1 Introduction

Ever since the first worker was employed, employers have concerned themselves with worker discipline. Workers agree to work but, having agreed, are unable or unwilling to supply their labour as reliably as the employer would wish. Ultimately, discipline can be enforced by sacking offending workers, but this is not necessarily the most profitable solution. Bryson (2007), for instance, observes that

> [Elizabethan] actors were subjected to rigorous contractual obligations, with graduated penalties for missing rehearsals, being drunk or tardy, failing to be 'ready apparelled' at the right moment, or – strikingly – for wearing any stage costumes outside the playhouse. Costumes were extremely valuable, so the fine was a decidedly whopping (and thus probably never imposed) £40. But even the most minor infractions, like tardiness, could cost an actor two days' pay.

On the other hand, Stone's (1950) account of work patterns in an Elizabethan coal mine between 1580 and 1582 indicates that efforts by the employer to control absence were non-existent:

> [T]he Sheffield accounts . . . offer a very different explanation than that of the ruthless employer sacking and hiring his workmen at will. Rather . . . the miner worked when and as long as he thought fit and the employer was obliged to content himself with methodically recording the rate of absenteeism. Involuntary absenteeism as shown in the accounts was extremely rare, though some of the unexplained short-time weeks may have been caused by illness. The only recorded illness was in the second week in August 1580, when production was halved, the marginal explanation being that the 'collyers were syck'. A fortnight later, the full output per shift was again being performed so that the malady could not have been serious. During the whole three and a half years only one industrial accident took place when one of the pickmen, Handcock, was hurt in the third week of June 1582 and was away for a total of seven weeks, no effort being made to replace him during this time. There can be no doubt, however, that in the vast majority of cases absenteeism at Sheffield was voluntary. One of the pickmen was continually taking a day or two off during

the week, thus breaking the regular production schedule of the pit. The bankman, as the foreman, was more responsible, but on one occasion at least all work had to stop for three days, because 'Copeley the barrower went forth'. Furthermore, the whole gang irregularly took itself off for celebrations, or for recuperation from such festivities. Sheffield fair day in November 1581 they took off. Again in May 1581, the entry runs: 'Monday after the fair day, they wrought not.' They took off the day before St James's Day, and Easter Saturday in different years, quite unexpectedly. Sometimes it is merely recorded 'the collyers was away besides' or some such phrase. All this was over and above the regular fixed holidays which they unfailingly allocated to themselves on the saints' days of the old pre-Reformation calendar. [. . .] In fact. . .they had between a month and five weeks' regular holiday every year – which is hardly what one has been led to expect of a defenceless working class freshly subjected to the full pressure of unbridled capitalism.

Taken with Sundays, this reduced the days in the year available for work to about 280, while normal interruptions due to other causes whittled down the figure still further. Damp in the mine, flooding and over-production all played their part, and in actual fact the days' work done were as follows:

1579 July–December, six months	83 days (flooding and damp)
1580 January–December, twelve months	256 days (12 days' flooding)
1581 January–December, twelve months	256½ days
1582 January–December, twelve months	227 days (24 days laid off)

Of course no payments were made when no work was done, and as a result the yearly earnings for a typical year 1580 were very small. The barrower and the two pickmen in continual employment each earned £6, and the bankman earned – officially and on his own reckoning – £9. The pickmen's and barrower's actual average weekly wage was therefore far below starvation level.

Although the accounts for maintenance and new mining show that the miners earned substantial sums from time to time doing extra work about the mine, yet even these do not suffice to compensate for the poverty of the annual wage. It seems therefore almost certain that economic factors as well as superstition drove the miners to this insistence upon what were by sixteenth-century standards wholly extravagant holidays. They must have been cottagers with casual employment on their own gardens or about the Earl's park. While there is no shred of evidence in the accounts to support this suggestion, it is difficult to see what other explanation is possible for the situation as it has been described.

