1 Introduction

Academic interest in the economics of professional team sports dates back as far as the mid-1950s. Since then, many books and journal articles have been written on the subject. Much of the academic literature originates in the United States (US). In common with trends that are evident throughout the subject discipline of economics, empirical research on the economics of sport has become increasingly sophisticated, both theoretically and in its use of econometric methodology. Papers on the economics of sport now appear regularly in many of the leading economics journals, and most economists would agree that in view of its social, cultural and economic importance, professional sport is a legitimate area of interest for both theoretical and empirical researchers. Indeed, many would argue that the unique configurations of individual and team incentives, and the interactions between cooperative and competitive modes of behaviour that professional team sports generate, make this particularly fertile territory in which to explore the perennial questions about incentives, effort, risk and reward that lie at the heart of all areas of economic inquiry.

This volume makes a contribution to the burgeoning literature on the economics of team sports, by providing a comprehensive survey of research that is focused on professional football. The spectacular recent increase in the size of football's audience is, of course, a strong motivating factor. Such a survey will recognise and reflect not only football's global popularity in the first part of the twenty-first century, but also the special historical significance of England as the original birthplace of the sport. Club football played in the English Premier League and Football League provides the laboratory for most of the original, empirical research that is reported in this volume.

Each chapter of this volume concentrates on a particular aspect of the economics of professional football. The previous theoretical and empirical literature that is relevant to each topic is reviewed, and new and original empirical analyses are presented. The review sections aim to convey an impression of the breadth and depth of previous academic research into the economics of professional team sports. Much has been written already about football, and much more has been written about other sports, especially in the US where attention naturally tends to

focus on the traditional major league sports of baseball, basketball, (American) football and hockey. Though football is the main subject of this volume, due attention and emphasis is devoted to insights that have been obtained from research into other sports, wherever these turn out to be of wider relevance.

As already emphasised, anyone who reads the academic literature on the economics of sport cannot fail to be struck by the sheer volume of column inches that have been devoted to this topic. Of all the articles that have been published, however, a few early contributions were particularly important in shaping the research agenda for the economics of team sports in general, and for the economics of football in particular. Many of the ideas contained in these articles are as relevant to researchers today as they were when the articles were originally published, many years ago. In Chapter 1, three such articles are highlighted and reviewed in some detail. There is, of course, an element of subjectivity in selecting such a small number of articles out of the many that have been published. Even so, a consensus seems to have evolved that regards the articles by Simon Rottenberg (1956) and Walter Neale (1964) as fundamental to the subsequent development of research on team sports in general. Both articles address various economic implications of the structural features of the markets within which professional sports teams operate. The article by Peter Sloane (1971) has also had a major influence on the developing research agenda, especially in respect of the economics of football. In Section 1.1, each of these articles is reviewed in turn, and the subsequent development of the economics of team sports as an academic discipline is outlined. The aim of Section 1.1 is to place the research that is reported in the rest of this volume into its proper context. This is followed in Section 1.2 by a summary of the contents of the remaining chapters of this volume.

1.1 The economics of professional team sports: three seminal contributions

Rottenberg: 'The Baseball Players' Labor Market', *Journal of Political Economy*, 1956

Rottenberg is widely credited with writing the first academic analysis of the economics of professional team sports. The paper was written at a time when US professional baseball players' contracts included a reserve clause. Once a player signed his first one-year contract with a team in Major League Baseball (MLB), he ceased to be a free agent. On expiry of his present contract, his team retained the option to renew his contract for another year.[1] This served to limit players' freedom of movement, by binding them to their present employers. Effectively, the reserve clause created a monopsony in the players' labour market: each contracted player could negotiate with only one potential buyer of his services. The baseball authorities defended the reserve clause on the grounds that it was necessary to ensure an equal distribution of playing talent among opposing teams. Without the reserve clause, the rich teams (with the largest potential markets) would outbid the poorer ones for the best available players. This would tend to reduce uncertainty of

outcome and spectator interest in the league competition as a whole, and depress the attendances and revenues of all teams.

Rottenberg's contribution was to argue that free agency in the players' labour market would not necessarily lead to a concentration of the best players in the richest teams. In other words, a reserve clause was not a necessary condition to ensure competitive balance. Professional team sports are intrinsically different from other businesses, in which a firm is likely to prosper if it can eliminate competition and establish a position as a monopoly supplier. In sports, it does not pay a rich team to accumulate star players to the extent that (sporting) competition is greatly diminished, because of the joint nature of 'production' in sports. Consequently, a team that attempts to accumulate all of the best available playing talent will find at some stage that diminishing returns begin to set in.

