1. An introduction to personnel economics and its application to sport

Neil Longley

1.1 A BASIC PRIMER ON THE APPROACHES OF PERSONNEL ECONOMICS

In order to provide a framework for studying personnel economics within the specific context of sport, it is important to first understand the approaches used in the broader (i.e. non-sport) field of personnel economics. With this base, one can then better appreciate how the concepts can be applied to sport, and how the institutional peculiarities of the sport industry sometimes complicate these applications.

The personnel economics literature has now become vast, so the overview provided here provides more of a high-level primer on the basic conceptual approaches in the field, rather than a detailed and exhaustive examination of specific topics and issues.[1]

Recruiting

The recruiting process involves the firm conducting a search to fill a vacant position – by its nature, it is the first step in the employer-employee relationship. The process involves the firm first generating a pool of applicants, and then using various screening mechanisms to select the best candidate from that pool. The goal of the firm is to hire the most productive employee, relative to the costs of finding that employee.

Since screening is costly, a key choice variable for the firm is to decide the amount of screening it will undertake for any given job. This decision will be driven by a number of factors. First, the nature of the job will matter. For example, lower-skilled jobs will tend to have more routine and structured tasks, so the potential variability in performance across employees will often be relatively small. In this case, one employee will be almost as productive as any other, and firms may not expend many resources on

the recruitment process. However, as one moves up the job hierarchy in an organization, jobs inherently become more complex, multi-dimensional, and less well-defined. With these jobs, subtle, difficult to measure differences across applicants may ultimately result in large performance differences. This greater dispersion in potential performance – compared to the situation with lower-skilled jobs – forces firms to adopt more intensive hiring processes so as to identify those who are most capable.

The firm's production technology – in the sense of how various jobs in the organization interact with each other – also impacts the search. In the most basic case, a worker's productivity is predetermined and set, and is independent of that individual's co-workers. While applicants may differ in productivity, that productivity is unrelated to where an applicant works. Thus, there is no firm-specific matching, in the sense that one candidate would be a better 'fit' for that particular organization than would another. Instead, the search merely involves the organization identifying the candidate(s) with the highest potential productivity.

A more complex scenario is where a candidate's productivity is a function of the organization for which they work – i.e. the candidate's value is higher to one organization than to another. The search task then becomes more than simply identifying the most productive applicant, and instead becomes more about determining whether the candidate is the best 'match'; Oyer and Schaefer (2010) term this 'match-specific productivity'. Match-specific productivity results from complementarities between the candidate and the organization – either complementarities between the candidate's skills and the skills of existing employees, or complementarities between the candidate's skills and the production processes of the organization. The hiring decision is now much more complicated, in that a firm must not only seek to determine the candidate's individual attributes/quality, but must now also estimate how well that candidate will interact with the firm's existing workers. In situations where the screening process fails to identify an effective employer-employee match, the result is often employee turnover. Such turnover is potentially costly to the organization, not only because it requires a new search, but because training costs incurred for the departed employee are not recoverable.

At the operational level, recruiting is essentially an exercise in forecasting – it seeks to predict the future job performance of applicants. For firms, an applicant's past performance is often one of the most critical screening mechanisms used to forecast future performance. However, information on a worker's past performance, particularly as it relates to the specific job in question, is often difficult to obtain. First, for younger workers, because they are relatively new to the labor market, little or no measures of past performance would even exist. Second, even when employees are

experienced, their past performance is with another firm, and is often difficult for other employers to evaluate. Job-seekers also have an incentive to overstate their past performance, particularly where any new employer would have difficulty verifying such information.

Because the screening process is inherently characterized by bilateral asymmetric information – i.e. the candidate knows more about their own skills than does the prospective employer, and the employer has more information about the specifics of the job and the work environment than does the prospective employee – the parties will often attempt to 'signal' their worthiness and quality to the other. The concept of signaling was developed by Spence (1973). An employer may use its public reputation as a signal that it is a good place to work. For job-seekers, a signal is a credential that, while not directly related to the duties of a job, is intended to proxy those workers' productivity. In most labor markets, education is perhaps the most common signal. Factors like education are considered 'signals', in that they do not necessarily measure attributes directly related to the job in question, but rather are proxies for other qualities that are presumed to be related to job performance. For a signal to be effective as a separating mechanism, high-quality candidates must be able to obtain the signal at a lower cost than low-quality candidates; for example, higher-aptitude individuals should be able to obtain a graduate degree at lower cost (in terms of effort) than lower-aptitude individuals. Signals tend to be most effective in situations where there is a limited amount of other information about the candidate; for example, with workers new to the labor force.

Firms can bring new employees into the organization in one of two basic ways. They can hire largely at the entry level, and eventually promote these individuals to fill higher positions in the organization. Or, for at least some higher-level positions, the firm can go outside and 'raid' workers from other firms.

When a firm hires/raids workers from other firms, it will do so only if the worker is more valuable to the raiding firm than to the individual's current employer. One possibility is that both firms (the raider and the current employer) assess the individual's future productivity similarly, but that the raider can convert this productivity into greater revenues, say due to firm size or market share. Another possibility is that the raider is more optimistic about the individual's future productivity than is the current employer – estimating future performance is an inexact science, particularly where job duties are complex, and where difficult-to-measure 'intangible' factors are critical factors to job success. Different firms may simply have different subjective assessments about the employee's potential.

However, raiding firms face the problems of both the 'winner's curse' and adverse selection. With the winner's curse, the raiding firm is, by

definition, the firm in the marketplace that places the most optimistic estimate on the individual's future productivity, and hence may overpay for that person. With adverse selection, the raiding firms may disproportionately attract lower-productivity candidates, since these individuals' current employers will make less effort to retain them, relative to that firm's more productive employees.

