Supervision and Monetary Incentives

STOCKHOLM SCHOOL OF ECONOMICS
EFI, THE ECONOMIC RESEARCH INSTITUTE

EFI Mission

EFI, the Economic Research Institute at the Stockholm School of Economics, is a scientific institution which works independently of economic, political and sectional interests. It conducts theoretical and empirical research in management and economic sciences, including selected related disciplines. The Institute encourages and assists in the publication and distribution of its research findings and is also involved in the doctoral education at the Stockholm School of Economics.

EFI selects its projects based on the need for theoretical or practical development of a research domain, on methodological interests, and on the generality of a problem.

Research Organization

The research activities are organized in nineteen Research Centers within eight Research Areas. Center Directors are professors at the Stockholm School of Economics.

ORGANIZATION AND MANAGEMENT	
Management and Organisation; (A)	Prof Sven-Erik Sjöstrand
Center for Ethics and Economics; (CEE)	Adj Prof Hans de Geer
Public Management; (F)	Prof Nils Brunsson
Information Management; (I)	Prof Mats Lundeberg
Man and Organisation; (PMO)	Acting Prof Jan Löwstedt
Industrial Production; (T)	Prof Christer Karlsson
ECONOMIC PSYCHOLOGY	
Center for Risk Research; (CFR)	Prof Lennart Sjöberg
Economic Psychology; (P)	Prof Lennart Sjöberg
MARKETING	
Center for Information and Communication Research; (CIC)	Adj Prof Bertil Thorngren
Center for Consumer Marketing; (CCM)	Associate Prof Magnus Söderlund
Marketing, Distribution and Industrial Dynamics; (D)	Prof Lars-Gunnar Mattsson
ACCOUNTING, CONTROL AND CORPORATE FINANCE	
Accounting and Managerial Finance; (B)	Prof Lars Östman
Managerial Economics; (C)	Prof Peter Jennergren
FINANCE	
Finance; (FI)	Prof Clas Bergström
ECONOMICS	
Center for Health Economics; (CHE)	Prof Bengt Jönsson
International Economics and Geography; (IEG)	Prof Mats Lundahl
Economics; (S)	Prof Lars Bergman
ECONOMICS STATISTICS	
Economic Statistics; (ES)	Prof Anders Westlund
LAW	
Law; (RV)	Prof Erik Nerep

Chairman of the Board: Prof Sven-Erik Sjöstrand. *Director:* Associate Prof Bo Sellstedt,

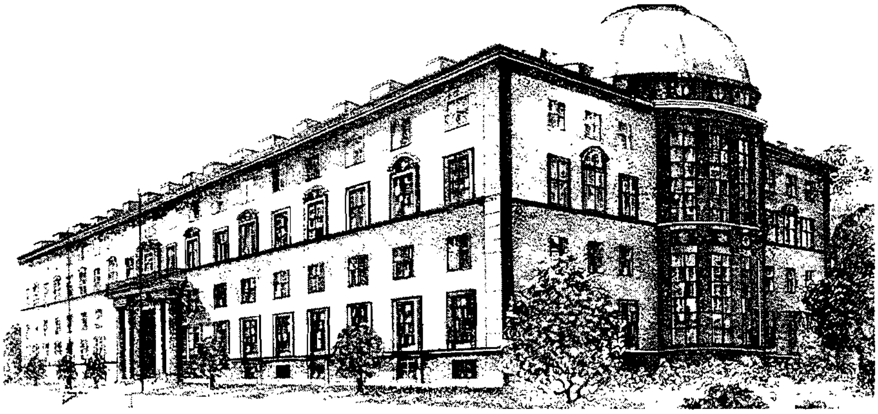
Adress

EFI, Box 6501, S-113 83 Stockholm, Sweden • Internet: www.hhs.se/efi/
Telephone: +46(0)8-736 90 00 • Fax: +46(0)8-31 62 70 • E-mail efi@hhs.se

SUPERVISION AND MONETARY INCENTIVES

Magnus Allgulin

Stockholm 1999



STOCKHOLM SCHOOL OF ECONOMICS
EFI, THE ECONOMIC RESEARCH INSTITUTE

Dissertation for the Degree of Doctor of Pilosophy, Ph.D.
Stockholm School of Economics, 1999

ISBN NR 91-7258-525-0

Keywords:
Monitoring, supervision, efficiency wages, incentive pay, implicit bond, entrance fee, quota, market, emission permits

Distributed by:
EFI, Stockholm School of Economics,
Box 6501, S-113 83 Stockholm, Sweden

Printed by:
HHS PrintCenter, Stockholm 1999

for Susanne

Contents

Preface

Fina lilla krummelur,
jag vill aldrig bliva stur.[1]

(Pretty little chililug,
I don't want to get bug.)

Chililug pills do not exist in real life, whatever anyone may think. Even as a child, when I first became acquainted with these magic yellow peas, I had my doubts because I knew that Pippi has a vivid imagination and that you should not believe everything she says. But the very idea of chililug pills fascinated me. As we know, grown-up people are very square beings, full of superstitions·and preconceived opinions. (They think that something terrible is going to happen if they happen to stick their knives in their mouths while they're eating, and things like that.) The insight that one day I would become a grown-up was frightening, and the thought of not just accepting this horrible metamorphosis appealed to me.

This is a critical motive for my decision to study for a Ph.D. This is because reading for a Ph.D. is synonymous with postponing the grown-up mentality, with the aim of resisting temptations to accept dogmatical truths. Something like a chililug pill. To work for a doctor's degree at the Stockholm School of Economics is probably an exceptionally good choice. Because here, it is as though everybody takes chililug pills all day long, and I am forever grateful for having been a part of this community during one period of my life. During this time I have had the opportunity to meet a large number of extraordinarily pleasant people.

First and foremost I must mention my advisor Tore Ellingsen. I cannot thank you enough for what you have meant to my academic process of maturing. First you were a teacher with such an ardent affection for microeconomics, it was impossible not to become absorbed by it. Then you were a responsible, diligent and scrupulous as well as encouraging and humorous thesis advisor. And finally, the very best coauthor there is. Thank you, Tore, for four highly inspiring and innovative years at the Stockholm School of Economics.

I further want to thank Jörgen W. Weibull and Lars Bergman for the enormous amount of effort they have put into providing a very stimulating environment for research at the department.

[1] Astrid Lindgren (1948): Pippi Långstrump i Söderhavet

You also helped me get the opportunity to locate part of the graduate program to Harvard University, for which I am very grateful.

My stay in the US really deserves a chapter of acknowledgments to itself, because so many people were involved in assisting me one way or another. Most of all, I am deeply grateful and touched by the cordial welcome I received from my hosts Christer and Brita Stendahl. You corresponded with me and tried to prepare me for my visit. You arranged for someone to come and meet me and my wife at Logan Airport. You lent us your condo the first couple of weeks while you were staying at your summer-house on Nantucket, etc, etc, etc. You also instructed Erik Berglöf at the Stockholm School of Economics to assist me if I needed any academic help. Thank you, Erik, for your generous offer of introducing me to your friends at Harvard. Whether you told Thomas Sjöström or he found out about my existence all by himself, I don't know, but as soon as I opened an e-mail account there, I received the message, "Let's get together and watch the Boston Bruins beat the Florida Panthers" from him. Marcus Asplund, Rickard Sandin, Lena Edlund, Rickard Friberg, Sara Johansson and Anders Paalzow who advised me on all kinds of practical matters and shared their own experiences from universities in the US also deserve my sincere gratitude. And finally, Kaj Martensen. Our long and fruitful discussions made every day at Harvard an enjoyable experience. Thank you ever so much.

The person who inspired me most to take on the task of pursuing a Ph.D. in Economics is definitely Sven-Olof Fridolfsson. Sven-Olof, David Sundén and I also worked very closely during the somewhat demanding first year of the Ph.D. program, and I would like to thank both of you for your help and great companionship during that time. Likewise, I would like to thank my dear friend Martin Hill with whom I have shared an office since the second year. You have discussed and helped me with every essay in this thesis. Our everyday conversations will probably be what I miss the most from my time as a Ph.D. student. You have truly made each day in the office something to look forward to, and I am immensely grateful for that.

Besides, thanks to the seminar participants in the Economic Theory-, the Industrial Organization- and the Environmental Economics group for comments and helpful discussions. I would further like to thank the Ph.D. students in the Micro II class 1995, 1996 and 1998. Apart from being such nice and tractable students, your candid and forthright questioning of established theories has been more valuable to me than you may know. In addition Kerstin Niklasson, Britt-Marie Eisler, Lisa Tilert, Majgund Fredholm and especially Pirjo Furtenbach and Azad

Saleh have played an important role in working out the many practical problems I have come across. Beyond this, I would like to express my sincere gratitude to my parents, my parents-in-law and the rest of my family for your never failing support and your ability to always be there for me.

Money also matters. Generous financial support from the Royal Swedish Academy of Sciences; C C Söderströms Donationfund, Handelsbanken's Research Foundations; Jan Wallander's and Tom Hedelius' Foundation for Social Research and Stockholm School of Economics; Louis Fraenckel's Grantfund is gratefully acknowledged.

Last, but definitely not least, Susanne. My sunshine. The best that has ever happened to me. You have been infinitely patient during the sometimes too time-consuming phases which the Ph.D. program has brought about. You have always encouraged me when I have lost heart. You inspire me with your happiness and spark of life and you have taught me to prize what is most important in life. You are my joy of living and my best friend. Thank you for your tremendous, unconditional love.

And now, as I write this very last sentence of my thesis, something overwhelming occurs to me:

Kära lilla krummelur,
jag har plötsligt blivit stur.

(Dear little chililug,
suddenly I've gotten bug.)

Stockholm, October 1999

Magnus Allgulin

1 INTRODUCTION AND SUMMARY

Introduction

"Isn't it funny what different implications efficiency wages and linear incentives have regarding monitoring and pay?" These are my advisor Tore Ellingsen's words a little more than three years ago. He referred to Shapiro and Stiglitz's (1984) result, that monitoring and pay are substitute instruments for motivating worker effort, and to Holmström and Milgrom's (1994) reversed result. "Why don't you write down a synthesis of these different mechanisms, explaining this peculiarity in a more general framework", he continued. Even though I didn't know Tore as well then as I do now, I had the feeling that he had given this question more than just a hasty thought. It also sounded like a clear and distinct problem to work with, so I decided to take the advice seriously.