> In baseball no team can be successful unless its competitors also survive and prosper sufficiently so that the differences in the quality of play among teams are not 'too great.' … At some point, therefore, a first star player is worth more to poor team B than, say, a third star to rich team A. At this point, B is in a position to bid players away from A in the market. (Rottenberg, 1956, p254, 255)

If teams are rational profit maximisers, the distribution of playing talent among the teams should be more or less equal. Neither a reserve clause nor explicit collusion is necessary in order to bring about this result. It is in each team's self-interest to ensure that it does not become too strong relative to its competitors.

> It follows that players will be distributed among teams so that they are put to their most 'productive' use; each will play for the team that is able to get the highest return from his services. But this is exactly the result which would be yielded by a free market. (Rottenberg, 1956, p256)

A reserve clause will therefore deliver almost the same distribution of playing talent between the competing teams as free agency. Whether players are free agents or not, the distribution of playing talent is determined by the incentive to maximise the capitalised value of the services supplied by individual players. If there is another team for which this capitalised value would be higher than it is for the player's present team, then there is a price at which it is advantageous for both teams to trade the player's contract.

Rottenberg also discusses the implications of the reserve clause, and the monopsony power it confers on teams as buyers of playing services, for players' salaries. Each player's reservation wage (the minimum salary he would accept to play baseball) is determined primarily by the next highest salary he could earn outside baseball adjusted to reflect his valuation of the non-pecuniary costs and benefits of playing baseball. Although theoretically the team has the contractual power to impose the reservation wage on all players, Rottenberg notes that in practice this does not seem to happen. Many players earn far more from baseball than they could in alternative employment. This is attributed to the fact that players as well as teams have bargaining power in salary negotiations: in an extreme case, a player can

simply threaten to withdraw his services. If a player's reservation wage is $10,000, but he is worth $20,000 to his team, then a salary anywhere between $10,000 and $20,000 is possible, depending on the 'shrewdness and guile of the parties in devising their bargaining strategies' (Rottenberg, 1956, p253). Competition among sellers, however, imposes limitations on players' bargaining power. A star player worth $40,000 to his team cannot extract a salary beyond $30,000 if a lesser player worth $20,000 is willing to accept a salary of $10,000 to fill the same position.

The main effect of the reserve clause is that players receive salaries below their value to the team that employs them. In other words, it tends to direct rents away from players and towards teams. The reserve clause does not achieve its stated aim of influencing the allocation of playing talent between teams. Rottenberg concludes by considering several alternative regimes that might produce a more or less equal distribution of playing talent between teams with free agency in the players' labour market. These include:

- Revenue sharing. If all revenues are shared equally, teams have no pecuniary incentive to spend on players to enhance their own performance and revenue. An equal distribution of mediocre playing talent is the most likely outcome.
- Imposition of a maximum salary. The effect on the distribution of talent depends on the ability of the teams to circumvent the maximum by offering players non-pecuniary rewards.
- Allocating multiple team franchises in large cities. If reasonable equality between each team's potential market size can be achieved, this is expected to create a more equal distribution of playing talent.

Neale: 'The Peculiar Economics of Professional Sports', *Quarterly Journal of Economics*, 1964

Neale's analysis begins by emphasising the joint nature of production in professional sports. Heavyweight boxing is used as an example to introduce what Neale calls the 'Louis Schmelling Paradox'. World champion Joe Louis' earnings were higher if there was an evenly matched contender available for him to fight than if the nearest contender was relatively weak. The same point applies also in baseball.

> Suppose the Yankees used their wealth to buy up not only all the good players but also all of the teams in the American League: no games, no receipts, no Yankees. When, for a brief period in the late fifties, the Yankees lost the championship and opened the possibility of a non-Yankee World Series they found themselves – anomalously – facing sporting disgrace and bigger crowds. (Neale, 1964, p2)

Does this imply that professional sport is an industry in which monopoly is less profitable than competition, contradicting what is taught to students and what can be read in any Principles text? Neale addresses this paradox by distinguishing between sporting and economic competition. Sporting competition is more profitable than sporting monopoly for the reasons outlined above, but sporting

competition is not the same as economic competition. Similarly, although in law the sports team is a firm (which may be motivated by profit), it is not a firm in the economist's sense. A single team cannot supply the entire market; if it did, it would have no opposition to play against. Teams must cooperate with each other to produce individual matches and a viable league competition, so there is joint production. The league's organising body exerts strict controls over a wide range of matters including competition rules and schedules, player mobility and the entry and exit of clubs. In short, the league rather than the individual team is the 'firm' in the economist's sense. A sports league should be regarded as analogous to a multi-plant firm, in which the individual teams are 'plants', subject to decisions which are taken and implemented collectively at league level.