The work contract in this Elizabethan coal mine was very simple. It seems to have specified six working days per week, except when certain specified religious holidays occurred. No sick pay was paid, and, despite the fact that the total pay was 'far below starvation level', voluntary absence was frequent, apparently organised and unpunished by any sort of disciplinary procedure. Stone does not report what the daily hours may have been, but it appears that subsistence was ensured by small-scale agriculture independently of this employment. Compare this with the typical worker in a developed economy at the start of the twenty-first century, whose single source of income is a five-day week; who will be paid during holiday leave; who will have some insurance against loss of income due to sickness; and whose absences are monitored and, if not satisfactorily explained, will trigger managerial action, usually disciplinary in nature. Furthermore, many employers have arrangements in place to enable absent workers' tasks to be covered in some way.

The study of worker absence over the last century or so has been dominated by the idea that absence is a problem of worker discipline. Many commentators take as given the proposition that low absence rates are good absence rates, while conceding that eliminating absence altogether is an impossibility. One theme of this book is that these ideas are misleading and, if they are used as the basis for human resources (HR) practice, can lead to policies that encourage inefficiency.

Since we are economists, our arguments use the tools of economics. These enable us to extend our purview beyond the existing literature's stress on absence as a problem in the psychological make-up or social circumstances of workers. We accept that these are important, but just as important, if firms' and governments' policies are to be well founded, are the interests of firms and governments. Taking these interests into account leads to more acute insights than are available when only the behaviour of workers is considered.

The Elizabethan coal mine gives an idea of what absence behaviour would be like without cover arrangements, managerial control, holiday entitlements and sick pay. Work still got done, but it was interrupted either by individual absence[1] or by coordinated group absence. It is also important to note one advantage of this arrangement from the employer's point of view: the mine could be closed when it was

[1] It would be interesting to know what their co-workers thought of Copeley the barrower when he 'went forth' and of the disruption caused by the pickman's absences. It is possible that they were able to reduce the cost of regular disruption by covering the pickman's work. Another possibility is that they were laid off, and could use their time more or less equally productively elsewhere.

unworkable, without any remuneration being due to disappointed workers. Under these work arrangements, disruptions could occur either because the capital was available unreliably, or because labour was available unreliably. Modern firms often insure workers against capital disruption, by paying them when machines fail, or negotiating terms for a layoff.

The coal mine described in Stone's article was a capital asset, which would remain unproductive unless it had labour input. The available work force clearly had outside activities, which they could exploit. Whatever the pickman was doing on the days when he failed to show up, he presumably thought he was devoting time to the most profitable activity available to him. The view of absence that we put forward in this book differs little from this. The theoretical models described here are an attempt to analyse the important considerations for firms confronted with workers whose attendance may be unreliable. We also examine the considerations surrounding the attendance decisions of workers, and particularly the effect of different contracts on those decisions.

While we ourselves are economists, and we use the methods of economics in our approach, we have attempted to make this book accessible to non-economists. The text has technical aspects, but throughout we have endeavoured to give sufficient non-technical explanation so that non-specialists should be able to follow the arguments. In particular, we believe that there are important messages in it for human resources practitioners, other managers and policy-makers. In particular, Chapter 8 pulls together the main messages of our argument for managers, and Chapter 9 attempts to outline how it might influence the thinking of policy-makers. We also hope that researchers in other disciplines will find food for thought in these pages, although we would not expect (nor do we think it desirable) that they would adopt our way of thinking about absenteeism unreservedly.

One of our main criteria is to provide a coherent framework within which to conduct analysis of these issues. This has taken a surprisingly long time to emerge. Early attempts to analyse absenteeism failed to be coherent at all. The difficulties can be illustrated with two early discussions of absence and alcohol by eminent economists.