My first very natural (and naive) idea was to proceed from Holmström and Milgrom's model and impose a limited liability restriction that had not previously been present, which was an important ingredient in the efficiency wage logic. I could never have imagined how this small perturbation literally made the model explode. I spent half a year, turning the model over, trying to make it small and attractive again, until Bengt Holmström advised me to throw away everything that I had done so far, and start all over again with a different approach. "Don't be sad", he said. "This is the most important lesson a Ph.D. student can learn. Most researchers *never* learn to let go of projects when they are not fruitful any more!" But he told me to stick to the fundamental idea and keep my aim high for a bit longer, since I was still at the start of my career: "If you insist on developing an existing theory, you should *remove* restrictions and try to present *more general* results. 90% of all theoretical papers are just perturbations of existing theories that the original authors had already thought of, but chose not to present."

Suddenly knowing more than most researchers, I dropped Holmström and Milgrom (1994), set my sights on Shapiro and

Stiglitz (1984) instead, and began my second attempt to attack the problem. Like a predator looking for prey, I was searching for a restriction to remove. Soon, I had made my choice: I would remove the simplifying restriction of discrete effort levels. This was an interesting approach because it would allow for a scale effect, which in turn ought to affect the correlation between monitoring and pay. "Brilliant", Tore replied when I presented my new angle and some results. A couple of weeks later he gave me an exquisite solution for the remaining Gordian knot, and the project was launched.

This is the background of the first essay, *Monitoring and Pay*, written with my advisor Tore Ellingsen. In this essay, we extend the standard shirking model of efficiency wages to a continuum of effort levels. It turns out that this generalization completely overturns previous intuitions. In particular, the characteristic feature of the earlier theory, that monitoring and pay are substitute instruments for motivating workers, no longer exists. This is of some importance since such a negative correlation is ordinarily used as the primary empirical test of the existence of efficiency wages.

Even if this surprisingly clear result took away one motive for performing the original task, that of developing a synthesis, the idea was still worth investigating. As a result, a common framework for efficiency wages and linear incentives is presented in the second essay, *Monitoring and Pay: General Results*, and it is shown that efficiency wages are strictly better for the employer than the use of linear incentives. Here, it is also demonstrated that the initial assumptions in the first essay can be made more general without altering the results.

The third essay, *Limited Liability and Dynamic Incentives*, is a dynamic version of the efficiency wage model presented in the first essay. It is influenced by the criticism that Shapiro and Stiglitz received from for example Bull (1985, 1987), that it is unrealistic to suppose that the firm and the worker do not know when the relationship is to end. Hence, the model has finitely many periods. Interestingly, once the wage is not forced to be stationary (by an assumption of infinitely many periods), the wage path will be very similar to the one that emerges from delayed payment/bonding contracts such as are found in Becker and Stigler (1974) and Lazear (1981). The major difference is that the worker rents associated with efficiency wages still exist, even if they are strongly diminished by the length of the relationship.

Finally, the last essay, *Do Market-Based Incentives Lower the Cost of Compliance?*, deals with policies for regulating polluting firms under imperfect monitoring. This is an example of the close relationship between the problem of setting wages and regulatory policies. It also shows how comprehensive the efficiency wage mechanism developed in the first essay is. In the same spirit, a regulatory agency uses both money and monitoring to persuade a polluting firm to comply with a regulatory standard. The framework in the second essay is used when comparing the optimal command and control policy with market-based incentives.

Thus, the thesis consists of three essays on labor economics, and one essay on environmental economics and government policy. These are two apparently very distinct areas but, as shown here, there is in fact a very strong connection between them. I would like to recommend the reader to read the chapters in succession (chronological order) even if each essay is self-contained. Following this advice, the reader will unquestionably understand and assimilate the efficiency wage model before taking on the regulatory policy in the last chapter, and I believe that the thesis may be easier digested this way.

The purpose of the remaining part of this first chapter is to provide the reader with a good insight into the thesis in very few pages so that even those who never have time to get past the introduction can understand and discuss the individual essays. So, whether you are satisfied with that or whether you try to make it to the end: enjoy reading it!

Summary of the essays

Essay 1

Monitoring and Pay Conventional wisdom says that monitoring and pay are substitute instruments for motivating workers: workers who are poorly monitored have to be well paid in order not to shirk, and conversely, workers who are closely monitored do not have to be particularly well paid. This version of the efficiency wage idea has a long and distinguished history. It is also the cornerstone of Bulow and Summers' (1986) theory of dual labor markets. In short, their theory says that there is a primary sector in which monitoring is difficult—hence working conditions are good and pay is high; and there is a secondary sector in which monitoring is easy—hence working conditions are poor and pay is low.

In this essay it is argued that the accuracy of monitoring and the level of pay are negatively related only under very restrictive assumptions. Notably, the above argument is correct when workers only have a choice between two effort levels (working and shirking), but incorrect when worker effort can be adjusted in a continuous fashion. Similarly, the argument is correct when the desirable level of worker effort is given exogenously, but incorrect when the desirable level of worker effort emerges as part of the solution to the firm's profit maximization problem.

The point is that monitoring costs affect not only the choice between the carrot (pay) and the stick (monitoring) in implementing a given level of effort, but also the level of effort itself. In general, there is a scale effect in addition to the substitution effect. For example, as monitoring becomes more costly, there will be a shift from monitoring toward wage incentives, just as Bulow and Summers claim, but there will also be a reduction in the desired effort level. Without further analysis, the total effect on the wage is therefore ambiguous. Indeed, as is shown, the wage is likely to go down, not up.

If primary sector jobs are jobs with high wages, the theory presented in this chapter says that monitoring is easier in the primary sector than in the secondary sector. The "technical" assumption that there are only two effort levels led Bulow and Summers (1986) to make exactly the opposite prediction. The leading example of a primary sector job offered by Bulow and Summers is Henry Ford's five-dollar day, the pay policy introduced by the Ford Motor Company in 1914, which represented almost twice the wage that similar workers could be expected to earn elsewhere. Reportedly, productivity increased by about fifty per cent, a feature which squares well with the model offered here. The key question is: Did the change in pay policy relate to changes in supervision technology, and if so how? The common view is that mechanization of production and standardization of tasks generally made it easier to deter shirking. According to this essay, then, the increase in pay could be due to easier monitoring.

Earlier theoretical literature has taken it for granted that workers are better off in firms where it is difficult to monitor and in firms which pay high wages. We show that neither of these propositions is true. Workers are best off in intermediate cases. In firms where monitoring is either very difficult or very easy, workers are paid close to their reservation wage. The difference is that when monitoring is easy, workers are required to work harder and therefore receive a higher wage. Indeed, the very highest wages will be paid by firms which monitor very accurately, and hence do not need to pay a wage premium.

The paper also identifies two serious problems associated with empirical implementation of the theory. First, the accuracy of monitoring is very hard to measure. Most studies proxy the accuracy of monitoring by the intensity of supervision, as measured by the number of supervisors per worker or the frequency of supervision. This is illegitimate. The number of supervisors may be high precisely because monitoring is difficult. Existing studies have essentially investigated the relationship between monitoring costs and pay. Unfortunately, this relationship is more complicated than the relationship between the monitoring level and pay. Here is a second problem. If there is sufficient variation in the underlying (unmeasured) parameters, there will be two possible wage levels for each level of monitoring costs. Thus, unless the effort level is properly controlled for, the empirical literature's regressions of wage premiums on monitoring costs are unjustified. But all is not lost. A more constructive contribution of our work is to show that it is still legitimate to run a regression of monitoring costs on wage premiums. In other words, existing data sets can still be of value in testing the theory if only regressions are run in reverse!

Thus, this chapter explains how a very natural generalization of the conventional shirking model completely overturns previous intuitions. The simple empirical prediction is: whenever the accuracy of monitoring is high, the level of pay should be high as well, and this is due to the existence of a scale effect which outweighs the substitution effect that earlier work has focussed on.

Essay 2

Monitoring and Pay: General Results In the previous essay it is calculated how monitoring and pay vary according to multiplicative shifts, and it is shown that they are positively correlated. This paper demonstrates the robustness of the result to a variety of generalizations. The main finding is that the result is true for any monotonic shift, not just the multiplicative.

Two extensions of the model are also investigated. The first is the introduction of an ex ante individual rationality constraint for the agent in addition to the previous ex post limited liability constraint. This will typically be of interest if unemployed workers bid over each other with entrance fee offers for a job opportunity. The implication is that a binding ex ante individual rationality constraint generalizes the result even further.

The second extension allows the principal to credibly use a mixed strategy when monitoring the agent. In that case, any concave monitoring cost function can be replaced by another monitoring cost function that is cheaper and linear. Hence, the monitoring technology used by the principal is always weakly convex, implying that even this extension will make the result more general.

Finally, a framework is constructed to compare the optimal non-linear incentive scheme solution in the paper to the best linear incentive scheme solution. The finding is that, for the principal, there is always a non-linear wage contract which strictly dominates any linear wage contract.

Essay 3

Limited Liability and Dynamic Incentives The objection against the shirking version of the efficiency wage model that most frequently recurs in the literature is that unemployed workers should offer to pay entrance fees for job opportunities, and such fees are very seldom observed. The best known paper which presents a shirking efficiency wage model is Shapiro and Stiglitz (1984). Indeed, Shapiro and Stiglitz were aware of this weakness and defended their model against the entrance fee criticism already in the original paper. They pointed out the two most obvious arguments against job purchases. The first is that workers may simply not be able to pay the entrance fee. The idea is that workers are wealth constrained due to imperfect capital markets. The second is that the resulting employment contract is not self enforcing. The firm would have an incentive to claim that a worker shirks, fire him once the fee is paid, and sell the employment opportunity again.