If the sports league is the 'firm' in the economic sense, this raises the question as to why it is unusual to observe direct competition between rival leagues operating within the same sport. Although the National League and American League do operate simultaneously in baseball, analytically they should be regarded as one larger 'multi-league' firm, since they come together at the end of each season to produce the World Series. Geographical division is a more common form of segmentation, though according to Neale, one that is inherently unstable. Where direct competition is precluded by geography, profit incentives tend to promote enlargement and the elimination of geographical boundaries. Competition between different sports is more common than competition between rival leagues within the same sport, though segmentation based on nation, region, season of the year or even social class is still common.[2] All such forms of segmentation tend to inhibit direct head-to-head competition.

Neale suggests that the general lack of competition between sports leagues arises because the cost and demand characteristics of the market for professional team sports tend to create conditions of natural monopoly, making it efficient for a single league to supply the entire market. On the cost side, Neale suggests that the long-run average cost curve is probably horizontal. Although an increase in the scale of production might entail the use of less efficient playing inputs, raising average costs, this tends to be offset by an 'enthusiasm effect'. If the sport operates on a larger scale, public enthusiasm encourages more people to take up playing, eventually raising the supply of players at the highest level. To some extent, the enthusiasm effect makes supply and demand interdependent: if more people play the sport, more will also want to attend matches at the professional level. Finally, the existence of rival leagues would effectively break the monopsony power of teams as buyers of playing services, enhancing players' bargaining power in salary negotiations. This would tend to make costs higher than they are when one league operates as a monopoly supplier.

On the demand side, Neale suggests that baseball teams produce a number of streams of utility: directly for spectators who buy tickets for seats in the stadium and for television viewers who watch the match at home; and indirectly for anyone who enjoys following the championship race as it unfolds. The closer the

competition, the greater the indirect effect. For newspapers and television companies in particular, the indirect effect is a marketable commodity that helps sell more of their product. The size of the indirect effect depends on the scale and universality of the championship, and is therefore maximised when the league is a monopoly supplier. Overall, 'it is clear that professional sports are a natural monopoly, marked by definite peculiarities both in the structure and in the functioning of their markets' (Neale, 1964, p14).

An important implication is that the peculiar economic characteristics of professional sports leagues and their constituent teams should be recognised by legislatures, by the courts and by the general public, whenever practices such as collective decision-making or other (apparently) anticompetitive types of behaviour come under scrutiny.

Sloane: 'The Economics of Professional Football: The Football Club as a Utility Maximiser', *Scottish Journal of Political Economy*, 1971

Sloane's paper questions Neale's conclusion that the league rather than the individual team or club is the relevant 'firm' (or decision-making unit) in professional team sports. In the case of English football, for example, the sport's governing bodies merely set the rules within which clubs can freely operate. Most economic decisions, such as how much money to spend on stadium development and how many players to employ, are made by the clubs. Although the total quantity of 'output' (the number of matches played by each team) is regulated, this clearly reflects the clubs' common interest. In cartels, it is not unusual for firms to reach joint decisions concerning price or production, but this does not imply that the cartel should be elevated to the theoretical status of a 'firm'. In short, Sloane suggests that Neale's argument tends to overemphasise mutual interdependence. 'The fact that clubs together produce a joint product is neither a necessary nor a sufficient condition for analysing the industry as though the league was a firm' (Sloane, 1971, p128).

Having argued that the club is the relevant economic decision-maker, Sloane goes on to raise a number of key questions concerning the objectives of sports clubs. Implicit in the reasoning of both Rottenberg and Neale is an assumption of profit maximisation. Despite the 'peculiarities' of sports economics elucidated by Neale, the behaviour of professional sports teams is analysed within a very conventional analytical framework. While this may be reasonable in the case of US professional team sports, where many teams do have an established track record of profitability, Sloane suggests that it may not be universally applicable. Throughout the history of English football, profit-making clubs have been the exception and not the rule. Most chairmen and directors of football clubs are individuals who have achieved success in business in other fields. Their motives for investing may include a desire for power or prestige, or simple sporting enthusiasm: a wish to see the local club succeed on the field of play. In many cases, profit or pecuniary gain seems unlikely to be a significant motivating factor. If so, it may be sensible to view the objective of the football club as one of utility

maximisation subject to a financial solvency constraint. The financial solvency constraint recognises that the benevolence of any chairman or director must reach its limit at some point.

