD3) of 24 teams each. Three or four teams per season are promoted and relegated between adjacent FL divisions, and between FLD1 and PL. A four-division structure, with total membership close to the current complement of 92 teams, has operated since the 1921-2 season.
2. L_i is discrete, while t_i is continuous. When n is large, however, L_i and t_i are approximately equivalent.
3. In practice, the number of foreign players employed in England before 1995 was small; the limit of three was only rarely reached, and only by a handful of teams.
4. While total revenues (from all sources) would be preferable in many respects to gate revenues, the former are only obtainable from company accounts, and most teams have not filed accounts consistently. For most of the period under investigation, gate revenues were by far the most important component of total revenues, though in recent seasons the proportion has fallen, in the PL and in all divisions of the FL, due to growth in income from television and other sources.
5. The classification criteria and the composition of the groups are detailed in Dobson and Goddard (2001). Populations are from the 1961 census, the closest to the mid-point of the observation period. To measure revenue divergence, it is important that the group classifications are based entirely on 'exogenous' characteristics of clubs, and not on their performance or revenues over the period under scrutiny. Perhaps the only contentious issue in this respect is the allocation of London teams between G1 and G3. By the early-1920s the four G1 London teams appear to have established a clear lead in popularity over the rest. Arsenal, Chelsea and Tottenham benefited from

early entry to the FL. West Ham joined after Fulham and Leyton Orient, but benefited from their unique East End location. By comparison the latter two have struggled, perhaps due to their geographical proximity to Chelsea and Tottenham respectively.

6. The polynomial functional forms are:

$$\alpha_{i,t} = a_0 + \sum_{m=1}^{M_1} a_m^{(g)} t^m + h_i \; ; \qquad \beta_t = b_0 + \sum_{m=1}^{M_1} b_m t^m$$

h_i are home team fixed effects, allowing for variation between teams in base levels of spectator demand. The polynomial trend component in the base levels of demand, $a_m^{(g)}$, is the same for all teams within each of the five groups (g=1...5). β_t, the elasticity of revenue share with respect to league position, is also subject to a polynomial trend, but does not vary between teams or between groups. F-tests for the joint significance of the six additional coefficients introduced by increasing M_1 in steps of one ($a_m^{(g)}$ for g=1...5 and b_m) indicate M_1=9 provides an adequate representation of the trend in $\alpha_{i,t}$ and β_t. t varies from 1 (1925-6 season) to 67 (1998-9 season). There is no break in t for the seven-year period when football was suspended during the Second World War. Teams with fewer than ten time series observations are excluded from the estimation.

7. The Football Association Challenge Cup (FA Cup) is the principal knock-out cup competition, open to all PL and FL teams, and non-league teams. The latter take part in a preliminary qualifying tournament. FLD2 and FLD3 teams enter in Round 1. Survivors are joined by PL and FLD1 teams in Round 3. Further rounds reduce the 64 Round 3 contestants to two finalists in Round 8. From the First Round onwards the draw determining who plays whom and which team initially has home advantage, is random; there are no seedings.

8. The polynomial functional form is: $\gamma_{k,t} = g_{k,0} + \sum_{m=1}^{M_2} g_{k,m} t^m$. In this expression $g_{0,0}$ is redundant and is set to zero. Ordered probit regression can be used to obtain estimates

of the parameters μ_1, μ_2, $g_{k,m}$, and therefore $\gamma_{k,t}$. Chi-square tests for the joint significance of the three additional coefficients introduced into [13] by increasing M_2 in steps of one ($g_{k,m}$ for k=0,1,2), based on the omitted variables version of Weiss's (1997) Lagrange Multiplier (LM) test, indicate that M_2=3 provides an adequate representation of the trend in $\gamma_{k,t}$.

9. If the first match produces a winner after 90 minutes' play, this result settles the tie. If the first match is level after 90 minutes, a replay is staged at the home of the team initially drawn away. If the replay produces a winner after 90 minutes, this settles the tie. If the replay is level after 90 minutes, 30 minutes' extra time are played. Until 1993-4, if the replay was still level after 120 minutes, further replays were staged until a winner emerged. Since 1993-4, penalty shoot-outs have settled ties level after 120 minutes of the (first) replay (210 minutes in total). The introduction of penalty shoot-outs in the 1993-4 season should not affect the parameters, since this change only affects what happens *after* 210 minutes' play.