As a supplement to pre-employment screening, employers can use an employee's 'probationary period' as an additional source of information – in essence, part of the screening occurs *after* the person is hired. This probationary information is particularly valuable because it is based on observing actual on-the-job performance; information that was clearly not available at the time of the hire. In these cases, a natural sorting occurs during the probationary period, where both employers and employees further evaluate the quality of the match. Thus, one would expect turnover to be higher for more recent hires than for those who have been with the organization for a longer period of time.

Furthermore, when firms require new employees to go through a probationary period, it facilitates a natural sorting and self-selection; it can discourage unproductive workers from applying, since they know they are likely to be 'discovered'. In conjunction with this, employers often pay probationary workers much less than those that have achieved permanent status; this low wage during probation may not be a problem for productive employees, since they can more than recoup this by earning higher wages after they have achieved permanency. However, for unproductive workers, i.e. those who are unlikely to make it past probation, this low wage is intended to discourage them from applying to the organization, and instead seek work with firms at which they are better suited (and, presumably, at which they will be paid more). In effect, then, the low probationary wage acts a screening mechanism, and serves to decrease adverse selection, a situation where workers self-select themselves into jobs for which they are unsuited. The key decision point for the employer is to set the probationary wage and the permanent wage at levels such that it encourages productive workers to apply and discourages unproductive workers from applying.

Because the exact future productivity of applicants is not known to the firm at the time of hiring, firms must make estimates of this productivity. Some applicants will inevitably present more 'risk' than others. For example, assume there are two candidates, A and B, and that both have an expected mean productivity of 100. Candidate A, however, is riskier, in that they have a higher standard deviation in productivity – assume they have a 50 percent chance of having a productivity of 150, and a 50 percent chance of having a productivity of 50. Candidate B, on the other hand, is a 'safe' choice, whose productivity is known with certainty to be 100. Lazear

and Oyer (2012) argue that organizations may prefer the riskier candidate, in that they provide a greater option value to the firm – if the worker turns out to be a high performer, the organization is better-off than if they had hired the safe candidate, but if the worker turns out to be a low-performer then they are terminated and the organization does not bear the costs of this low performance.

This advantage to hiring riskier workers is lessened to the extent that organizations must engage in long-term employment contracts, and to the extent that termination and severance costs are incurred. This potential preference for riskier workers often translates into a preference for younger and/or inexperienced workers, since their future productivity is much more uncertain and difficult to predict compared to workers that have been in the labor force for many years – in effect, younger workers have a greater potential 'upside'. Risky workers who are younger are particularly valuable to the firm, since the firm can potentially benefit from the upside for a longer period of time, compared to workers who are closer to the end of their careers. The value to the firm of a risky worker is also a function of the length of time the firm can retain the employee at below-market costs. If, for example, workers that turn out to be high performers are quickly bid-away in the market, firms will be less able to capture the benefits of the worker's upside. This possibility is lessened where there is asymmetric information about the worker's abilities – i.e. the current employer is more aware of the high-quality of the worker than are other firms – and/or where the worker's productivity is heavily firm-specific, making the worker more valuable to the current employer than to other firms.

Training and Development

While the recruiting process is a means to evaluate a worker's potential productivity at the time of hiring, that productivity is not necessarily static or fixed, but can be enhanced after the worker enters the firm. In HR management terms, this process is known as training and development (T&D). Analyzing T&D decisions within firms involves applying human capital theory.

There are two types of training – general and firm-specific. General training provides the employee with skills that can be used with any employer. In contrast, firm-specific training only increases the worker's productivity with their current firm, not at any other firms. In reality, most training programs fall somewhere between the two extremes. General training might involve, for example, teaching employees to use basic computer software packages, like Excel or Word, the types of skills that they can use

with a wide variety of employers. An example of firm-specific training would be where an employer teaches its employees to use accounting software that is uniquely customized to that particular organization. The skills learnt using this software would typically be non-transferable to another employer. An example of a somewhat less-tangible form of firm-specific training occurs when the firm educates its employees in the norms, values, and culture of the organization. This type of training is intended to increase the social cohesiveness of the workgroup, and convey a set of behavioral expectations on the employee.

The cost of acquiring general skills is paid for by the individual. If a firm were to pay, any employee that left the firm could fully transfer those skills to the new employer. Thus, there is no incentive for the firm to pay – its optimal choice is to wait for other firms to pay. However, all firms have this same incentive structure, so none will pay.

When firms provide workers with general training, it raises the workers' productivity across all organizations, not just with the firm providing the training. Since the firm providing the training cannot capture the benefits of providing such training, it will force the worker to pay for such training, either directly through compensating the firm, or, more typically, through reduced wages. This is one reason why younger (i.e. inexperienced) workers are paid less than their older counterparts – part of the young worker's total compensation is recaptured by the firm to cover training costs incurred on the worker's behalf. In essence, workers receive lower wages early in their career, i.e., while they are in their training period, so that they can receive higher wages later in their career.

Firm-specific training increases the worker's productivity at their current employer, but not with other employers. While the firm providing the training benefits from the increased productivity of the worker, the firm has an incentive to share some of these benefits (in the form of salary increases) of this increased productivity with the worker; otherwise the worker is indifferent between remaining with the current employer and moving to another employer. If the firm shares some of these benefits, the worker can earn more with the current employer than with other employers, and thus has an incentive to not move. However, the firm will not share all the benefits; otherwise, it would have no incentive to train workers. Thus, workers will be paid more than they could earn elsewhere, but less than their value to the firm.

With firm-specific training, where both the firm and worker share the benefits of increased productivity, both parties have an incentive to continue the employment relationship. Firms are better off because they are able to capture some of the benefits of the worker's increased productivity; correspondingly, workers have an incentive to remain with the firm

because they can earn a higher wage with the current employer than with other employers. The manifestation of firm-specific training is reduced turnover – both the firm and the worker have an incentive to maintain the relationship. There is also a related age factor at work. Since younger workers have gained less firm-specific human capital than older workers, the former are more likely to switch jobs than the latter.