However, many economists have not been convinced by these explanations of the lack of entrance fee observations. Carmichael (1985) in his reply to Shapiro and Stiglitz and Carmichael (1990) argue that the presence of an imperfect capital market, or the moral hazard problem, is not sufficient to prevent the emergence of entrance fees. For example, the moral hazard problem can be solved by the introduction of a third party. Moreover, in the Shapiro and Stiglitz model, the firm and the worker do not know when the relationship is to end. Bull (1985, 1987) argues that this is unrealistic and continues to show that unemployment will not be sufficient to generate the surplus required for a self-enforcing contract in a model with a last period. As a consequence, the efficiency wage theory has lost advocates to a related theory, referred to as delayed payment/bonding contracts. In these contracts, firms initially pay wages below the alternative wage and later pay wages above the alternative wage, to discourage shirking when monitoring is imperfect, making sure that the worker's individual rationality constraint is binding. This way one avoids the entrance fee criticism since the present value of compensation is not altered from the first best full information level.

The main purpose of this paper is to show how a firm should adjust the level and timing of compensation in the best possible way. Incentives can be constructed by both efficiency wages and back-loaded compensation, and it turns out that a rational firm will use both of them. It will be demonstrated that in a self-enforcing efficiency wage model with a finite number of periods and wealth constrained workers, the shape of the optimal wage path may look very much like the one derived from a delayed payment/bonding contract, such as for example Lazear (1981) and Becker and Stigler (1974). It will typically be constituted by a high wage in the last period and a lower and stationary wage for all preceding periods. The worker rents associated with efficiency wages still exist, but they are strongly diminished by the length of the relationship. This result is perhaps most remarkable for not fully eliminating worker rents; Lazear (book) and others have claimed that delayed payment is a perfect substitute for entrance fees. A notable major difference is that it is optimal to pay the higher wage in the last period because of the efficiency wage mechanism, not because it compensates the worker for previous periods but because it yields the correct incentives in the last period.

Bonding contracts and entrance fees might in principle eliminate the remaining worker rents. However, both bonds and entrance fees are inferior means of extracting worker rents. Investment in firm-specific human capital solves the moral hazard problem more efficiently. Shapiro and Stiglitz briefly mention that when the cost of losing job-specific human capital is substantial, workers may have an incentive to exert effort even under conditions of full employment. That is true in this paper also. The point is that a firm-specific human capital investment increases the worker's liability and hence increases the firm's profit at the investing worker's expense. Unemployed workers or workers employed by other firms bid over each other for a job opportunity with higher education levels. This may go on until the hired worker is totally extracted. The reason why bids are in education levels and not in money (bond or entrance fee) is simply that it solves the moral hazard problem in the simplest possible way; if the employer fires the worker he will lose the firm-specific human capital investment together with the worker.

Essay 4

Do Market-Based Incentives Lower the Cost of Compliance? Ignoring monitoring and enforcement costs while designing a regulatory standard will most certainly lead to the implementation of an inferior policy. Either the regulatory agency will impose a too expensive policy, or it will not succeed in making firms comply with the standard. In fact, the consequences may be disastrous, especially in the latter case. This is a lesson both academics and policy makers have learned over the last few years. Consequently, monitoring and enforcement have been the focus in many recent theoretical and empirical studies, and regulatory agencies have read the proposed prescriptions carefully. The most obvious question they search an answer for is of course: What is the socially most desirable regulation? Or more precisely, what is preferable, a command and control policy or more market-based incentives?

This essay tries to answer that question, primarily with reference to any negative external effect on the environment, but the reasoning throughout the paper can easily be extended to embrace most kinds of negative external effects. The paper starts out by deriving the optimal command and control policy. This will be constituted by an emission quota, a penalty if the emission level exceeds the quota and a compliance fee otherwise. The firm has to be given correct incentives in order to obey the quota. The mechanism for this is first described in Becker's (1968) economic analysis of crime. The basic insight of that article is that potential criminals respond to both the probability of detection and the severity of punishment if detected and convicted. Thus, deterrence from criminal activity may be enhanced either by raising the penalty, or by increasing monitoring activities to raise the likelihood that the offender will be caught. That is true here also. On the other hand, here it will be assumed that there is an exogenous limit to how high a penalty is feasible, which will be binding. The reason may be political or due to a wealth constraint of the regulated firm. Instead, the net penalty, i.e. the difference between the penalty and a compliance fee, can be varied.

The first striking result is that the optimal compliance fee may be negative, i.e. some polluters should not carry any tax burden at all but should rather be subsidized. This follows from the fact that a smaller fee is optimally used as a complement to monitoring in order to create incentives, and consequently certain firms, which need strong incentives, are monitored very accurately and are imposed a negative fee in equilibrium.

The idea of a compliance subsidy is found earlier in the environmental economics literature: Downing and Kimball (1982) note that a cost subsidy will make a regulated firm more inclined to comply with a regulatory policy, because the firm has more to lose by not complying. The same message is put forward in for example Sullivan (1987), Fullerton and Kinnaman (1995) and Sigman (1998). A somewhat different cause for a subsidy is presented in Nowell and Shogren (1994), who study its effects on liability evasive activities.

The paper continues with finding the optimal emission-fee regulatory policy, and this policy is shown to be at least as good as the best market-based policy. Finally, the command and control and the emission fee policies are compared. This leads directly to a second very interesting result. Despite the fact that the major part of the environmental economics literature touts the use of market-based incentives, the finding here is the opposite: a certain emission target is always more costly to enforce if there exists a market for emission permits than if it does not. Thus the answer to the title is: No, market-based incentives do not lower the cost of compliance. The intuition behind the result is simply that a market restricts the regulatory agency to using an incentive scheme for the polluting firms that is linear in their emission levels, and that such a restriction makes the incentive scheme less powerful.

Sammanfattning på svenska

Uppsats 1

Monitoring and pay Den vedertagna uppfattningen är att övervakning och lön är substituerbara redskap för att motivera arbetare: Arbetare som är dåligt övervakade måste vara bra betalda för att inte maska i arbetet. Omvänt menar man att hårt övervakade arbetare inte behöver vara särskilt bra betalda. Denna version av effektivitetslöneidén har en lång och ansenlig historia. Den är också grundvalen i Bulow och Summers (1986) teori om dubbla arbetsmarknader. Sammanfattningsvis säger deras teori att det finns en primär sektor i vilken övervakning är svår - och därmed arbetsförhållandena är goda och lönen hög; och att det finns en sekundär sektor i vilken övervakning är lätt - och därmed arbetsförhållandena är dåliga och lönen låg.

I detta kapitel anförs det skäl för att noggrannheten i övervakning och lönenivån är negativt korrelerade under väldigt inskränkta förhållanden. Speciellt gäller att ovanstående resonemang är riktigt om arbetare bara kan välja mellan två prestationsnivåer, men felaktigt om prestationsnivån kan anpassas kontinuerligt. På liknande sätt är resonemanget riktigt när den önskvärda prestationsnivån är exogent given, men felaktigt om den önskvärda prestationsnivån uppstår som del av lösningen till ett företags vinstmaximerings problem.

Saken är den att övervakningskostnader inte bara påverkar valet mellan morot (lön) och piska (övervakning) för en given prestationsnivå, utan också prestationsnivån själv. Generellt gäller att det finns en skaleffekt utöver den vedertagna substitutionseffekten. Till exempel, om övervakning blir mer kostsam växlar man över från övervakning till löneincitament, precis som Bulow och Summers gör gällande, men det kommer också att resultera i en lägre önskad prestationsnivå. Utan grundligare analys är det därför svårt att säga hur lönen påverkas. Faktum är att det är sannolikare att lönen går ner än upp.

Om arbetstillfällen i den primära sektorn är arbeten med höga löner så säger teorin som presenteras i den här essän att övervakning är lättare i den primära sektorn än i den sekundära sektorn. Det "tekniska" antagandet att bara två prestationsnivåer är möjliga ledde Bulow och Summers (1986) till precis den omvända förutsägelsen. Det ledande exemplet av ett primärsektorsarbete som framförs av Bulow och Summers är Henry Fords "five dollar day", lönepolicyn som infördes av the Ford Motor Company 1914, vilken motsvarade nästan dubbelt så mycket som liknande arbetare kunde tjäna någon annan stans. Det påstås att produktiviteten ökade med ungefär femtio procent, något som stämmer väl överens med modellen som anförs här. Huvudfrågan är: hade den nya lönepolicyn något samband med förändringar i övervakningsteknologi, och i så fall hur? Den allmänna uppfattningen är att mekanisering av produktion och standardisering av arbetsuppgifter generellt gjorde det lättare att hindra arbetare från att maska. Enligt den här essän kan löneökningen ha berott just på att övervakningen blev lättare.

Tidigare teoretisk litteratur har tagit det för givet att arbetare har det bättre i företag där det är svårt att övervaka och i företag som betalar höga löner. Vi visar att ingen av dessa påståenden är sanna. Arbetare har det bäst i mellanliggande fall. I företag där övervakning är antingen väldigt lätt eller väldigt svårt kommer arbetare att betalas nära deras reservationslön. Skillnaden är att när övervakning är lätt kommer företaget att kräva en högre prestationsnivå, men samtidigt kompensera för de höga kraven med en hög lön. De allra högsta lönerna kommer att betalas av företag som övervakar väldigt noggrant och som därmed inte behöver betala någon lönepremie alls.

Uppsatsen identifierar också två allvarliga problem förbundna med empirisk prövning av teorin. Det första är att övervakningsprecisionen är väldigt svår att mäta. De flesta studier låter övervakningsprecision representeras av övervakningsintensitet, mätt som antalet övervakare eller arbetsledare per arbetare. Detta är illegitimt. Antalet övervakare eller arbetsledare kan mycket väl vara stort just därför att det är svårt att övervaka. Existerande studier har i huvudsak undersökt relationen mellan övervakningskostnader och lön. Olyckligtvis är denna relation mer komplicerad än den mellan övervakningsnivå och lön. Detta leder till det andra problemet: om det är tillräckligt stor variation i de underliggande parametrarna kommer det att finnas två möjliga lönenivåer för varje given övervakningskostnad. Den empiriska litteraturens regressioner av lönepremie på övervakningskostnader är därför oberättigade om inte prestationsnivån är rätt kontrollerad för. Men allt är inte förlorat. Ett mer konstruktiv bidrag från vårt arbete är att det visar att det fortfarande är legitimt att köra en regression av övervakningskostnader på lönepremier. Med andra ord, existerande dataset kan fortfarande vara värdefulla för att testa teorin om bara regressionerna körs omvänt!