Non-profit-maximising models of the firm had received considerable attention in the economics literature during the decade prior to the publication of Sloane's article.

> A major drawback to the general introduction of the utility maximisation assumption in the theory of the firm is that it may be rationalised so that it is consistent with almost any type of behaviour and therefore tends to lack operational significance. (Sloane, 1971, p133)

In the case of football clubs, however, it is not too difficult to identify several plausible and easily quantifiable objectives. Sloane suggests the following:

- Profit. The expectation that profit is not the sole or even the most important objective does not preclude its inclusion as one of several arguments in the utility function.
- Security. Simple survival may be a major objective for many clubs. Decisions (concerning, for example, sales of players) may aim more at ensuring security than at maximising playing success.
- Attendance or revenue. A capacity crowd enhances atmosphere and a sense of occasion, and may in itself be seen as a measure of success. Recently, an increasing willingness to charge whatever ticket prices the market will bear suggests that revenue (or profit) carry a heavier weight in the utility function than in earlier periods, when it was usual to charge the same price for all matches, irrespective of the level of demand.
- Playing success. This is probably the most important objective of all, and one to which chairmen, directors, managers, players and spectators can all subscribe.
- Health of the league. This enters the utility function in recognition of clubs' mutual interdependence.

Formally, the club's objective is to maximise:

$$U = u(P,A,X,\pi_R - \pi_0 - T) \qquad \text{subject to } \pi_R \geq \pi_0 + T \qquad [1.1]$$

where P = playing success; A = average attendance; X = health of the league; π_R = recorded profit; π_0 = minimum acceptable after-tax profit; and T = taxes.

It is important to note that the utility maximisation model has implications that are very different to those that follow from the profit-maximising assumptions of Rottenberg and Neale. In particular, if the weighting of P in the utility function is heavy relative to that of X and π_R, the argument that diminishing returns would prevent the accumulation of playing talent in the hands of a small number of rich clubs does not necessarily hold, unless there are binding financial constraints preventing expenditure on new players. The notion that profit incentives should help maintain a reasonably even allocation of playing talent between richer and poorer teams breaks down. The case for regulation to override the 'free market' outcome, whether in the form of a reserve clause, revenue sharing or the taxation of transfer

fees, therefore seems to be enhanced if clubs are pursuing non-profit rather than profit objectives.

1.2 Outline of this volume

Since the appearance of the pioneering work of Rottenberg, Neale and Sloane, there has been a proliferation of published academic research on the economics of team sports, in journal articles and in books. Many of these contributions are reviewed in this volume. By way of an introduction, this section contains a brief outline of the contents of each of the following chapters.

Rottenberg's argument – that market mechanisms can be relied upon to maintain a reasonable degree of competitive equality among the member teams of a sports league – provides the motivation for a number of theoretical contributions to the sports economics literature, examining the implications for resource allocation within a professional sports league of several regulatory mechanisms, including the reserve clause, payroll caps, the reverse-order-of-finish draft, revenue sharing, and rules governing the award of franchises and league membership. According to the theoretical analysis, most of these devices promote the survival of small-market teams; but in many cases they fail to reduce competitive inequality because they do not create the necessary profit incentives. Chapter 2 reviews some of the key findings from this literature.

The theoretical models of the economics of sports leagues are concerned with the degree of competitive balance or competitive inequality between the member teams of a league competition, which in turn determines the extent of uncertainty of outcome for individual match results and for the destination of the league championship. In Chapter 3, the emphasis shifts away from this theoretical analysis at the level of the league championship, towards the measurement of competitive inequality and uncertainty of outcome for individual match results. Certain other empirical patterns and regularities in football match results data are investigated, including the phenomenon of home-field advantage, and 'streaks' or persistence effects in sequences of consecutive match results.

Uncertainty of outcome seems to be an essential ingredient for any competitive sport. If the match result were completely deterministic, what would be the point for either competitors or spectators? The observation that the result of any individual match is uncertain, however, does not immediately consign any forecasting exercise to irrelevance. If prediction is interpreted as assessment of the probabilities for home-win, draw and away-win outcomes, the results of individual football matches are highly predictable, as evidenced by the wide variation from one match to the next in the betting odds quoted by bookmakers and online betting firms for these outcomes. Chapter 4 examines the specification and performance of econometric forecasting models for match results in football, which generate probabilistic forecasts of the results of forthcoming matches, by number-crunching large volumes of past match results data and other relevant information.