Of course, turnover can still occur amongst longer-tenured employees. This turnover reflects a decrease in the quality of the employer-employee match, to the point where one or the other party seeks to end the relationship. For example, changes in management, production processes, or co-worker groups might all lead an employee to seek a different position. Turnover can be costly to the organization, particularly for positions that require high amounts of firm-specific knowledge – whether it be knowledge of production processes, management structures, cultural norms, etc. This firm-specific knowledge is usually only developed over a considerable length of time, and, by definition, cannot simply be purchased in external labor markets, making it difficult for organizations to quickly and seamlessly replace a departed employee.

As workers get older, the amount of T&D that firms devote to them will decline. This occurs for two reasons. First, T&D exhibits diminishing returns – the more T&D a worker has, the lower the marginal gains from providing more training. Second, the older the employee, the fewer the number of years the firm will have to recapture its investment, and the less willing it will be to invest in T&D for the worker.

Empirically, the relative importance of general skills versus firm-specific skills can be ascertained by examining the salary returns to 'experience' versus 'tenure'; experience measures an individual's total length of time in the labor market, whereas tenure measures the individual's time with the firm. The greater the returns to tenure, relative to experience, the more important is firm-specific training relative to general training.

Incentives, Compensation, and Performance

Personnel economics models the employee-employer relationship as an agency problem. The firm's goal is to maximize worker output; this output, however, is based not only on the skills/abilities the worker possesses, but also on the worker's effort level. Since effort is costly to the worker, the worker may have a tendency to provide less effort than is optimal for the firm. In other words, workers have a tendency to shirk. One way the firm can enforce effort by the worker is to directly monitor and control the worker's behavior. Another way is to create a monetary incentive structure that aligns the interests of the worker with the interests of the firm.

In this regard, so-called piece-rate compensation schemes directly connect output with pay – the worker gets paid per unit of output they produce. An example of such a scheme might be a sales job that is 100 percent commission-based. For piece-rate compensation to be most effective, a worker's output must be both measureable, and be largely independent of the efforts of co-workers. However, most jobs in today's modern economy do not meet this criteria – jobs, particularly those requiring higher skills, tend not to have easily measureable output; relatedly, the nature of work processes in many organizations requires employees to work as a team, meaning that one person's output, even if measureable, is a function of factors beyond simply that person's individual effort. As a result, most workers today are salaried, where their compensation is set in advance, and where it is independent of their current-period output – in other words, compensation is input-based, with the input being the worker's time.

The design of compensation structures can also be related to recruiting issues. For example, highly productive workers might be attracted to firms that utilize a piece-rate compensation scheme (where such high productivity is directly rewarded), whereas low-productivity workers may be attracted to firms that employ fixed-rate compensation schemes. Thus, piece-rate firms may be more productive, not necessarily because piece-rates encourage workers to give greater effort, but because piece-rate firms attract higher-quality workers.

When workers are on fixed-rate (i.e. input-based) pay schemes, different mechanisms must be used to discourage shirking. One such mechanism can be found in the overall pay structure of typical firms, where workers with more seniority get paid more. This positively sloped age-earnings profile can act as a motivator, both to younger workers and older workers. For younger workers, they have an incentive to reduce their shirking in the present, in hopes that they will retain their employment with the firm and enjoy advantages later in their career, where their pay will exceed their productivity. For older workers, they are already enjoying such benefits (of pay exceeding productivity), and thus have an incentive to minimize their shirking so as to be able to maintain these benefits.

Firms and workers will also sometimes enter into longer-term employment contracts. These contracts will specify a predetermined level of compensation for a predetermined time period. They create a potential disconnect between the employee's productivity and what the employee gets paid. The contract sets compensation based on the worker's expected productivity over the life of the contract; however, this productivity may deviate from expectations, due not only to the effort put forth by the worker (i.e. the worker may shirk now that they have a long-term contract), but also because of luck, or because of macro factors beyond the worker's

control. By providing a contract, the firm is essentially guaranteeing the worker a set payment regardless of productivity, and hence is reducing the worker's risk. In return, workers are likely to accept a salary penalty – i.e. they will agree to a compensation level that is lower than what they could expect to earn by continuously contracting on the spot market.

The Organization of Work: Work Teams and the Role of Management

Firms often seek to utilize work teams to accomplish production tasks. In designing and constructing these teams, the firm's goal is always to maximize team productivity. The specific nature of work teams differs greatly across organizations depending on the production technology of the firm in question. Sometimes, a work group may be referred to as a 'team', but in reality the individual workers will have very little interaction with each other. An example here might be a team of salespeople in a furniture store – each salesperson can perform their job with little to no interaction or support from co-workers. In this instance, the total production of the sales team will simply be additive across the various salespersons.

However, in more complex production technologies, the productivity of the team is dependent not only on the sum of the skills of the individual workers, but on how these workers interact with each other – the greater the complementarities across workers, the greater the team output. Firms will often try to mix workers – whether by age, experience, skills, nationality, cultures, etc. – to enhance the team's collection of capabilities. For example, older workers can help mentor and train younger workers – this is particularly important where the position relies heavily on firm-specific skills. Older workers are especially important in situations where their skill set remains current, and where factors like technological change do not quickly erode their productivity.

Also important to team production is the need to develop and maintain a cohesive work group. For example, workers in teams need to be able to effectively communicate with each other, and such communication can be affected by age differences, language differences, etc. Team members also need to have a sense of equity – i.e. the sense that all team members are contributing their share, and that team members are rewarded fairly. With the latter, one area that has been extensively studied is determining the effects on team production of relative pay inequality within the team. Two schools of thought have dominated the discussion – the so-called tournament model, and the fairness model. The tournament model, developed by Lazear and Rosen (1981), argues that large compensation differences within a work group – where a few workers at the top earn disproportionately high salaries for relatively small ability differences compared to their

peers – encourages maximum effort in that it will incentivize 'rank-and-file' workers to achieve these heights. In contrast, the fairness model argues that large pay disparities, particularly when not supported by comparable ability differences, are detrimental to team cohesiveness and performance.