Denna uppsats förklarar således hur en väldigt naturlig generalisering av den sedvanliga maskningsversionen av effektivitetslönemodellen fullkomligt omkullkastar tidigare intuitioner. Den enkla empiriska förutsägelsen är: när övervakningsprecisionen är hög är också lönerna höga, och detta på grund av existensen av en skaleffekt som dominerar den substitutionseffekt tidigare arbeten koncentrerat sig på.

Uppsats 2

Monitoring and Pay: General Results I det föregående kapitlet analyseras hur övervakning och lön varierar efter multiplikativa förändringar och det visar sig att de är positivt korrelerade. Den här uppsatsen demonstrerar att detta resultat gäller även under mer generella antaganden. Den viktigaste poängen är att resultatet gäller för alla monotona förändringar; inte bara de multiplikativa.

Två utvidgningar av modellen undersöks också. Den första introducerar en ex ante individuell rationalitetsrestriktion för arbetaren utöver den tidigare ex post begränsade ansvarsrestriktionen. Detta är typiskt intressant när arbetslösa bjuder under varandra för att få ett arbetstillfälle. Slutsatsen av detta är att en bindande individuell rationalitetsrestriktion gör det tidigare resultatet än mer generellt.

Den andra utvidgningen tillåter arbetsgivaren att trovärdigt använda sig av en mixad strategi när han övervakar en arbetare. Då kommer varje konkav övervakningsfunktion domineras och ersättas av en ny som är billigare och linjär. En svagt konvex övervakningsfuntion visar sig i sin tur också göra det tidigare resultatet mer generellt.

Dessutom presenteras ett ramverk i vilket optimala ickelinjära incitament kan jämföras med de bästa linjära incitamenten. Det huvudsakliga resultatet är att för företaget finns det alltid ett icke linjärt lönekontrakt som strikt dominerar varje linjärt lönekontrakt.

Uppsats 3

Limited Liability and Dynamic Incentives Den invändning mot den ovan diskuterade varianten av effektivitetslönemodellen som oftast förekommer i litteraturen är att om modellen stämmer borde arbetslösa arbetare erbjuda sig att betala en sorts inträdesavgift för ett arbetstillfälle, men sådana avgifter förekommer väldigt sällan. Det mest kända artikel som presenterar denna variant av effektiviteslönemodellen är Shapiro och Stiglitz (1984). Faktum är att Shapiro och Stiglitz var medvetna om inträdesavgiftskritiken redan när de skrev artikeln. De framhäver de två mest uppenbara argumenten mot att arbetare skulle förvärva arbetstillfällen. Det första är att arbetare helt enkelt kanske inte har råd att betala en relevant inträdesavgift på grund av en ofullkomlig kapitalmarknad. Det andra är att det resulterande anställningskontraktet inte kommer bli sådant att det upprätthåller sig själv. Företaget skulle ha incitament att hävda att en arbetare maskar för att avskeda henne efter det att avgiften är erlagd och sälja arbetstillfället ytterligare en gång.

Många nationalekonomer har emellertid inte blivit övertygade av dessa förklaringar för varför inträdesavgifter inte förekommer. Carmichael (1985), in hans replik på Shapiro och Stiglitz, och Carmichael (1990) hävdar att varken förekomsten av en ofullkomlig kapitalmarknad eller det beskrivna moral hazard problemet är tillräckligt för att förhindra förekomsten av inträdesavgifter. Moral hazard problemet kan till exempel lösas med hjälp av en tredje part. Utöver detta vet varken företag eller arbetare om när deras förhållande kommer ta slut i Shapiro Stiglitz modellen. Bull (1985, 1987) påstår att detta är orealistiskt och fortsätter med att visa att arbetslöshet inte kan generera det överskott som krävs för ett självupprätthållande anställningskontrakt i en modell med en bestämd sista period.

Följden har blivit att effektivitetslönemodellen förlorat förespråkare till förmån för den besläktad teorin om fördröjd betalnings/obligationskontrakt (Arbetaren ger företaget pengar vid anställningstillfället mot en revers/obligation som kan lösas in senare). I dessa kontrakt betalar företaget initialt löner under arbetarens reservationslön och senare löner över reservationslönen, för att skapa incitament och motverka maskning när övervakningen är ofullständig. På detta sätt kan företaget pressa arbetaren mot hennes individuella rationalitets restriktion och man undviker därmed inträdesavgiftskritiken.

Det huvudsakliga syftet med den här essän är att visa hur ett företag bäst avpassar koordinationen mellan tidpunkt och nivå för löneutbetalningar till en arbetare. Incitament kan skapas med både effektivitetslöner och framskjuten betalning och det visar sig att det rationella företaget kommer använda sig av båda metoderna. I en självupprätthållande effektivitetslönemodell med ett ändligt antal perioder och förmögenhetsbegränsade arbetare, visas det att den optimala löneutvecklingen mycket väl kan getalta sig som enligt ett fördröjd betalnings/obligationskontrakt, som till exempel Lazear (1981) och Becker och Stigler (1974). Det kommer typiskt utgöras av en hög lön i sista perioden samt en lägre och stationär lön för övriga perioder. Arbetaren kommer få del av den genererade vinsten som i effektivitetslönemodeller, men denna del snabbt mindre desto längre anställningstid kontraktet avser. Detta resultat är kanske mest uppseendeväckande för att det inte fullt ut eliminerar arbetarens del av vinsten; bland andra hävdar Lazear (book) att fördröjd betalning är ett perfekt substitut för inträdesavgifter. Anmärkningsvärd är den stora skillnaden att den höga lönen i den sista perioden som härstammar från effektivitetslönemekanismen är optimal att betala ut, inte för att den kompenserar för den låga lönen i tidigare perioder, utan för att den ger upphov till korrekta incitament just i den sista perioden.

Obligationskontrakt och inträdesavgifter kan i princip avlägsna arbetarens resterande del av företagets vinst. Både obligationer och inträdesavgifter är emellertid underlägsna medel för att göra detta. Investeringar i företagsspecifikt humankapital löser moral hazard problemet mer effektivt. Shapiro och Stiglitz nämner kortfattat att när kostnaden av att förlora företagsspecifikt humankapital är påtaglig kan arbetare ha incitament att anstränga sig även under full sysselsättning. Detta är sant även i den här essän. Kärnpunkten är den att företagsspecifikt humankapital ökar en arbetares ansvar. Därmed ökar företagets vinst på arbetarens bekostnad. Arbetslösa och arbetare anställda i andra företag bjuder över varandra med högre och högre utbildningsnivåer för ett arbetstillfälle. Detta pågår ända tills den arbetare som slutligen blir anställd är helt extraherad. Skälet varför buden är i utbildningsnivåer och inte i pengar är helt enkelt att det löser moral hazard problemet på enklast möjliga sätt; om företaget avskedar arbetaren kommer det att förlora det företagsspecifika humankapitalet tillsammans med arbetaren.

Uppsats 4

Do Market-Based Incentives Lower the Cost of Compliance? Tar man inte hänsyn till övervakningskostnader och kostnader för upprätthållande av en regleringsstandard när man utformar en regleringspolicy kommer den med stor säkerhet att bli underlägsen vad som annars kunnat åstadkommas. Antingen kommer regleringsmyndigheten använda sig av en för dyr policy eller så kommer den inte lyckas med att få företag att rätta sig efter policyn. Faktum är att konsekvenserna kan bli närmast katastrofala, speciellt i det senare fallet. Detta är en läxa både den akademiska världen och regleringsmakare tagit lärdom av de senaste åren. Som en följd har övervakning och frågan om upprätthållandet av en regleringspolicy hamnat i centrum i många nyligen utgivna teoretiska och empiriska studier, och regleringsmyndigheter har läst de föreslagna ordinationerna noggrant. Den mest uppenbara frågan de vill ha besvarad är självklart: Vilken är den socialt mest önskvärda regleringsformen? Eller mer precist, vad är att föredra, en kvotreglering eller mer marknadsbaserade incitament?

Detta kapitel försöker svara på den frågan, i första hand för en negativ extern miljöeffekt, men resonemanget genom hela uppsatsen kan lätt utsträckas till att omfatta de flesta typer negativa externa effekter. Uppsatsen börjar med att härleda den optimala kvotregleringen som kommer att bestå av en utsläppskvot, ett vite ifall utsläppsnivån överstiger utsläppskvoten samt en skatt.

Företaget måste få korrekta incitament för att lyda. Mekanismen för detta är tidigast beskriven i Beckers (1968) ekonomiska analys av kriminalitet. Den huvudsakliga insikten från den artikeln är att potentiella brottslingar reagerar på både sannolikheten att bli upptäckta och hur strängt straff de skulle få om de blev upptäckta och dömda. Man kan med andra ord avskräcka från kriminell aktivitet antingen genom högre straff eller genom ökad övervakning så att sannolikheten att en förövare blir fångad ökar. Samma sak gäller i den här uppsatsen. Emellertid antas det att det finns en exogent bestämd gräns för hur strängt straffet kan vara, och den gränsen kommer att vara bindande. Skälet kan vara politiskt eller på grund av en förmögenhetsbegränsning för det reglerade företaget. Istället kan nettostraffet varieras, dvs mellanskillnaden mellan straff och skatt.

Det första slående resultatet är att den optimala skatten mycket väl kan bli negativ, dvs vissa företag som förorenar kommer inte åläggas någon skatt alls utan får istället en subvention. Detta följer av att en skatteminskning optimalt kommer att användas som ett komplement till övervakning för att skapa incitament. Vissa företag som man vill ge starka incitament (t ex för att utsläppet är väldigt skadligt), kommer att övervakas väldigt noggrant och därmed få en negativ skatt.

Iden om en subventionering har funnits tidigare i miljöekonomi litteraturen: Downing och Kimball (1982) noterar att en kostnadssubventionering gör reglerade företag mer benägna att rätta sig efter en regleringspolicy för att företagen då har mer att förlora på att inte underkasta sig. Samma budskap förs fram i till exempel Sullivan (1987), Fullerton och Kinnaman (1995) och Sigman (1998). Ett något annorlunda skäl för en subvention presenteras i Nowell och Shogren (1994). De studerar hur en subvention påverkar företags aktivitet att försöka undvika ansvar.