In 1919 Irving Fisher (cited by Benge, 1923) took a normative view:

> Another force for prohibition is the force of industrialism and of the modern desire for efficiency both on the part of industries and on the part of

individuals. Industry is applying modern science. From two to four glasses of beer will reduce the output of typesetters by 8 per cent. These and other experiments demonstrate that we will increase, by enforcing prohibition, the economic productivity of this nation from 10 to 20 per cent, and will add to the national output of the US between 7 to 15 billion dollars' worth of product, every year, reckoned at the moderate level of prices. These forces, the ideals of work, the requirements of modern industrial competition, the findings of modern science and the ideals of morality in American life are the forces which have put over prohibition, and it must be on these forces that we shall depend to enforce prohibition.

In the same year Paul Douglas published the first survey of the absenteeism literature (Douglas, 1919). In discussing the relationship between alcohol and absenteeism, he says:

(l) *Liquor.* The influence of liquor in causing absenteeism cannot be accurately measured. The majority of employment managers, however, state that from their observation, drinking men are absent far more frequently than abstainers. It is also true that shipyards in dry states have somewhat better attendance records than those in wet states. Although the whole matter is one on which no absolute statement can be made, it seems reasonably certain, all other things being equal, that complete prohibition will bring with it a decided improvement in working attendance.

(m) *Wage income higher than standard of living*. When a workman receives more money than he wishes to spend or save, he will stop working and thus bring his income down to his standard of living. If . . . real wages increase faster than the standard of living, absenteeism necessarily results. This was the situation created in several war industries where wages increased faster than the wants of the worker. It was undoubtedly a factor in certain sections of the shipbuilding and munition industries. A higher standard of living, therefore, decreases absenteeism as indeed does an increase in prices. It should be clearly realized that the term 'higher standard of living' carries with it no ethical implications.

It might perhaps more accurately be called 'more expensive standard of living'. The new wants which go to form it may be vicious. The Hawaiian planters are said to prefer dissolute laborers, addicted to drinking and gambling, to sober men, because they work harder and more constantly to satisfy their wants. From the standpoint of human values, therefore, absenteeism with all its waste may be far more profitable than some additional wants. Those who believe that progress consists in the multiplication of wants would do well to consider the qualitative as well as the quantitative aspects of the standard of living.

One central point that Douglas seems to be making is that drunks are more unreliable (or more given to absenteeism) than more sober types. When he discusses 'standard of living', though, he argues (or, at least, relays a rumour about pineapple growers) that, because drunks require a larger income to maintain their habit, they are more reliable at work than sober types. Douglas's confusion here arises partly from the idea (which has now largely been jettisoned by labour economists) that 'standard of living' is somehow exogenous to the labour market. According to this view, people have a target income and stop working when they reach it[2] – an assumption that was made unnecessary by Robbins' (1930) exposition of the roots of labour supply in choice theory, which explains exactly how an increase in the wage rate can generate a lower supply of labour, when workers view their welfare as dependent on their total income and available non-work time. Robbins' argument, which has recently been clarified by Sorauren (2008), relies on the now commonplace prediction of choice theory that an increase in a wage rate has two effects. It changes wealth as well as relative values. While the latter change has a predictable effect, the former does not, so the total impact is unpredictable. This model is now a key part of every course in microeconomics, but in 1919, when Douglas was writing, it was certainly not a familiar part of economists' analyses of the labour market.

A second issue raised by Douglas's remarks is not so easily disposed of. It is claimed not only that 'dissolute workers. . .work harder', but also that they work 'more constantly'. The 'constancy' of work (or of other kinds of productive inputs) has not been the subject of much economic theorising. Indeed, we believe that one of the chief contributions of this book is to point out the inadequacy of deterministic models in explanations of absenteeism. Such models may have some power in explaining average effects, but they fail when asked to deal with questions associated with contracting between firms and workers, which must necessarily take into account the facts that illness is to be expected, that it has negative impacts on the worker as well as on the employer and that these impacts are at least partially ameliorable, by the use of substitute inputs or by insurance.