The roles and effectiveness of management is another area that has been of interest to personnel economists. Managers serve a number of potential roles in overseeing workers, ranging from the technical (overseeing work-flow patterns, assigning workers to tasks, etc.) to the more interpersonal (motivator, coach/mentor), etc.

As Lazear et al. (2015) discuss, managers/leaders have the ability to impact the work performance of many other workers, so their impact can be multiplicative. The importance of leaders will vary across organizations and functional areas. In general, leaders will be more important in situations that are high-variance – that is, where the outcome will depend significantly on the decision made by the leader (Lazear, 2012). The best leaders are those that can utilize their judgment and experience to adopt the most effective course of action for their organization.

Empirically measuring the effectiveness of leaders can be difficult, because it generally requires detailed firm-level data. However, one common approach in the literature has been to examine the impacts of the sudden departure of CEOs due to their unexpected death. To the extent that such events impact the firm's stock price, it is taken as an indication of the extent to which the CEO's leadership impacted firm success.

1.2 THE UNIQUE CHARACTERISTICS OF THE EMPLOYMENT MARKET IN PROFESSIONAL SPORTS

The specifics of any particular employer-employee relationship are always a function of the market structures within which that relationship occurs. The sport industry in North America is particularly unique, in that the four major professional leagues (the 'Big 4') operate both as monopolists in the output market and monopsonists in the input market; leagues like the National Football League (NFL), for example, are the sole suppliers of elite-level professional football entertainment in the US, and are simultaneously the sole buyers of elite-level football talent.

While all of the Big 4 have been challenged by rival leagues at certain points in their histories, none of these rival leagues survived over the long term, and any competition in the output market was always temporary and fleeting, eventually reverting back to monopoly. In recent decades, the monopoly positions of the Big 4 now seem completely entrenched

and impenetrable, with no legitimate rival league existing in US sports since the collapse of the United States Football League (USFL) in 1985. Without the threat of rival leagues, the established leagues have used their monopoly position to limit the supply of their output, particularly as it relates to ensuring that the number of franchises in the league is kept at an artificially low level.

This scarcity of franchises means that the number of jobs (i.e. roster spots) is also scarce, and leads to a tournament-style process in the labor market whereby large numbers of prospective employees compete to secure one of the relatively few positions in the major pro league. For example, the NBA employs only about 390 players (30 teams with 13 players each) in any given season, but thousands of players compete each year at the Division I college level and in European leagues. Competition is intense, as small differences in performance can result in large differences in compensation. For example, consider two basketball players of almost equal ability – one earns the final roster spot on a National Basketball Association (NBA) team and makes the NBA minimum salary of over $800,000 per year, while the other gets sent to the D-League, the NBA's developmental league, where salaries average about $25,000 per year.

This windfall payoff that goes to those reaching the elite level incentivizes aspiring professional players to invest heavily – both monetarily and in terms of time commitments – in the development of their (athletic) human capital. However, working counter to this incentive is the fact that the athletic skills needed to reach the elite levels of a sport are almost completely non-transferable outside of the sport sub-sector in which they play. For example, the unique set of skills needed to play quarterback in the NFL are not adaptable in any meaningful way to jobs outside of sport, nor, for that matter, are they even adaptable to other sports, like hockey or basketball. This non-transferability increases the risk of investing heavily in these skills, and introduces an all-or-nothing type of outcome – the player either reaches the elite-level league and earns the commensurate windfall payoffs, or does not, and hence gets no return on their investment of time and money.

The lack of transferability of athletic skills to other jobs, combined with the tournament-style system within sport labor markets, also means that those players that do reach the elite leagues in their sport will earn considerable economic rents – in other words, a player's next-best alternative to their job in the elite league will pay a much lower salary, usually by several orders of magnitude. In unregulated or non-unionized environments, employers (i.e. team owners) could potentially take advantage of such a situation in negotiations, knowing that players could take salary cuts and still earn much more than in alternative (non-sport) occupations.

A Brief History of the Employer-Employee Relationship in Sports

The unique nature of the sport labor market described above inherently puts players at a natural disadvantage when dealing with team owners. For players, the lack of transferability of their athletic skills to other types of employment, the potential economic rents they can earn in the sport industry, combined with the monopsony power of team owners, all serve to drastically reduce their bargaining power

In the early decades of professional sport in North America, owners used their superior bargaining position to unilaterally implement several institutional mechanisms to even further reduce the bargaining power of players. These mechanisms ensured that individual franchises within a league did not compete against each other for players. Two mechanisms were particularly important in this regard: (i) the player draft, for new players entering the league, and (ii) the reserve clause, for veteran players already established in the league.

With the draft, teams sequentially select from a group of prospective incoming (amateur) players; once a team selects a player, that team holds exclusive negotiating rights to that player. Draft systems take away from the player any ability to decide the team for which they will play. The NFL was the first of the Big 4 to implement a draft, in 1936; the NBA followed suit in 1947, with the National Hockey League (NHL) and Major League Baseball (MLB) implementing their drafts in 1963 and 1965, respectively.

While the draft system removed the bargaining power of incoming players, the bargaining power of established (i.e. veteran) players was removed by the so-called 'reserve clause'. The reserve clause was first employed in 1884, in baseball, and essentially bound players to their original teams, unless the team decided to trade or release the player. When a player's contract (typically, one year) expired, the reserve clause provided the player's team the right to unilaterally renew (usually with a small salary increase) that contract for another year. This gave a team perpetual control over the player, and prevented any voluntary player movement within a league. Versions of the reserve clause were ultimately adopted in basketball, hockey, and football.