Uppsatsen fortsätter med att ta reda på den optimala utsläppsavgiftspolicyn och det visas att denna policy är minst lika bra som den bästa marknadsbaserade policyn. Slutligen jämförs kvotpolicyn med utsläppsavgiftspolicyn, vilket leder direkt till ett andra väldigt intressant resultat. Trots att merparten av litteraturen inom miljöekonomi förespråkar användandet av marknadsbaserade incitament är slutsatsen här det omvända: ett utsläppsbegränsningsmål är alltid mer kostsamt att genomdriva om det finns en marknad för utsläppsrättigheter än annars. Svaret på frågan i titeln är med andra ord: nej, marknadsbaserade incitament sänker inte kostnaden för att få företag att rätta sig efter en policy. Intuitionen bakom resultatet är helt enkelt att en marknad begränsar regleringsmyndigheten till att använda incitament som är linjära i det utsläppande företagets utsläppsnivå, och att en sådan restriktion ger försvagade incitament.

References

Becker, Gary and George Stigler (1974): Law Enforcement, Malfeasance, and the Compensation of Enforcers, *Journal of Legal Studies* III, 1-18.

Becker, Gary (1968): Crime and Punishment: An Economic Approach, *Journal of Political Economy* 76, 169-217.

Bull, Clive (1987): The Existence of Self-Enforcing Implicit Contracts, *Quarterly Journal of Economics* February, 147–159.

Bull, Clive (1985): Equilibrium Unemployment as a Worker Discipline Device: Comment, *American Economic Review* September, 890–891.

Bulow, Jeremy I. and Lawrence H. Summers (1986): A Theory of Dual Labor Markets with Applications to Industrial Policy, Discrimination, and Keynesian Unemployment, *Journal of Labor Economics* 4, 376–414.

Carmichael, H. Lorne (1990): Efficiency Wage Models of Unemployment - One View, *Economic Inquiry* April, 28, 269-295.

Carmichael, H. Lorne (1985): Can Unemployment be Involuntary?: Comment, *American Economic Review* December, 75, 1213-1214.

Downing, P. and J. Kimball (1982): Enforcing Pollution Control Laws in the United States, *Policy Studies Journal* 11, no. 1.

Fullerton, D. and T. C. Kinnaman (1995): Garbage, Recycling, and Illicit Burning or Dumping, *Journal of Environmental Economics and Management* 29, 78-91.

Holmström, Bengt and Paul Milgrom (1994): The Firm as an Incentive System, *American Economic Review* 84, 972–991.

Lazear, Edward (1981):Agency, Earnings Profiles, Productivity and Hours Restrictions, *American Economic Review* LXXI, 606-620.

Nowell, C. and J. Shogren (1994): Challenging the Enforcement of Environmental Regulation, *Journal of Regulatory Economics* 6, 265-282.

Shapiro, Carl and Joseph E. Stiglitz (1984): Involuntary Unemployment as a Worker Discipline Device, *American Economic Review* 74, 433–444.

Sigman, H. (1998): Midnight Dumping: Public Policies and Illegal Disposal of Used Oil, *RAND Journal of Economics* 29(1), 155-178.

Sullivan, A. M. (1987): Policy Options for Toxics Disposal: Laissez-Faire, Subsidization, and Enforcement, *Journal of Environmental Economics and Management* 14, 58-71.

2 MONITORING AND PAY

with Tore Ellingsen

Abstract The shirking model of efficiency wages has been thought to imply that monitoring and pay are substitute instruments for motivating workers. We demonstrate that previous results hinge on unduly restrictive assumptions regarding workers' choice of effort – for example that there are only two possible effort levels. Under more reasonable assumptions, monitoring and pay are complementary instruments. Another result is that there is a non–monotonic relationship between the wage level and the workers' rents. Moreover, much of the empirical literature on the monitoring–pay relationship is shown to be misguided. Only two existing studies use an appropriate methodology. They reject the model's main implication.

JEL CLASSIFICATION: J31, J41.
KEYWORDS: Monitoring, efficiency wages, incentive pay.

Introduction

Conventional wisdom says that monitoring and pay are substitute instruments for motivating workers: workers who are poorly monitored have to be well paid in order not to shirk, and conversely, workers who are closely monitored do not have to be particularly well paid. This version of the efficiency wage idea has a long and distinguished history. It is also the cornerstone of Bulow and Summers' (1986) theory of dual labor markets. In short, their theory says that there is a primary sector in which monitoring is difficult—hence working conditions are good and pay is high; and there is a secondary sector in which monitoring is easy—hence working conditions are poor and pay is low.

In this paper we shall argue that the accuracy of monitoring and the level of pay are negatively related only under very restrictive assumptions. Notably, the above argument is correct when workers only have a choice between two effort levels (working and shirking), but incorrect when worker effort can be adjusted in a continuous fashion. Similarly, the argument is correct when the desirable level of worker effort is given exogenously, but incorrect when the desirable level of worker effort emerges as part of the solution to the firm's profit maximization problem.

The point is that monitoring costs affect not only the choice between the carrot (pay) and the stick (monitoring) in implementing a given level of effort, but also the level of effort itself. In general, there is a scale effect in addition to the substitution effect.[1] For example, as monitoring becomes more costly, there will be a shift from monitoring toward wage incentives, just as Bulow and Summers claim, but there will also be a reduction in the desired effort level. Without further analysis, the total effect on the wage is therefore ambiguous. Indeed, as we shall show, the wage is likely to go down, not up.

Briefly stated, our main contribution is to extend the simple shirking model to a continuum of effort levels and to characterize the profit maximizing levels of monitoring and pay. Interestingly, complementarity holds not only for changes in monitoring costs, but for changes in any other parameter as well. The main finding is that, under weak assumptions, monitoring and pay are complementary instruments. A parameter change which causes an increase in the accuracy of monitoring

[1] The possibility that scale effects could outweigh substitution effects has been mentioned earlier by Rebitzer (1993,1995). However, as far as we know, the issue has not been analyzed in any detail.

also causes an increase in the level of pay and vice versa.

Earlier theoretical literature has taken it for granted that workers are better off in firms where it is difficult to monitor and in firms which pay high wages. Neither of these propositions is true here. Workers are best off in intermediate cases. In firms where monitoring is either very difficult or very easy, workers are paid close to their reservation wage. The difference is that when monitoring is easy, workers are required to work harder and therefore receive a higher wage. Indeed, the very highest wages will be paid by firms which monitor very accurately, and hence do not need to pay a wage premium.

The paper also identifies two serious problems associated with empirical implementation of the theory. First, the accuracy of monitoring is very hard to measure. Most studies proxy the accuracy of monitoring by the intensity of supervision, as measured by the number of supervisors per worker or the frequency of supervision.[2] This is illegitimate. The number of supervisors may be high precisely because monitoring is difficult. Existing studies have essentially investigated the relationship between monitoring costs and pay. Unfortunately, this relationship is more complicated than the relationship between the monitoring level and pay. Here is a second problem: if there is sufficient variation in the underlying (unmeasured) parameters, there will be two possible wage levels for each level of monitoring costs. Thus, unless the effort level is properly controlled for, the empirical literature's regressions of wage premiums on monitoring costs are unjustified. But all is not lost. A more constructive contribution of our work is to show that it is still legitimate to run a regression of monitoring costs on wage premiums. In other words, existing data sets can still be of value in testing the theory if only regressions are run in reverse!

Our basic model can be extended in a number of directions. The extensions we pursue involve imposing plausible restrictions on the feasible set of monitoring levels. These extensions enable us to link the theoretical model quite closely to the empirical work of Groshen and Krueger (1990) on supervision and pay in hospitals and to that of Krueger (1991) on franchising. As it turns out, these two studies, which are arguably the only studies which properly address the predictions of our model, both reject the predictions. Unless these predictions are also artefacts of remaining simplifying assumptions, the shirking model looks to be in bad shape.

The paper is organized as follows. The next section sets up the model and the main result is derived. Welfare implications are discussed in the third section. In the following section, we

[2] See e.g. Leonard (1987), Gordon (1990,1994), Kruse (1992), Neal (1993).

investigate what happens if we allow for other sources of variation than the cost of monitoring. The fifth section discusses the model's implications and how it should be tested. Here we also discuss an issue which is important in some applications: discrete monitoring technologies (organizational forms). The last section concludes.

Model

The set-up is a simple efficiency wage model in the spirit of Shapiro and Stiglitz (1984). A risk neutral principal employs a single risk neutral agent. The agent can exert effort $e \in \mathbb{R}_+$ which affects the principal's benefit, $\beta B(e)$, at some cost to the agent $\gamma C(e)$. By assumption, the parameters β and γ are both positive. The principal motivates the agent through a compensation contract $w(e)$, where w is the wage that the agent receives if the principal observes e.

The principal can observe and verify the agent's effort with probability p. This probability is affected by the principal's choice of monitoring technology. In order to attain probability p of observing the agent's effort, the principal has to pay $\mu M(p)$, where $\mu > 0$. We make the following assumption regarding functional forms.

Assumption 1

(i) $B'(e) > 0$, $B''(e) \leq 0$,

(ii) $C(0) = 0$, $C'(e) > 0$, $C''(e) > 0$,

(iii) $M'(p) > 0$.

The ex post utility of the principal can now be written

$$U = \beta B(e) - w - \mu M(p), \tag{2.1}$$

and the ex post utility of the agent is

$$V = w - \gamma C(e). \tag{2.2}$$

A key assumption of the efficiency wage model is that there is a lower limit, $w_0 \in \mathbb{R}$, to the payment. The limit may be due either to legal rules or to a wealth constraint.