10. There is some ambiguity in the results for $\hat{\gamma}_{1,t}$ in the replays estimation, which is based on fewer observations than the estimation for first matches. In this case there is a form of selection bias against teams from the lower reaches of the FL, relatively few of which survive to contest a replay having been drawn away in the first match.

11. During spells with Huddersfield and Arsenal, Herbert Chapman, the most successful manager of the inter-war period, saw his teams achieve previously unprecedented levels of physical fitness and tactical acumen.

12. Eckard (2001) argues along similar lines that a lack of player mobility prevented the NA major league baseball players' labour market from functioning in accordance with the EHQ model prior to the introduction of free agency in 1976.

References

Bennett, R.W. and Fizel, J.L. (1995) Telecast deregulation and competitive balance, *American Journal of Economics and Sociology* 54, 183-199.

Dobson, S.M. and Goddard, J.A. (1998) Performance, revenue and cross subsidisation in the English Football League, 1927-94, *Economic History Review* 51, 763-785.

Dobson, S.M. and Goddard, J.A. (2001) *The economics of football*. Cambridge: Cambridge University Press.

Dunning, E., Murphy, P., and Williams, J. (1988) *The roots of football hooliganism: an historical and sociological study.* London and New York: Routledge and Kegan Paul.

Eckard, E.W. (1998) The NCAA cartel and competitive balance in college football, *Review of Economic Organization* 13, 347-369.

Eckard, E.W. (2001) Free agency, competitive balance, and diminishing returns to pennant competition, *Economic Inquiry* 39, 430-443.

El-Hodiri, M. and Quirk, J. (1971) An economic model of a professional sports league, *Journal of Political Economy*, 79, 1302-1319.

Football Trust (various) *Digest of football statistics*. Leicester: University of Leicester.

Fort, R. (2000) European and North American sports differences (?) *Scottish Journal of Political Economy* 47, 431-455.

Fort, R. and Quirk, J. (1995) Cross-subsidization, incentives, and outcomes in professional team sports leagues, *Journal of Economic Literature* 33, 1265-1299.

Hoehn, T. and Szymanski S. (1999) The Americanization of European football, *Economic Policy*, 205-240.

Késenne, S. (1996) League management in professional team sports with win maximising clubs, *European Journal for Sport Management* 2, 14-22.

Késenne, S. (1999) Player market regulation and competitive balance in a win maximising scenario, in: Késenne, S. and Jeanrenaud, C. (eds.) *Competition policy in professional sports: Europe after the Bosman case*. Antwerp: Standaard Editions.

Késenne, S. (2000) The impact of salary caps in professional team sports, *Scottish Journal of Political Economy* 47, 422-430.

Koning, R.H. (2000) Balance in competition in Dutch soccer, *The Statistician* 49, 419-431.

Quirk, J. and Fort, R. (1999) *Hard ball: the abuse of power in pro team sports.* Princeton: Princeton University Press.

Rottenberg, S. (1956) The baseball player's labor-market, *Journal of Political Economy* 64, 242-258.

Russell, D. (1997) *Football and the English: a social history of association football in England, 1863-1995.* Preston: Carnegie.

Scully, G. (1989) *The business of Major League Baseball*. Chicago: University of Chicago Press.

Sloane, P. (1971) The economics of professional football: the football club as utility maximiser, *Scottish Journal of Political Economy* 17, 121-146.

Szymanski, S. (2001) Income inequality, competitive imbalance and the attractiveness of team sports: some evidence and a natural experiment from English soccer, *Economic Journal* 111, F59-F84.

Vrooman, J. (1995) A general theory of professional sports leagues, *Southern Economic Journal* 61, 971-990.

Walvin, J. (1994) *The people's game: the history of football revisited.* Edinburgh, Mainstream.

Weiss, A.A. (1997) Specification tests in ordered logit and probit models, *Econometric Reviews* 16, 361-391.