Team owners took a very paternalistic approach to players, promoting the notion that owners and players were a family, and assuring players that they would be 'looked after' and that owners always had their best interests in mind. Simultaneously, owners never missed an opportunity to remind players that they were privileged to be able to play professional sports, and that they earned much higher incomes than most other Americans.

This one-sided employer-employee relationship ultimately provided the impetus for players to unionize. During the 1950s, all of the Big 4 leagues

began to see their players become more vocal in their discontent and to begin to mobilize into collective action; by the mid-1960s, players' associations had formed in all four leagues. However, in that era the players' associations were 'company unions', in that they were largely controlled by the owners; the players' associations were a vehicle where players could air grievances, but within a framework where owners could still control and contain the discussion.

The relatively benign nature of players' associations changed abruptly in 1966 with the appointment of Marvin Miller as Executive Director of the Major League Baseball Players' Association (MLBPA). Miller adopted a much more confrontational approach to owners, ultimately resulting in the first-ever strike in US professional sport, when MLB players walked out for 13 days at the start of the 1972 season. As part of the settlement of that dispute, players gained improved pensions, and also the right to binding arbitration for salary disputes with owners – the latter gain seemed relatively modest at the time, but it would soon turn out to be the foundation for the most radical transformation of employer-employee relations in baseball history.

December 23, 1975: The Beginning of the Modern Era in Player-Owner Relations

On the advice of Marvin Miller and the MLBPA, pitchers Andy Messersmith and Dave McNally elected not to sign contracts for the 1975 season, and instead chose to 'play out their option' on their 1974 contracts. The standard baseball player's contract of that era (almost all of which were for one year) contained an option clause, which allowed teams to unilaterally renew a player's contract for the following year, his so-called option year. Of course, without free agency, the player would still ultimately have to re-sign with the team once the option year was completed.

The MLBPA took the Messersmith and McNally cases to arbitration and argued that, by playing out their option year, the players should be declared free agents and thus should no longer be bound to their current teams, the Los Angeles Dodgers and Montreal Expos, respectively. On December 23, 1975, arbitrator Peter Seitz rendered his decision, and agreed with the players; Messersmith and McNally were suddenly free agents, as was almost every player in baseball.

Miller and the MLBPA eventually negotiated a deal with the owners whereby only those players with at least six years of Major League service would be eligible for free agency; while this seems counterintuitive in the face of such a landmark victory, Miller knew that it was critical to limit the supply of free agents on the market at any one time if player salaries were

to rise. Players' salaries did rise; in fact they skyrocketed. In the five years preceding free agency, salaries rose an average of 8.8 percent per year; in contrast, under the first five years of free agency, salaries rose an average of 26.3 percent per year. Work by Scully (1974) found that MLB players in the reserve clause era earned only about 20 percent of their marginal revenue products (MRPs).

In the 20 years subsequent to the Seitz decision, players in the other three major pro leagues also eventually gained free agency. The players' associations in these three leagues were not as strong as the MLBPA, and the battle to gain free agency was often long and difficult.

Rival Leagues

In the era prior to players winning free agency rights in their sports, the only instances where teams did not have full monopsony power over their players was when a rival league existed.

Rival leagues were particularly prominent in football – the NFL faced four different rival leagues in the 40 years following the Second World War. The most successful of these was the American Football League (AFL), which started play in 1960, and within six years negotiated a merger agreement with the NFL that saw all eight AFL teams join the NFL. The last rival league to challenge the NFL was the United States Football League (USFL), in the mid-1980s. The USFL had several high-profile owners and was established with much fanfare, but the league survived only three seasons, and failed to gain any type of merger with the NFL.

In basketball, the NBA began play in 1946 under the name Basketball Association of America (BAA), and was itself a rival league to the established National Basketball League (NBL). The BAA proved to be too much competition for the NBL, and the NBL was absorbed into the BAA in 1949, with the league becoming known as the NBA. The NBA faced serious competition for nine seasons beginning in 1967, with the formation of the American Basketball Association (ABA). The two leagues merged in 1976, with four ABA franchises absorbed into the NBA.

At about the same time, the NHL faced competition from a rival league in the form of the World Hockey Association (WHA). The WHA existed for seven years during the 1970s, and had four teams absorbed into the NHL when the two leagues merged in 1979.

Major League Baseball (MLB) is the only one of the Big 4 leagues not to face competition in the post-Second World War era. In fact, no rival leagues have existed in baseball in over a century – the last being the Federal League, which ceased operations in 1917. The lack of rival leagues in baseball is largely attributable to a 1922 Supreme Court decision

that granted baseball an exemption from antitrust law. This allowed baseball to potentially engage in activities that would otherwise be illegal to prevent the formation of rival leagues. For example, in the late 1950s the Continental League was formed to challenge MLB, but the upstart league never did play a game, in part because MLB suddenly decided to expand (after 60 years of not expanding) to some of the cities targeted by the Continental League, and in part because MLB had hundreds of players in its minor leagues who were perpetually tied to their MLB clubs, thus ensuring that potential rivals would be unable to access a sufficient number of quality players.

Rival leagues were created on the basis of opportunities in both the output and input market. In the output market, rival leagues would attempt to capitalize on the monopolistic tendencies of the established leagues to severely limit expansion, thus leaving many viable markets unserved. On the input side, the monopsonistic power of the established leagues resulted in players being highly underpaid, relative to their MRPs, and made players amenable to switching leagues to increase their salaries. Not coincidentally, no rival league has ever emerged in a sport after players in the established league gained free agency rights.