Game theory is defined as the study of decision-making in situations of conflict and interdependence. Some economists have argued that sports such as football offer promising opportunities for the investigation of propositions of game theory, concerning the formulation by the players of optimising strategies. For example, the high-pressure 'game' played out between the kicker and the goalkeeper whenever a penalty kick is awarded mirrors closely the textbook conditions for a two-person non-cooperative zero-sum simultaneous game. The kicker decides in which direction to shoot and the goalkeeper decides simultaneously in which direction to dive; and either the kicker scores, or a goal is averted because the goalkeeper saves or the kicker misses the target. Viewed more broadly, the football match in its entirety has many of the characteristics of a strategic and dynamic 'game'. The two teams are pitched into direct opposition, and the team managers select and adjust continuously their styles of play as the match evolves. Chapter 5 reviews the theoretical and empirical literature, and presents some new empirical evidence on applications of game theory in football.

The sporting attributes of professional football and its characteristics as a business have always been closely connected. Chapter 6 presents an overview of the historical development of English club football as a business, and an analysis of its contemporary economic, financial and commercial structure. The analysis covers English club football's competitive structure, trends in spectator demand reflected in match attendances and gate revenues, the market for TV broadcast rights, the development of the labour market for playing talent, and the ownership, governance and financing of English football clubs.

The level of the top professional footballers' remuneration has raised concerns over the effect on the competitive structure and finances of football, and over the morality of such disproportionately high financial rewards accruing to footballers relative to other occupations whose contribution to society's well-being is, arguably, much greater. Chapter 7 examines theoretical explanations that have been put forward for the exceptionally high remuneration of the leading stars in modern-day professional football, and other professional sports. The theoretical analysis of the economics of superstars notes that, especially since the arrival of pay-TV, sports stars are capable of servicing very large paying audiences simultaneously, incurring little or no incremental cost as the audience size increases. Other occupational groups, such as nurses and teachers, service strictly finite number of users. The rank-order tournament model views the remuneration of the highest-paid stars as a means of providing all professionals with incentives to invest in the development of their skills, in the hope of reaching the top of their profession and being rewarded accordingly.

Chapter 8 discusses several other topics concerned with the economics of the professional footballers' labour market. Patterns of migration, employment mobility and career development among the professional footballers employed by the ninety-two member-clubs of the Premier League and Football League are examined, including the impact of the arrival of large numbers of overseas players on

the career prospects of locally born footballers. The factors influencing patterns of international migration by footballers are examined. Empirical evidence on racial discrimination in English football suggests that a form of hiring discrimination affects the opportunities for indigenous black players to progress to professional status. It seems likely, however, that any such effect has diminished over time.

The job description of the football team manager includes the selection, supervision and coaching of players, and the formulation of tactics and strategies. Many football managers, especially in professional football's lower tiers, are also responsible for the buying and selling of players, wage negotiations, and various administrative duties. Chapter 9 examines various aspects of the role and contribution to team performance of the football manager. In principle, the football manager's contribution subdivides into a direct and an indirect component. The direct contribution is to maximise performance through astute team selection, superior tactics and powers of motivation; while the indirect contribution is to coach players so as to enhance their skills, and to strengthen the team through effective dealings in the transfer market. The indivisibility of the team effort which ultimately determines performance poses a major challenge for any researcher seeking to isolate and measure the manager's contribution. The modern-day football manager's position is renowned for its chronic insecurity, which raises several interesting questions for researchers concerning the relationship between managerial turnover and team performance.

While it is the football manager who usually pays the ultimate price for perceived underachievement on the part of the players under his direction, the football referee often serves as a convenient scapegoat during the immediate aftermath of a poor result or performance. Football referees, who are the subject of Chapter 10, are routinely criticised by managers, players, journalists and spectators for being incompetent, inconsistent and biased. Several recent academic studies have examined whether there is any substance to claims of favouritism and bias on the part of football referees. Due to technological advances in broadcasting, the actions of referees have never been more intensely scrutinised than they are today. Split-second decisions taken by referees can have enormous financial consequences, due to the fine line between spectacular success and catastrophic failure that exists in football.

Social history and sociology provide many useful insights into the causes of fluctuating football attendances. Econometric modelling of variations in the attendances of individual clubs, both season-by-season and match-by-match, has been a subject of attention for sports economists since the 1970s. Chapter 11 provides a non-technical review of the empirical literature, focusing on the issues of variable definition, model specification, estimation and interpretation that are faced by researchers in this area, and presents an analysis of the variation in the average attendances per season of English football clubs during the post-Second World War period.

Research concerning the relationship between the market prices for bets on the outcomes of sporting contests, and the probabilities associated with these