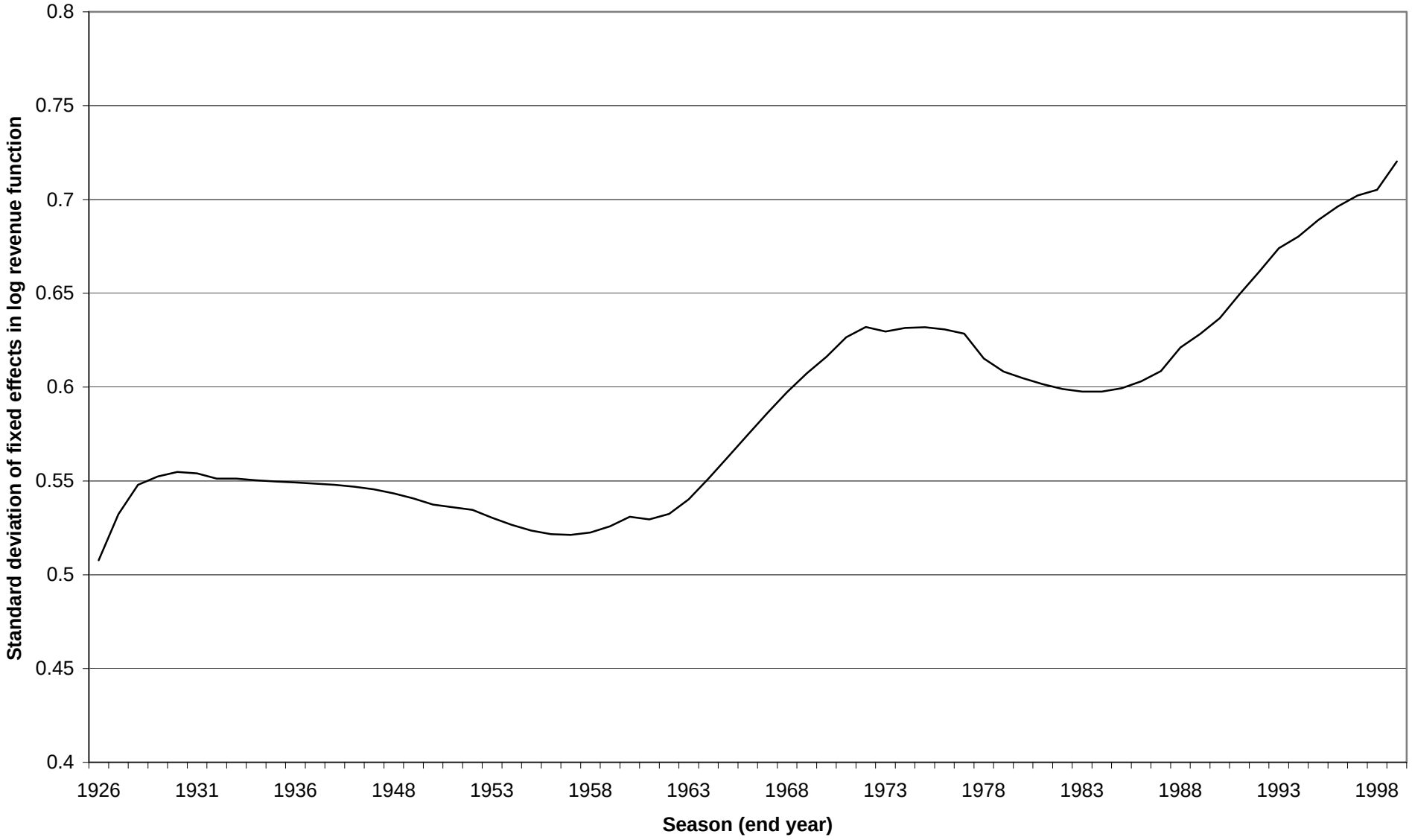
Table 1 Estimation results: revenue function