1.3 THE CURRENT ENVIRONMENT: COLLECTIVE BARGAINING AGREEMENTS AND THEIR IMPACTS ON PERSONNEL DECISIONS IN SPORT

Despite differences in specifics, the history of employer-employee relationships has been remarkably similar across all of the Big 4 leagues: in the earlier years, the complete monopsony control of players through the reserve clause, alleviated occasionally and temporarily by the formation of rival leagues (except in baseball); followed by the more modern era, with the development of stronger unions, the eventual demise of the reserve clause, and the corresponding absence of any further rival leagues. With these similar histories, it is not surprising that the *current* state and structure of employer-employee relations is also very similar across the Big 4.

In all four leagues, collective bargaining agreements (CBAs) form the foundation for player-owner relations. CBAs have an impact across a wide range of personnel functions – from recruiting, to pay and performance, to training and development. The discussion below examines some of the specific areas in which CBAs set the institutional rules for the player-owner relationship.

Roster Size and Make-Up

In non-sport businesses, one of the core personnel decisions that individual firms must make is to decide *how many* employees to hire. This decision is removed for sport firms, as roster sizes are strictly set at the league level and are identical across all teams in the league. In addition, given the structured nature of the game being played on the field, how teams use these specific rosters spots, in terms of how many workers are hired at each playing position, tends to vary only minimally across teams.

In hockey, for example, NHL active rosters are limited to 23 players at any given time, only 20 of which can actually 'dress' for any particular game. Of these 20, teams will almost always carry 12 forwards – divided equally amongst centers, right-wingers, and left-wingers – six defensemen, and two goaltenders. In football, NFL rosters are limited to 53 players – of these, 22 are 'starters', with 11 on offense and 11 on defense. Rosters in basketball are small – only 13 players, relatively evenly divided between front-court players and guards. In baseball, MLB teams will generally use their 25-player roster to carry four or five starting pitchers and eight or nine relief pitchers, with the remainder being position players.

What this means is that all personnel decisions that ultimately follow – who to hire, how much to pay, etc. – must be undertaken within these constraints of a fixed roster size. NBA teams, for example, need never consider the relative costs and benefits of expanding their rosters to 20 players, for such decisions are beyond the individual teams' purview.

Entry Rules: Player Drafts

The typical entry point for players into the Big 4 is through the draft system

The draft system allocates incoming players to specific teams. Drafts are generally conducted in some type of reverse order, whereby the poorest performing teams of the prior season receive the first selections. As such, leagues justify the draft on the basis that it contributes to competitive balance. However, there is considerable evidence, both theoretical and empirical, to question such claims. First, the invariance principle states that resources – in this case, players – will flow to their highest-value use, regardless of who initially owns the resource. This suggests that, while the draft may initially allocate the best incoming players to the poorest performing teams (often small-market teams), these players will eventually find their way to the large-market teams, meaning that any competitive balance benefits of the draft are, at best, temporary and fleeting. Empirical evidence has generally supported these theoretical predictions, and has uncovered little to no evidence that drafts improve competitive balance.

To economists, the more plausible explanation as to why all four leagues have drafts is that drafts limit the bargaining power of incoming players. The draft system creates a monopsony situation, whereby players can only negotiate with the club that drafted them. This causes the salaries of incoming players to be lower than under an open bidding system. However, unlike the pre-free agency era, drafts can now only bind incoming players to particular teams for a specified length of time, and not in perpetuity.

The number of rounds in the draft varies by league – 40 in MLB, seven in both the NFL and NHL, and two in the NBA.

Player Mobility: Free Agency

In all of the Big 4 leagues, veteran players (i.e. those not on rookie/entry-level contracts) are able to gain free agency rights after reaching a pre-specified number of years of service in the league.

There is a critical distinction between what are typically referred to as 'restricted' versus 'unrestricted' free agents. Restricted free agents can sign with other clubs, but the signing club must compensate the player's former club, either through draft picks or other players (or, in more modern-day versions of being 'restricted', the player's current club has the right to match the offer and retain the player). Compensation requirements greatly reduce demand for the player – often to the point of zero – as other clubs are reluctant to lose players or draft picks. In contrast, with unrestricted free agency, the signing club does not compensate a player's former club, creating a much more active market for that player. Because the compensation requirements of restricted free agency are so harsh, most economists consider only unrestricted free agents to be 'true' free agents, in that they are the only players not facing a monopsonistic employer.

In the days before the owners and players formally negotiated the specifics of free agency through the CBA process, a few players would become restricted free agents by 'playing out their option', essentially challenging their team's perpetual claim on their services. The leagues were, of course, strongly opposed to such actions, and devised methods to limit voluntary player movement. In the NFL, for example, R. C. Owens of the San Francisco 49ers played out his option and signed with the Baltimore Colts. Pete Rozelle, the NFL commissioner at the time, intervened and unilaterally adopted a policy where the commissioner could arbitrarily award compensation (in the form of a player or players) to the team losing a free agent – the policy became known as the Rozelle Rule. The other three major pro leagues all followed suit and soon adopted some form of their own Rozelle Rule. Any instances in the 1960s and 1970s of players challenging the monopsony control of their teams were isolated and infrequent, and

often resulted in long court battles. As players' associations became powerful, free agency provisions began to become codified into CBAs.

Baseball was the first sport of the Big 4 where players gained true free agency, which was the result of the aforementioned decision of arbitrator Peter Seitz in December 1975. Following the decision, baseball owners and the MLBPA quickly agreed on a set of rules that would govern free agency. Starting for the 1976 season, all players that had accrued six years of Major League service were eligible to become unrestricted (i.e. no compensation) free agents. Remarkably, this six-year time frame to free agency still holds today, over 40 years later. In addition, players with three years of service can now file for salary arbitration, providing them some degree of bargaining power, even without full free agency rights.