Since effort is not always observed, the compensation contract also needs to specify some payment $\bar{w} \in \mathbb{R}$ that the agent is to receive in this case. The agent is assumed to maximize expected utility,

$$\mathrm{E}[V] = pw(e) + (1-p)\bar{w} - \gamma C(e). \tag{2.3}$$

Discrete effort levels

For a moment assume that the effort level is restricted to take only the values $e = 0$ and $e = \bar{e}$. This is the case in earlier versions of the efficiency wage model. Then the following incentive constraint must be satisfied

$$pw(\bar{e}) + (1-p)\bar{w} - \gamma C(\bar{e}) \geq pw(0) + (1-p)\bar{w}.$$

The principal will optimally punish the agent as much as possible if he is detected cheating, thus $w(0) = w_0$. We impose the standard assumption that an indifferent agent takes the action that the principal favors. The principal can then lower the wage down to the level where the incentive compatibility constraint becomes an equality and the incentive constraint can be simplified to

$$p(w - w_0) = \gamma C(\bar{e}), \tag{2.4}$$

where $w(\bar{e})$ is denoted by w. This constraint tells us that, as long as the right hand side is constant, an increase in the equilibrium value of monitoring must be followed by a decrease in the equilibrium value of the wage and vice versa. In Figure 1 we have plotted the points of w and p which are just sufficient to implement the effort level $\bar{e}$. After a decrease in μ the principal will move from the initial equilibrium A along the iso-effort curve to point B, substituting wage for monitoring accuracy in order to minimize costs. That is, monitoring and pay are substitute instruments.

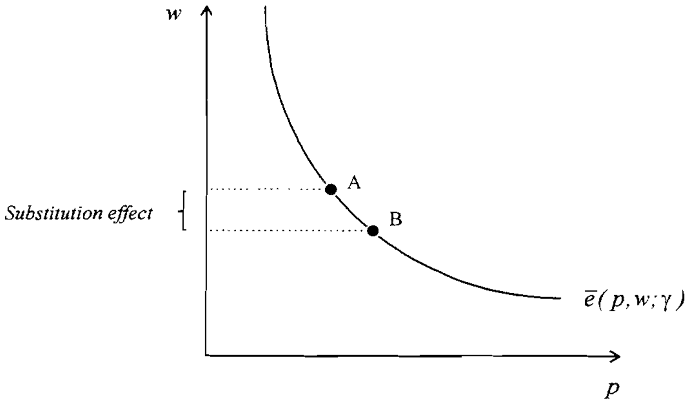


Figure 1

Continuous effort levels

Fortunately for empirical work, this result is an artefact of the very strong restriction that the effort level can only take two possible values. We will now show that, under weak assumptions, there is a scale effect that dominates the substitution effect, i.e. p and w are complementary instruments with respect to changes in μ. Figure 2 illustrates the result. Now, we have plotted two iso-effort curves for different levels of effort. After a decrease in μ, cost minimization implies a move from the initial equilibrium A, along the iso-effort curve to point B. However, at this point the marginal revenue is not equated to the marginal cost of effort, which is required for profit maximization. In order to maximize profits, the principal has to increase the effort level, ending up at point C. The substitution of pay for monitoring accuracy is dominated by the principal's complementary use of monitoring accuracy and pay to increase the implemented level of effort.

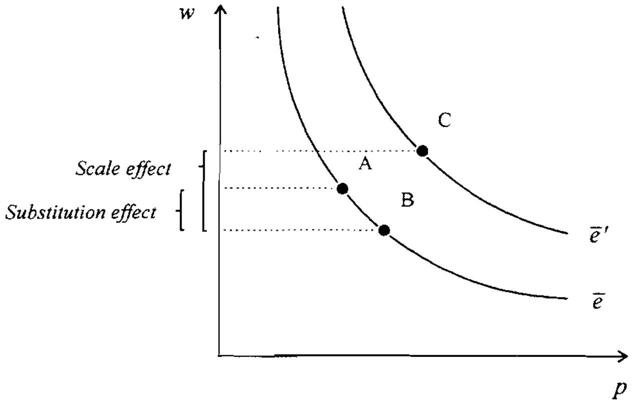


Figure 2

Let us now prove this claim more formally. When the principal is free to induce the agent to take any level of effort $\hat{e} \in \mathbb{R}_+$, the following incentive compatibility constraint must be satisfied for all e:

$$pw(\hat{e}) + (1-p)\bar{w} - \gamma C(\hat{e}) \geq pw(e) + (1-p)\bar{w} - \gamma C(e).$$

We see that any incentive compatible contract that implements $\hat{e}$ can, without loss to the principal, be replicated by a step function of the form $w(e) = w_0$ for $e < \hat{e}$ and $w(e) = w$ for $e \geq \hat{e}$. I.e., the principal sets an effort target, $\hat{e}$. The agent gets w if he meets or exceeds the target and the minimum payment

w_0 otherwise.[3] This contract is illustrated in Figure 3.

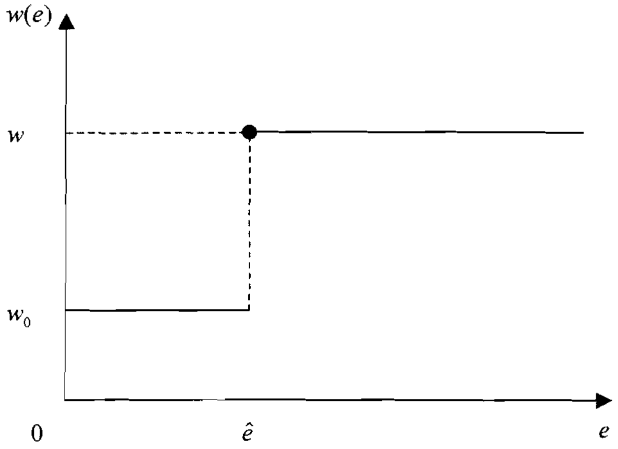


Figure 3

Given this kind of a contract, if an agent ever wants to deviate, he will deviate to $e = 0$. Thus, the incentive compatibility constraint becomes

$$p(w - w_0) \geq \gamma C(\hat{e}). \tag{2.5}$$

Finally, the assumption that an indifferent agent takes the action favored by the principal enables the principal to lower the wage down to the level where the incentive compatibility constraint becomes an equality. Inverting this equality, we obtain an expression for the actual effort which the agent will exert,

$$e(p, w) := C^{-1}((w - w_0)p/\gamma). \tag{2.6}$$

Note that $\bar{w}$, the wage in the case that the effort level is not observed, is irrelevant for the agent's incentives. It is common to assume that $\bar{w} = w(\hat{e})$, and we follow this practice. One reason why the assumption is plausible is that if $\bar{w} < w(\hat{e})$, the principal would have had an incentive not to monitor (or to falsely claim that he has not monitored).

Before we turn to the detailed analysis, it is useful to define

$$r(e) := \frac{\beta B'(e)}{\gamma C'(e)},$$

i.e. the ratio of the marginal benefit to the marginal cost. Needless to say, this ratio would be equal to one with perfect monitoring.

[3] As shown by Demougin and Fluet (1997), the optimality of effort targets is quite general. It is not an artefact of the simple monitoring technology.

The principal's problem is to find a probability p and a wage w to maximize

$$U(p,w) = \beta B(e(p,w)) - w - \mu M(p) \tag{2.7}$$

subject to the constraints $w \geq w_0$ and $p \in [0,1]$. This is a straightforward maximization problem in two variables. Let a solution to this problem be denoted (p^*, w^*), and let $e^* := e(p^*, w^*)$ denote the associated effort level. The first–order conditions for the solution can then be written

$$p^* r(e^*) - 1 \leq 0, \tag{2.8}$$

with equality if $w^* > w_0$, and

$$(w^* - w_0) r(e^*) - \mu M'(p^*) \geq 0, \tag{2.9}$$

with equality if $p^* < 1$.

Equation (2.8) tells us that the marginal benefit from increased effort will be larger than the marginal cost whenever the principal chooses to monitor imperfectly. The reason for the distortion away from the socially optimal effort level is that the agent must be paid a rent in order not to shirk. Also, equation (2.8) confirms that there must be a positive level of monitoring in order to induce any effort. (If $p^* = 0$, then $w^* = w_0$.)

Let U_{ij} denote the second derivative of U and let U^*_{ij} denote the second derivative of U evaluated at a solution (p^*, w^*). The second–order conditions are then $U^*_{ww} < 0, U^*_{pp} < 0$ and $U^*_{pp} U^*_{ww} - (U^*_{pw})^2 > 0$. To state the conditions in full, note that

$$U_{ww} = \frac{\beta (p)^2 h(e)}{\gamma^2}, \tag{2.10}$$

$$U_{pp} = \frac{\beta (w - w_0)^2 h(e)}{\gamma^2} - \mu M''(p), \tag{2.11}$$

$$U_{pw} = \frac{\beta p (w - w_0) h(e)}{\gamma^2} + r(e), \tag{2.12}$$

where

$$h(e) := \frac{B''(e) C'(e) - C''(e) B'(e)}{(C'(e))^3}.$$

Main results

Our main objective is to characterize how the wage level and the accuracy of monitoring vary with the parameter μ. Given that the solution is interior, this is a standard comparative static exercise.

The effect of a change in the cost of monitoring, μ, is found by differentiating the two first-order conditions (2.8) and (2.9) to get the two equations

$$\begin{aligned} U^*_{ww}dw^* + U^*_{pw}dp^* &= 0 \\ U^*_{pw}dw^* + U^*_{pp}dp^* &= M'(p^*)d\mu. \end{aligned}$$

Using Cramer's rule, we then have

$$\frac{dw^*}{d\mu} = \frac{-U^*_{pw}M'(p^*)}{U^*_{ww}U^*_{pp} - (U^*_{pw})^2} \tag{2.13}$$

and

$$\frac{dp^*}{d\mu} = \frac{U^*_{ww}M'(p^*)}{U^*_{ww}U^*_{pp} - (U^*_{pw})^2}. \tag{2.14}$$

As one would expect, an increase in μ always lowers the level of monitoring. In the other case, the sign of the cross-derivative U^*_{pw} plays a crucial role. Following common usage, we say that monitoring and pay are Edgeworth complements (substitutes) if U^*_{pw} is positive (negative).

Thus, our main result can be stated as follows.

Lemma 1 *If $U_{pw} > 0$, monitoring and pay are complementary instruments.*

We are left with the question: How realistic is it that monitoring and pay are Edgeworth complements?