[2] This idea has resurfaced in the 'behavioural economics' literature. See, especially, Camerer *et al.* (1997) and the riposte by Farber (2005).

The above quotation by Douglas begins with the remark that '[t]he influence of liquor in causing absenteeism cannot be accurately measured'. Recent work using theoretical ideas and empirical techniques similar to those described in this book is beginning to make this remark seem like an untruth (Johannsson, Böckerman and Uutela, 2009). We believe that this is evidence of the progress that the use of economic concepts can make in the understanding of absence. In particular, this book argues that absenteeism cannot be fully understood without appeal to the idea of a market.

We refer to the decisions of employers as 'supply-side' decisions and of workers as 'demand-side' decisions. Workers demand absence from time to time, even though they are parties to a contract according to which they undertake to attend. The extent to which workers will do this depends not only on their contract, their sense of responsibility, commitment, honour, and other socio-psychological factors, but also on how they value time spent other than at work. This varies considerably, and often in an unpredictable way. Sickness, sick children, jury service, pastimes and a whole host of possibly unpredictable[3] events can change a worker's valuation of time spent at work relative to time spent elsewhere. This idea is a key part of the book. If a worker's valuation of time spent in different activities did not change, the problem of labour supply would be simple. A job offer specifying a particular work pattern would either be acceptable or not. If it were not acceptable, it would be rejected. If it were acceptable, there is no reason why the worker would not abide by it.[4] Uncertainty is an inalienable part of any explanation of absence rates.

The idea that there is a 'supply' of absence is harder to swallow. It becomes easier to accept, perhaps, once it is made clear that 'supply' is a technical term used by economists to mean not just an amount of something offered by a supplier – economists call that 'quantity

[3] Unpredictable, or simply unusual. One of us, John Treble, can remember being required to take time off from work so that his co-workers could watch the live television broadcast of Princess Anne's wedding. Although their relative valuation of time had changed so much that it would have been unwise for the manager to try to enforce their contract, Treble's had not changed at all, so he got a good deal out of it.

[4] There is an argument due to Reza (1975) and Dunn and Youngblood (1986) that workers sign a contract for more hours than they would ideally like, and then take absences to compensate for the difference. We do not find this argument very convincing for a number of reasons, which are discussed in Chapter 3.

supplied' – but the relationship between quantity supplied and price. When we speak of the 'supply of absence' in the following pages, we mean that employers will generally tolerate a particular level of absence among their workers. (They would rather not have to, but they do because they know that without such tolerance employing workers would be very expensive.) The extent to which they will do so depends on how costly absence is to them, and how costly it is for them to monitor and control it. A firm's 'supply of absence' is the relationship between different levels of absence and the marginal cost to the firm of each of them.

Sickness is a source of uncertainty for workers. An employer who can help them handle that uncertainty will generally be able to pay lower wages than one who cannot, since risk-averse workers will be prepared to accept lower wages in return for reduced uncertainty. The costs to employers of providing such services depend on many things, but one thing is sure: these costs will not be the same for all employers. A market supply curve will thus be generated.

It would be naïve, though, to claim that the market for absence can be studied in isolation. Apart from the obvious fact that absence is a labour market phenomenon and cannot be studied separately from the labour market as a whole, the argument in the last paragraph suggests that employers might also supply certain kinds of insurance, so that the markets for these kinds of insurance become entangled with the study of absence as well.

While absenteeism seems to have become a major concern for employers sometime around the end of the nineteenth century, it was not until the middle of the twentieth that the insurance aspects became matters of managerial and regulatory concern. The US debates have been described and discussed in detail by Murray (2007), although, rather frustratingly, he ends his narrative before the post-war adoption by some US states of state-administered insurance against loss of income due to sickness. In most European countries, systems similar to the one initiated in 1883 by Count Otto von Bismarck in Germany were adopted, and in many of them these remain the basis for modern arrangements. A striking exception is the United Kingdom, where the state withdrew from its administrative role during the 1980s, and from most of its regulatory role in the early 1990s.