Currently in the NFL, players can become unrestricted free agents after four years of NFL service. However, unlike baseball, football players have earned free agency rights only relatively recently. The NFL's first foray into free agency occurred in 1989 with the so-called 'Plan B' free agency system. Plan B was implemented unilaterally by the NFL (i.e. rather than being negotiated with the NFL Players' Association (NFLPA)). The 1987 NFL players' strike had ended disastrously for the NFLPA, with the NFL bringing in replacement players, forcing the regular players back to work without a settlement. In an apparent effort to avoid antitrust issues, the NFL decided to provide players with limited free agency, whereby each team 'protected' 37 players who were not eligible for free agency, but all others were free agents when their contract expired. Of course, the protected players were those who were of the highest quality, so Plan B only benefited the more marginal NFL players. This changed with the Freeman McNeil court case, where Plan B free agency was ruled a violation of antitrust laws, ultimately paving the way for NFL players to acquire full free agency rights in 1994.

With the other two leagues, NBA players won unrestricted (i.e. no compensation) free agency in 1988. Current NBA players generally acquire unrestricted free agency rights after completing their second NBA contract. Hockey was the last of the Big 4 sports where players gained unrestricted free agency. Currently, NHL players achieve unrestricted free agency after they have reached 27 years of age, or have played in the league for at least seven years.

Team Payrolls and Individual Player Salaries

Three of the four major professional leagues have some form of a salary cap, with MLB being the only league without a cap. The term 'salary' cap is somewhat of a misnomer because generally what are capped are team

payrolls, rather than the salaries of individual players (although the NBA CBA does restrict individual salaries for some types of players).

Both the NHL and NFL have hard caps, while the NBA has a soft cap. The difference is that hard caps provide a strict and unyielding upper limit on team payrolls; a limit that cannot be exceed under any circumstances. A soft cap also specifies a numerical upper-limit on team payrolls, but there are several allowable ways in which team can exceed this limit.

Luxury taxes are different from salary caps. With luxury taxes, an upper-limit payroll threshold is established and teams that exceed that threshold must pay a tax on the amount by which they exceed the threshold. Both the NBA and MLB have a luxury tax system.

The Case of European Soccer

The above discussion has focused on the Big 4 leagues in North America, simply because in these leagues much of the employer-employee relationship is regulated by CBAs. In contrast, European soccer is a more open and competitive market, and functions much closer to the way most non-sport firms would operate.

In Europe, the top domestic leagues compete with each other for players, drastically reducing the monopsony power of individual clubs or leagues. While Europe had its own version of the reserve clause at one time, the 1995 'Bosman' ruling essentially granted soccer players the same free agency rights that North American players first gained with baseball in 1975.

Players in European soccer leagues do not collectively bargain with their employers over monetary issues like salary or payroll caps. Furthermore, there are no roster limitations, so clubs can determine unilaterally the size of their squad.

In many ways, we would expect European soccer, rather than North American leagues, to be able to provide better insights into the non-sport world. However, a limitation for researchers is that individual performance in soccer is notoriously difficult to measure; in addition, salary data for individual players has generally not been publicly available. Thus, any potential gains to researchers from studying a more unregulated market have often been nullified by these measurement and data restrictions.

1.4 APPLICATIONS AND IMPLICATIONS FOR SPORT

As Oyer and Schaefer (2010) note, strategic management researchers ask whether persistent differences in performance across firms are (at least

partially) attributable to differences in HR practices. Applied to sport, this question becomes whether certain clubs in a league can consistently outperform its peers (either on the field and/or in terms of profitability) because of superior HR practices. Given the highly structured nature of the closed North American leagues – with roster limitations, the draft system, and salary caps – individual clubs have fewer choice variables on which they can differentiate themselves, compared to, say, European soccer clubs, or to firms in non-sport industries.

In thinking about applying personnel economics to sport, many broad observations are possible, and several questions arise that are important to ponder. Below are a few summary thoughts in each of the functional areas of HR management.

Recruiting and Hiring

- In non-sport industries, firms often use higher pay than their competitors to attract better quality applicants. However, in the Big 4 leagues, entry-level (i.e. rookie) contracts are generally dictated by CBAs, and even salaries for veteran free agents will often be constrained by salary caps.
- How many resources should clubs devote to the screening of entry-level players? Most teams in a league have the same public information about prospective players, sometimes through league-wide screening processes like the NFL draft 'combine', or the NHL's Central Scouting rankings of draft-eligible players. Individual clubs are able to gain a competitive advantage over other clubs only if they have consistently better private information than others, which can only be gained by superior scouting techniques.
- Even if a club were to consistently draft better players, two questions arise. First, how costly was it to achieve the better draft results? Second, how long can the player be retained by the club at a below-market wage? With the latter, free agency provisions generally allow players to leave their drafting clubs after four to seven years, depending on the league, meaning that any economic rents achievable through superior drafting methods are relatively fleeting.
- Is drafting 'risky' players an effective HR strategy? It would seem that it could be, in that players can be easily released if they fail to perform to expectations – there is little to no protection for players in this regard. Risky players would seem to be more likely found when clubs go beyond their usual geographic recruiting territories (as in the early days of Europeans entering the NHL and NBA, where these players were higher risk because of the lack of reliable information on their

talent levels) or when they recruit college players who have played in the lower divisions (say, Division II or III football players). However, despite the potential benefits of recruiting riskier players, agency issues may lessen the likelihood of this occurring. In preserving their own jobs, general managers may be very sensitive to not drafting players – particularly with early picks – who have a higher probability of being 'busts'. Their own survival sense may incentivize them to 'play it safe'.

- With veteran free agents, as opposed to entry-level players, the issues are different. Issues of asymmetric information, adverse selection, and the winner's curse all increase the risk to teams when signing free agents.

Training and Development

Sport clubs must determine their investment level in training and development activities. In North America, these activities typically occur within the club's minor league teams, while in Europe they occur within the club's youth training academy. The NHL and MLB have long maintained very extensive minor league systems to house players who have been drafted and signed by the club. Player development in these sports, particularly baseball, can be lengthy, with most 18-year-olds, for example, not ready to play at the top level. In Europe, the top soccer clubs all maintain youth academies whereby they bring players into the club at a young age, so the clubs have control and influence over a player's development for much longer than teams in North America sports.