We claim that monitoring and pay are always Edgeworth complements if the principal's benefit function, $B(e)$, and the agent's cost of effort function, $C(e)$, are both represented by any kind of power function satisfying our initial assumptions. To see this, just let $B(e) = b_0 e^{b_1}$ and $C(e) = c_0 e^{c_1}$. Now, it follows from (2.6) and (2.7) that

$$U = \beta b_0[(w - w_0)p/\gamma c_0]^{b_1/c_1} - w - \mu M(p),$$

so

$$U_{pw} = \frac{\beta b_0 b_1^2}{\gamma c_0 c_1^2}\left[\frac{(w - w_0)p}{\gamma c_0}\right]^{\frac{b_1 - c_1}{c_1}},$$

which is clearly positive if both b_0 and c_0 are positive.

In the commonly studied special case that the principal's benefit function is linear, a more general statement can be made, namely that monitoring and pay are Edgeworth complements, and hence complementary instruments, if and only if the relative growth of cost of effort is decreasing in the effort level.

To prove this formally, let $B(e) = b_0 e$. It follows that $B'(e) = b_0$ and that $B''(e) = 0$. Now monitoring and pay are Edgeworth complements if and only if (2.12) is positive in equilibrium, i.e. if and only if

$$U^*_{pw} = -\frac{\beta b_0 p^*(w^* - w_0)C''(e^*)}{\gamma^2(C'(e^*))^3} + \frac{\beta b_0}{\gamma C'(e^*)} > 0.$$

Use the IC-constraint, (2.6), to rewrite the above condition to

$$U^*_{pw} = \frac{\beta}{\gamma}\frac{(C'(e^*))^2 - C(e^*)C''(e^*)}{(C'(e^*))^3} > 0. \tag{2.15}$$

That is, $U^*_{pw} > 0$ if and only if

$$\frac{\partial\frac{C'(e^*)}{C(e^*)}}{\partial e^*} < 0. \tag{2.16}$$

This leads to the following proposition.

Proposition 1 *Sufficient conditions for monitoring and pay to be complementary instruments are (each effective individually):*

1) The principal's benefit function, $B(e)$, and the agent's cost of effort function, $C(e)$, can both be represented by power functions.

2) The principal's benefit function, $B(e)$, is linear and the relative growth of cost of effort is decreasing in the effort level.

The latter condition is a property common for a large set of functions. For example, it can be shown that if the agent's cost of effort is represented by any polynomial function with positive coefficients, monitoring and pay are Edgeworth complements. On the other hand, it should be noted that rational functions and exponential functions can be constructed so that the relative growth is increasing over some range; thus for these forms of functions we have to assume that inequality (2.16) holds in equilibrium.

In the literature, two special cases of the efficiency wage model prevail. Either a linear benefit function and a quadratic cost of effort function are assumed, or alternatively, workers only have a choice between two effort levels. It is noteworthy that a linear benefit function with a quadratic cost function satisfies *both* the conditions in Proposition 1, so there is no question that complementarity applies here!

The case when the level of effort is restricted to take only two values can be thought of as a special case of our continuous model with an extremely concave cost of effort function. For example, in the above exercise where the benefit function and the cost of effort function are represented by power functions, let the parameter c_1 go to infinity. It follows that $C(e > 1) = \infty$, $C(e = 1) = c_0$ and $C(e < 1) = 0$; thus the only two relevant levels of effort are $e = 1$ and $e = 0$. To find the sign of U_{pw}, just let $c_1 \rightarrow \infty$ in the expression (2.15), whereupon it is seen that $U_{pw} \rightarrow 0_+$, i.e. converges to zero *from above.* Thus U_{pw} is indeed strictly positive, and monitoring and pay are complementary instruments. Consequently, the result in Shapiro and Stiglitz (1984) can *not* be interpreted as a special case of our result.

Welfare implications

Most work on dual labor markets has taken for granted that workers are better off in jobs that are poorly monitored. Here, this is not true at all. From (2.8) and (2.9), we have that

$$w^* = w_0 + p^* \mu M'(p^*), \tag{2.17}$$

and from (2.5) we know that

$$\gamma C(e(p^*, w^*)) = p(w^* - w_0). \tag{2.18}$$

The agent's utility can hence be written

$$\begin{aligned} V &= w^* - \gamma C(e(p^*, w^*)) \\ &= w_0 + p^*(1 - p^*)\mu M'(p^*). \end{aligned} \tag{2.19}$$

Thus, for $p^* = 0$ and for $p^* = 1$ the agent's utility is at its minimum level, w_0. For all other levels of monitoring, the agent earns rents. The point is very simple: if the principal cannot monitor at all, the agent will always shirk, and there is no reason to pay an efficiency wage. Conversely, if the principal can monitor perfectly, any effort level can be implemented as long as the agent is paid the corresponding reservation wage.

By implication, the agent's utility is a non-monotonic function of w^*. The wage is increasing in the level of monitoring, and hence utility is at its lowest for very low and very high wages. The best paid workers are required to put in a lot of effort. Hence, the true rents are largest in the intermediate wage range.

If the model is an appropriate description of reality, it means that workers need not be worse off when monitoring technologies improve, at least as long as monitoring remains reasonably imperfect. This may be one reason why the mechanization of production, contrary to the claims of Braverman (1974) and other Marxian scholars, does not seem to have implied increased exploitation of workers.[4]

Other sources of variation

Another obstacle to empirical testing of the model is that μ is not the only parameter of the model. What if γ changes instead? Even in a model with only two effort levels we can show that monitoring and pay are complements in this case. Figure 4 illustrates the outcome. An increase in the cost of effort, represented by a move from γ to γ', means that the agent requires a stronger incentive to maintain the effort level $\bar{e}$. This can be accomplished by higher wages or by more monitoring. It turns out that the principal optimally uses a little of both (see Appendix 1). In this case monitoring and pay are complementary instruments and the new incentive contract is represented by the contract B.

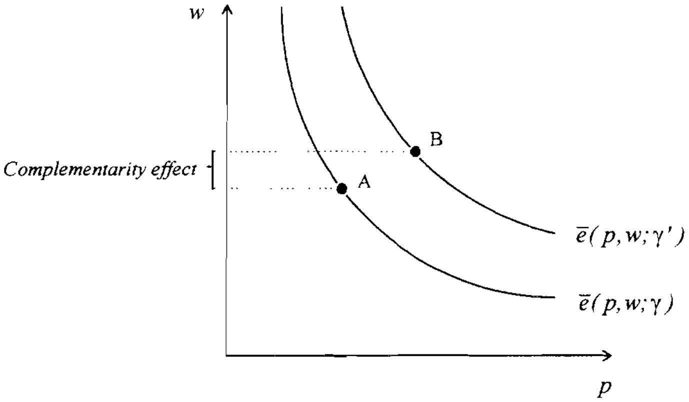


Figure 4

Moreover, if the source of variation is β, nothing will happen with the equilibrium values of p and w provided that it is still profitable for the principal to remain in business.[5] Consequently, the discrete model yields a different implication for any possible source of variation.

[4] Of course, this does not imply that the level of monitoring is efficient. As Bowles (1985) has pointed out, in the shirking model a slight reduction in monitoring accompanied by a slight increase in the wage yields a second-order loss to the principal and a first-order gain to the agent.

[5] See Appendix 1 for details.

Let us now investigate how the wage level and the accuracy of monitoring vary with the parameters β and γ in the case with continuous effort. The effect of a change in the principal's benefit from effort, β, is found by differentiating the two first-order conditions (2.8) and (2.9) to get the two equations

$$\begin{aligned} U^*_{ww}dw^* + U^*_{pw}dp^* &= -\frac{p^*}{\beta}r(e^*)d\beta, \\ U^*_{pw}dw^* + U^*_{pp}dp^* &= -\frac{w^* - w_0}{\beta}r(e^*)d\beta. \end{aligned}$$

Again, using Cramer's rule, we have that

$$\frac{dw^*}{d\beta} = \frac{[-p^*U^*_{pp} + (w^* - w_0)U^*_{pw}]r(e^*)}{[U^*_{ww}U^*_{pp} - (U^*_{pw})^2]\beta} \tag{2.20}$$

and

$$\frac{dp^*}{d\beta} = \frac{[-(w^* - w_0)U^*_{ww} + p^*U^*_{pw}]r(e^*)}{[U^*_{ww}U^*_{pp} - (U^*_{pw})^2]\beta}. \tag{2.21}$$

The sign of $dp^*/d\beta$ is now easily determined. By the second-order condition and Assumption 1, we have

$$\operatorname{sign}\frac{dp^*}{d\beta} = \operatorname{sign}[-(w^* - w_0)U^*_{ww} + p^*U^*_{pw}] = \operatorname{sign} r(e^*) > 0.$$

The intuition is quite clear. When the principal has more to gain from increased effort, he will find it optimal to engage in more monitoring – this being one way of making sure that the agent does not shirk. A similar argument would seem to guarantee that the wage is increasing in β. However, this is *not* true without further assumptions. More precisely, we have from (2.20) that $dw^*/d\beta > 0$ if and only if

$$\mu p^* M''(p^*) + (w^* - w_0)r(e^*) > 0. \tag{2.22}$$

Using equation (2.9), an alternative statement of the condition is

$$-p^*\frac{M''(p^*)}{M'(p^*)} < 1.$$

Thus, a sufficient, but not necessary, condition is that $M(p)$ is convex. Generally, we may allow some increasing returns to monitoring. Only when the returns to monitoring increase sufficiently fast could increased profitability of effort lead to a switch away from monetary incentives towards monitoring. Note that if the principal can commit to random monitoring, the relevant monitoring cost function does become convex even if $M(p)$ is concave.

Finally, we investigate the impact of a change in the cost of effort. The formulas in this case become

$$\frac{dw^*}{d\gamma} = \frac{-p^*\mu M''(p^*) - (w^* - w_0)r(e^*)}{[U^*_{pp}U^*_{ww} - (U^*_{pw})^2]\gamma}U^*_{pw} \tag{2.23}$$

and

$$\frac{dp^*}{d\gamma} = \frac{-p^*r(e^*)}{[U^*_{pp}U^*_{ww} - (U^*_{pw})^2]\gamma}U^*_{pw}. \qquad (2.24)$$

We see directly that, even here, monitoring and pay are complementary instruments if and only if condition (2.22) is fulfilled. The above results can be stated as follows.

Proposition 2 *If the source of variation is the β or γ parameter, monitoring and pay are complementary instruments if and only if $-p^*M''(p^*)/M'(p^*) < 1$.*

Since this is a very weak condition, we can safely conclude that w and p are complementary with respect to changes in β and γ.