What kind of protections can employers give to workers against the uncertain consequences of illness? There are essentially three: they can provide job security, they can provide full or partial replacement income, or they can provide insurance against the extra costs of illness. The extent to which employers provide these things varies enormously between different economies, and in many economies it is heavily influenced by government provision or regulation. In England and Wales, for instance, job protection is governed largely by case law (employers are expected to handle dismissals due to illness in a 'fair and reasonable' manner). Replacement income is practically unregulated, although it is subject to small statutory minimum amounts. Health care is provided by a national health service, funded largely from general taxation. In the United States, on the other hand, federal law prohibits the dismissal of workers for taking small amounts of absence for reasons related to their own or their family's health. Replacement income is not federally regulated, although five states have sick pay arrangements funded from hypothecated tax receipts. Health plans are often provided by employers, but provision is by no means universal.

The reasons for the widely different mixes of private as opposed to public provision of the two insurances mentioned in the last paragraph are not well understood. The debate about the public provision of health care is moribund in Britain, since a consensus exists in that country that the National Health Service, while imperfect, is better than the alternatives (or, at least, that sticking with it is better than switching to any alternative). The debate in the United States rages, since there is a widespread suspicion that the current arrangements are not sustainable, allied to an equally widespread and deep suspicion of government intervention in markets. The final chapter of this book is an attempt to open a debate about the provision of sick pay, on the basis of the claim that, if experience rating were to be adopted by firms providing it, the classic problems raised by asymmetric information in insurance markets need not arise.

The idea of absence as the outcome of a market process is central to this book. Although it is no more than an idea, it has important practical implications. In particular, it invites reconsideration of the idea of absenteeism as an avoidable cost. The total costs of absence are themselves complex, and very hard to measure. In Britain, the Confederation of British Industry (CBI) attempts to do this every year, and regularly comes up with a figure well in excess of £10 billion. However, these estimates can be taken seriously only as an indication of the order of magnitude of the total costs. Because it would be unwise for a firm to try to impose zero levels of absence, there is no suggestion that there is a free lunch worth £10 billion that can be consumed whenever managers can summon up the energy to do so. In fact, absence control is an expensive business. Even the unsophisticated Elizabethan way of keeping track of and recording worker attendance would have taken some time, which the foreman could have used doing more valuable things had his workforce not been so fickle. In modern developed countries, those same tasks involve the installation and maintenance of time-keeping equipment, setting up and maintaining computer-based records, calculating the effect of absence on pay, keeping the relevant government administrative departments informed and handling disputes when things go wrong with any of these things. One British firm that we have studied in detail employs one full-time clerical worker in each of its plants simply to handle the day-to-day administration of its absence records and their interface with the firm's pay systems. In addition to this, the attendance system needs to be maintained, and its use, development and application needs to be guided and supervised by managers.

One of our own studies (discussed in more detail in Chapter 6) has developed an alternative approach to the measurement of absence costs. It suggests that the balance between the gain from managerial effort put into absenteeism reduction by our sample of French firms and the costs of that effort is well struck. Apparently, there is little in the way of free lunches to be had for French managers, given the technology that they are operating and the regulatory framework within which they work. As far as we can tell, they do a good job of identifying and acting on opportunities for absence reduction that cost less than they return.

The supply side of our models is dominated by the idea that not all production processes are equally disrupted by absence. In particular, teamwork makes absences costly, because an absence affects the productivity of an entire team; similarly, capital use makes absences costly, because the equipment falls idle as well as the worker. Once again, the detail of these arguments is complex, and their appearance in the actual world of work is not easy to disentangle from other differences between firms. If team work is more expensive, it will be worthwhile for firms that cannot avoid it to make adjustments to hiring practices, pay and conditions in an effort to reduce the expense. This means that