The critical question here is the extent to which a player's professional prospects with the club are influenced by their skills related to the sport itself, versus their club-specific knowledge and skills. The former are transferable to other competitor clubs, while the latter are not.

While teams always talk about seeking incoming players that are a good 'fit' with the club's culture, current players, and/or coaching style, all of these factors can change abruptly, particularly the latter, as coaches are hired and dismissed with great frequency. When that occurs, is the player that was drafted, or was taken into the youth academy, no longer as valuable to the club as they once were? Casual observation would suggest that while club-specific factors are not necessarily unimportant in some situations, the fact that most veteran players switch clubs without significant performance changes would suggest that it is the general sport-related skills, rather than any club-specific match, that are the most important factors in their personal performance.

To the extent, then, that this is the case – i.e. that the club-specific match is secondary to the sport-related skills – this would suggest that

development systems, like the minor leagues in North America and the training academies in Europe, are more valuable to teams as long-term screening mechanisms (i.e. probationary periods) than as knowledge-impartment mechanisms. This is certainly a fruitful area for further research.

Performance, Pay, and Incentives

The output of a professional athlete is generally much easier to measure than it is for workers in non-sport firms. However, this ability to measure player output varies greatly by sport – where on-field teammate interaction is low (baseball), it is easier to measure individual contributions, compared to sports where interaction is high (football, hockey, soccer). Each of these sports has now been subject to the analytics revolution – the attempt to better measure the performance of players through the use of advanced technological tools that more effectively capture on-field activities, and through using this information to create new, 'advanced', performance metrics. This (presumed) increased ability to better measure player performance will continue to encourage clubs to find market inefficiencies – i.e. to find disconnects between player value and player pay. Exploiting these inefficiencies is profit-generating for the clubs, at least until other clubs gain knowledge of the inefficiency and bid it away in the market.

The better the understanding a club has about true player performance, the more likely it is to make effective decisions in the free agent market. For example, understanding age-performance profiles lessens the chances of clubs offering long-term contracts to players who are 'past their prime' and whose performance is likely to decline very soon.

Directly attaching pay to performance at the micro level is difficult in the sport industry, in that 'performance bonuses' are not a major component of players' compensation packages. In other words, to use broader HR terms, most pay is base pay, not piece-rate pay. There are at least two important reasons for this. First, players' associations are generally opposed to performance pay, as are most unions in the non-sport world. Second, the clubs themselves must be careful here, since performance bonuses, by their very nature, reward individual players, and may have negative impacts on teamwork and cohesive behavior.

The Roles of Teammate Effects and Management

In sports that require high interaction among teammates, like soccer, hockey, and basketball, team output is much more than the sum of the individual talents of the players. The most effective teams are those that,

for a given level of individual talent, generate the most output. Clubs can gain a competitive advantage over other teams in the league by essentially being better at team-building – in other words, finding the right mix of players. Several questions arise here:

- How large is this team effect? To what degree can finding the right combination of players overcome individual talent deficiencies?
- All else equal, are teams more effective if they are homogenous or if they are heterogeneous? Teams can be heterogeneous along many dimensions – age, experience, geographic origin, race, language, national origin, etc. For example, holding average age of the team constant, is it better to construct a team of primarily mid-career players, or a team where there is a mix of younger and older players? As another example, are players better when they have teammates that are similar to them in terms of geographic origin, or language, or culture? Or, does the effect work the other way? Or, does it not matter either way?
- For a given level of talent, how important is coaching to the team's success? Coaches perform multiple roles – selecting the specific mix of players, training, motivation, game strategy, etc. Given that coaches are paid much less than average players in most professional sports leagues, does this then imply that the coach's role is much less significant to team success? It this is the case, then coaches should be relatively interchangeable with each other, without any discernible performance change in the team.
- As professional sports have become more sophisticated – with advances in training methods, in-game strategic decision-making, motivation techniques, etc. – has the requirements to be an effective coach changed? Are educated generalists more effective in today's game than more narrow specialists (for example, former players)?

Several of the questions raised in this section have been examined, to various degrees, in the literature, while others have received little attention to date. The chapters that follow in this book touch on several of these issues, and help to provide a larger context for the discussion of personnel economics in sport.

NOTE

1. Since much, if not most, of the intellectual foundations of personnel economics are attributable to Ed Lazear, this section relies heavily on his (and his co-authors') countless

contributions to the discipline over the past three decades. For more detail and discussion on each of the topics discussed in this section, see Lazear and Gibbs (2014).

REFERENCES

Lazear, E. (2012) 'Leadership: a personnel economics approach', *Labour Economics* **19** (1), 92–101.
Lazear, E. and M. Gibbs (2014) *Personnel Economics in Practice*, Hoboken, NJ: John Wiley & Sons.
Lazear, E. and P. Oyer (2012) 'Personnel economics', in Robert Gibbons and John Roberts (eds), *The Handbook of Organizational Economics*, Princeton, NJ: Princeton University Press, pp. 479–519.
Lazear, E. and S. Rosen (1981) 'Rank-order tournaments as optimum labor contracts', *Journal of Political Economy*, **89**, 841–864.
Lazear, E., K. Shaw and C. Stanton (2015) 'The value of bosses', *Journal of Labor Economics*, **33** (4), 823–861.
Oyer, Paul and S. Schaefer (2010) 'Personnel economics: hiring and incentives', in O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Amsterdam: Elsevier, pp. 1770–1823.
Scully, G. (1974) 'Pay and performance in Major League Baseball', *American Economic Review*, **64** (4), 915–930.
Spence, M. (1973) 'Job market signaling', *The Quarterly Journal of Economics*, **87** (3), 355–374.