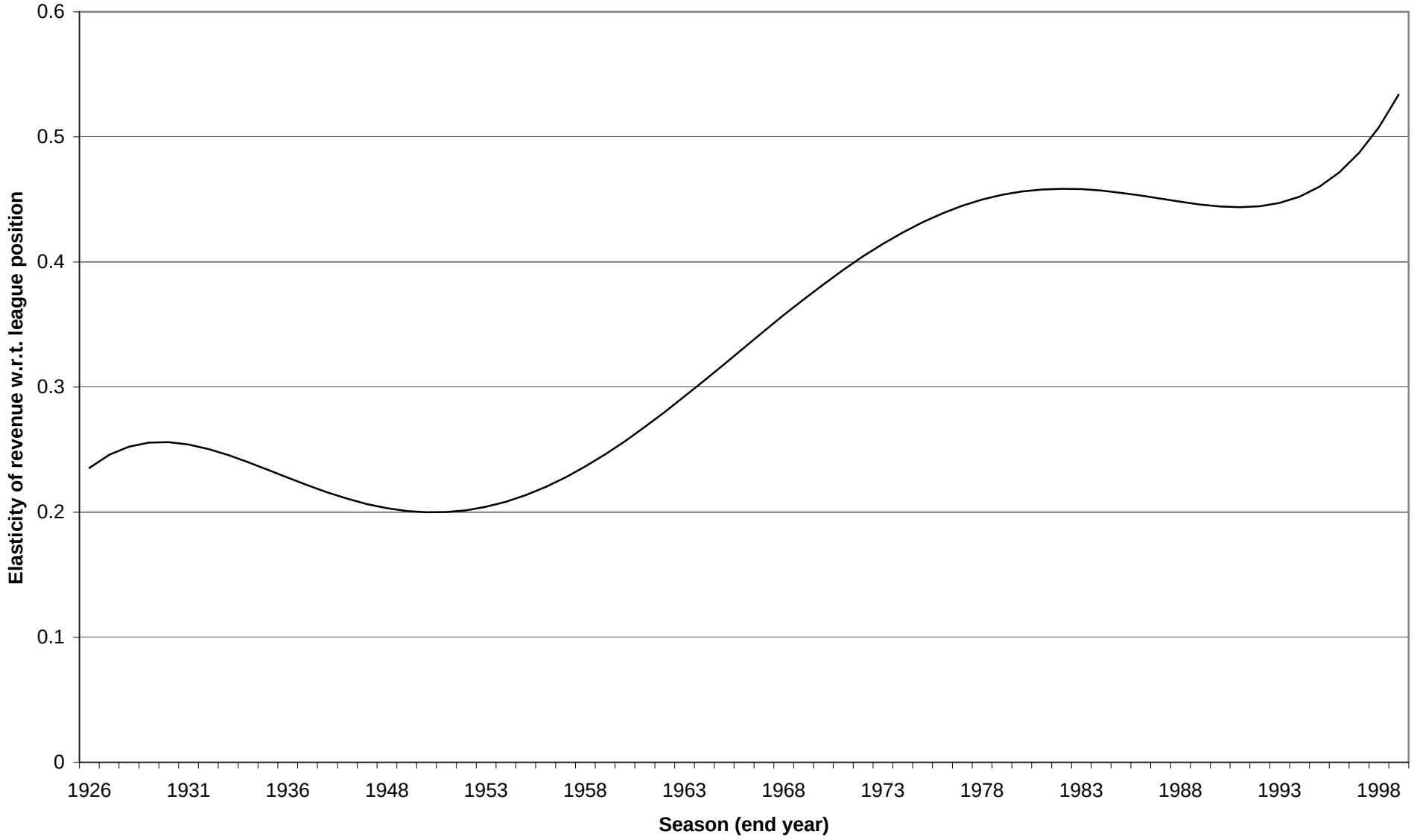
Season end-year = t:								
	1926	1936	1949	1959	1969	1979	1989	1999
(a) League performance score (%) by group:								
G1	26.9	25.5	25.1	25.7	26.3	25.0	24.1	25.9
G2	19.0	19.6	17.7	16.0	16.4	16.4	17.8	17.3
G3	18.4	22.4	24.1	27.0	30.7	35.1	36.2	30.0
G4	21.3	21.4	20.7	18.6	15.9	13.0	12.3	17.2
G5	14.4	11.1	12.4	12.8	10.7	10.5	9.6	9.7
(b) % share of gate revenue by group:								
G1	32.8	34.9	29.0	31.7	40.2	39.3	39.9	43.5
G2	18.6	17.9	21.3	17.6	19.2	16.5	18.3	20.8
G3	21.6	23.8	24.2	25.9	24.8	27.6	27.4	19.6
G4	16.9	15.9	16.3	15.3	9.4	10.2	8.7	12.5
G5	10.1	7.4	9.1	9.5	6.4	6.5	5.7	3.5
(c) Group mean values of $\exp(\hat{\alpha}_{i,t})$ (see [11]):								
G1 (g=1)	0.025	0.024	0.022	0.023	0.028	0.031	0.030	0.030
G2 (g=2)	0.016	0.016	0.017	0.016	0.016	0.016	0.016	0.019
G3 (g=3)	0.012	0.011	0.011	0.011	0.010	0.010	0.010	0.008
G4 (g=4)	0.012	0.011	0.012	0.011	0.009	0.009	0.008	0.009
G5 (g=5)	0.007	0.006	0.006	0.006	0.005	0.006	0.005	0.004
(d) Cross-sectional standard deviation of $\hat{\alpha}_{i,t}$ (see [11]):								
$s(\hat{\alpha}_{i,t})$	0.5077	0.5492	0.5407	0.5258	0.6074	0.6082	0.6283	0.7202
(e) $\hat{\beta}_t$ = elasticity of revenue w.r.t. league position (see [3] and [11]):								
$\hat{\beta}_t$	0.2354	0.2277	0.2010	0.2462	0.3700	0.4536	0.4459	0.5337

Note: League performance scores (panel (a)) are calculated by expressing the sum of $L_{i,t}$ over all teams in Group g as a percentage of the sum of $L_{i,t}$ over all teams.

Table 2 Estimation results: competitive balance

Season end-year = t:								
	1926	1936	1949	1959	1969	1979	1989	1999
(a) Proportion of home wins, draws and away wins, cup matches, 10 seasons up to and including t:								
Home wins	n/a	0.530	0.531	0.464	0.485	0.500	0.487	0.453
Draws	n/a	0.218	0.199	0.244	0.245	0.268	0.250	0.259
Away wins	n/a	0.252	0.270	0.293	0.271	0.232	0.263	0.288
(b) Ordered probit estimation results – first match (see [12]): $\hat{\mu}_1$ = -0.881; $\hat{\mu}_2$ = -0.089; obs. = 5438								
$\hat{\gamma}_{0,t}$	-0.013	-0.112	-0.146	-0.172	-0.167	-0.145	-0.119	-0.101
$\hat{\gamma}_{1,t}$	1.160	1.314	1.335	1.301	1.246	1.242	1.363	1.682
$\hat{\gamma}_{2,t}$	-1.141	-1.417	-1.458	-1.407	-1.314	-1.302	-1.498	-2.024
(c) Ordered probit estimation results – replay (see [12]): $\hat{\mu}_1$ = -0.916; $\hat{\mu}_2$ = -0.545; obs. = 1498								
$\hat{\gamma}_{0,t}$	-0.021	-0.183	-0.242	-0.289	-0.290	-0.265	-0.234	-0.217
$\hat{\gamma}_{1,t}$	1.303	1.042	1.022	1.136	1.338	1.515	1.554	1.342
$\hat{\gamma}_{2,t}$	-1.352	-1.230	-1.210	-1.255	-1.380	-1.567	-1.798	-2.055
(d) Selected win probabilities conditional on (re-scaled) league position:								
$w_t(0.75,0.25)$	0.747	0.781	0.787	0.784	0.779	0.783	0.809	0.862
$w_t(0.90,0.10)$	0.853	0.892	0.897	0.895	0.888	0.891	0.915	0.957
(e) Cross-sectional standard deviation of win probabilities against median team:								
$s(w_t)$	0.1504	0.1712	0.1733	0.1719	0.1686	0.1715	0.1883	0.2261


0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
Standard deviation of fixed effects in log revenue function
1926
1931
1936
1948
1953
1958
1963
1968
1973
1978
1983
1988
1993
1998
Season (end year)


0.6
0.5
0.4
0.3
0.2
0.1
0
Elasticity of revenue w.r.t. league position
1926
1931
1936
1948
1953
1958
1963
1968
1973
1978
1983
1988
1993
1998
Season (end year)


0.25
0.225
0.2
0.175
0.15
0.125
0.1
Standard deviation of win probabilities against median team
1926
1931
1936
1948
1953
1958
1963
1968
1973
1978
1983
1988
1993
1998
Season (end year)