At first blush, then, it seems that there are two important conditions, namely condition (2.22) and the sign of U_{pw}. However, there is in fact a close relationship between the two. To see this, consider the condition

$$U_{ww}U_{pp} - (U_{pw})^2 > 0,$$

which must hold at (p^*, w^*) (this is one of the second-order conditions). A few computations reveal that the inequality can be written as

$$\mu pM''(p) + (w - w_0)r(e) > r(e)\left(\frac{-\gamma^2 r(e)}{\beta ph(e)} - (w - w_0)\right).$$

It is straightforward to check that the right hand side of this inequality is positive if and only if $U_{pw} > 0$. Hence, if $U_{pw} > 0$, (2.22) is implied by the second-order condition.

Empirical implications

Prendergast (1999) argues that it is difficult to test for the existence of efficiency wages because we do not know if we would expect to see wages and supervisors as substitutes or complements. He writes (page 45): "The problem is that either may easily arise in a world of efficiency wages and depends critically on the source of variation across firms. On the one hand, if the source of variation across firms is the cost of supervisors, then the two instruments are likely to be substitutes, where firms substitute away from high-cost supervisors into wages. On the other hand, if the source of variation across firms is in the return to effort (so some firms value effort exertion more than others), those firms that want more effort will use more of both instruments relative to those that do not value such high effort."[6]

In contrast, the theory in this essay has a simple empirical implication: whenever the accuracy of monitoring is high, the level of pay should be high as well.[7]

Apart from Rebitzer (1995), existing empirical studies do not control for effort level effects. Hence, at first sight, the findings of Leonard (1987), Gordon (1990,1994), Kruse (1992), Neal (1993) and Arai (1994a, b) are directly relevant to our model. The evidence from these investigations of the monitoring/pay relationship is mixed, with about as many positive correlations as negative. Nonetheless, for reasons that we shall now give, we do not think these empirical results are very informative.

The measurement of monitoring

A major drawback of most empirical work to date is that it uses monitoring intensity as a measure of monitoring accuracy. Monitoring intensity is proxied by the number of supervisors per employee or by the frequency of supervision. But the whole point of Bulow and Summers' (1986) theory is that there are substantial differences in monitoring opportunities. The number of supervisors may therefore reflect the difficulty rather than the accuracy of monitoring.[8]

To be specific, suppose that each supervisor hour costs one dollar, and that the required number of supervisor hours needed to attain accuracy p is $\mu M(p)$. In other words, μ is a measure of how difficult it is to monitor. How does the optimal number of supervisor hours relate to μ? Differentiating and evaluating at the point p^* we have

$$\frac{d\mu M(p^*)}{d\mu} = M(p^*) + \mu M'(p^*)\frac{dp^*}{d\mu}. \tag{2.25}$$

We know from Section 2.3 that $dp^*/d\mu < 0$. Hence, the sign is ambiguous in general. However, we can say more. As μ tends to zero, the second term vanishes (at least if $M'(p)$ is bounded

[6] For the complementary effect, Prendergast refers to Athey and Stern (1997). It relies on the marginal cost of each instrument to be increasing in its quantity.

[7] This conclusion is based on the fact that monitoring and pay move in the same direction regardless of which parameter we change. The only additional assumption we need in order to take the prediction to data is that the parameters β, μ, γ are affiliated random variables – see Holmström and Milgrom (1994).

[8] Or, as Kruse (1992) puts it, the measure captures the quantity but not all aspects of the quality of supervision.

above), so in this case the number of supervisor hours goes up as monitoring becomes more difficult. On the other hand, as μ tends to infinity, p^* tends to zero and the first term vanishes. Indeed, when monitoring is sufficiently costly, it is not worthwhile doing any monitoring at all. Hence, over some interval the number of supervisor hours must decrease as monitoring becomes more difficult.

We have thus proved that the number of supervisor hours is a non-monotonic function of μ. It follows that there will also be a non-monotonic relationship between the wage and the number of supervisor hours. Very low wages and very high wages should both be associated with few hours of supervision. The highest wages should be paid when monitoring is so accurate that very few supervisor hours are needed; the lowest wages should be paid when supervision is so costly that the principal cuts back on supervision for this reason. The big difference between the two cases is that in the former jobs, workers are expected to work very hard, and in the latter jobs they are not expected to work much at all.

If we think of primary sector jobs as jobs with high wages, our theory says that monitoring is easier in the primary sector than in the secondary sector. On the contrary, the "technical" assumption that there are only two effort levels led Bulow and Summers (1986) to make exactly the opposite prediction. However, the leading example of a primary sector job offered by Bulow and Summers is Henry Ford's five dollar day, the pay policy introduced by the Ford Motor Company in 1914, which represented almost twice the wage that similar workers could be expected to earn elsewhere. Reportedly, productivity increased by about fifty per cent, a feature which squares well with the model offered here. The key question is: did the change in pay policy relate to changes in supervision technology, and if so how? The common view is that mechanization of production and standardization of tasks generally made it easier to deter shirking. According to our theory, then, the increase in pay could be due to easier monitoring.

Given the above observations, we should not be surprised that earlier studies have been unable to find a strong relationship between hours of monitoring and the level of pay. In the terminology of our model, most of the empirical work consists of regressions of the wage level w or the wage premium $w - w_0$ on the monitoring cost $\mu M(p)$. Our analysis suggests that there will be (at least) two values of w for every value of $\mu M(p)$. Hence, it is invalid to estimate any functional relationship with w as the dependent variable. On the other hand, the model does not in general rule out the possibility of estimating $\mu M(p)$ as a function of $w - w_0$. A natural direction for future research is to reexamine existing data with the monitoring cost rather than the wage level as a dependent variable, and allowing for the kind of non-linear relationship predicted by the theory.

Minimum monitoring restrictions

In some professions, employers are not free to decide the ratio of supervisors to workers. For example, Krueger and Groshen (1990) argue that regulation of working conditions makes the supervisor/staff ratio largely exogenous in the case of hospital nurses in the United States.[9] These authors go on to show that there is a negative empirical relationship between supervision and pay for hospital nurses. If effort is freely flexible, this finding contradicts our model.

Proposition 3 *Suppose the level of monitoring, p, is subject to binding regulations. Then the optimal wage, w^*, is an increasing function of p if and only if $U_{pw}(p, w^*) > 0$.*

PROOF: Suppose that regulations force p to be greater than the principal would have wanted. Hence, the principal has only one instrument: the wage. The solution to the principal's problem is given by the first-order condition $U_w(w^*, p) = 0$. Differentiating this condition gives

$$\frac{dw}{dp} = -\frac{U_{pw}(p, w^*)}{U_{ww}(p, w^*)}.$$

Since the denominator is negative by the problem's second-order condition, it follows that the wage should increase with monitoring if and only if $U_{pw} > 0$. ■

As we have argued above, the cross-derivative is indeed positive for reasonable specifications. If monitoring is in fact largely exogenous as the authors claim, the data used by Krueger and Groshen is immune to the criticisms raised above, because there will be a monotonic relationship between whatever is regulated and actual monitoring, p. Hence, their empirical finding rejects the present model.[10]

Direct measures of monitoring

[9] Note that Krueger and Groshen are primarily concerned with an endogeneity problem that we have ruled out here: the possible substitution between supervisors and workers in the production function.

[10] It is worth noting that Krueger and Groshen appear to be aware that scale effects could possibly overturn the prediction that monitoring and pay are substitutes. They write (page 138-S): "Thus, holding workers' effort level constant, the efficiency wage model predicts that increases in monitoring would be associated with lower wages." But they then proceed to test the hypothesis without any controls for worker effort.

Another paper which credibly identifies the quality of monitoring and not just the monitoring cost is Krueger's (1991) study of the fast–food industry. Krueger compares the level of pay in franchise outlets with the level of pay in company–owned outlets, finding that pay is lower in the former. He argues that there should be more monitoring in the franchise outlets, because there the residual claimant can monitor workers directly, and that this might explain the lower pay for non–supervisory staff in the franchise outlets. As shown in equation (2.13)[11], this conclusion is incorrect when the scale effect is taken into account.[12] Hence Krueger's finding contradicts the shirking model, rather than confirming it.

Final remarks

We have shown that a very natural generalization of the conventional shirking model completely overturns previous intuitions. This is due to the existence of a scale effect which outweighs the substitution effect that earlier work has focussed on. Moreover, much of the empirical literature fails to properly address the model's implications. The only two studies which do represent valid tests of the model without controlling for scale effects, namely Groshen and Krueger (1990) and Krueger (1991), reject its predictions. How can we explain this?

One possibility is that effort levels are in fact regulated; there is some maximum effort that an employer can ask for, and any attempt to exceed this level is prohibited. We think that this explanation is unconvincing for both hospital nurses

[11] Krueger also argues that it is the different contractual arrangements that explain why franchisees expend more effort monitoring their workers than do managers of company-owned restaurants. He writes (page 76): "An owner-manager of a franchise has a strong incentive to expend effort supervising and monitoring his workers because he receives the residual profit generated by the enterprise; whereas a manager of a company-owned establishment is usually not paid a share of the establishment's profit, and his actions are not perfectly observed by his principal, the parent company." In our model this may be more appropriately captured by different βs rather than different μs. Hence, for those believing in the argument above, this reference should be to equation (2.20) instead.

[12] The fact that Krueger only looks at two levels of monitoring does not affect the qualitative conclusion. It is straightforward to show that our results carry over to the case in which the level of monitoring is chosen from a finite set P of alternatives: When the principal chooses to jump to a higher level of monitoring, at which point the discreteness of P implies that the new p is "too high", he will also choose to raise the wage. This follows directly from Proposition 1.

and fast–food personnel. Hence, with respect to the studies by Groshen and Krueger (1990) and Krueger (1991) we maintain our view that scale effects are potentially present, and that these studies therefore reject rather than confirm the efficiency wage hypothesis—at least in its current form.

Another possibility is of course that Edgeworth complementarity of monitoring and pay in the present model is due to some of the simplifying assumptions that we have made. A more realistic monitoring technology or a multi–period extension of the model could conceivably lead to another outcome. These extensions await further study.

The apparent empirical failure of the efficiency wage theory to explain the monitoring/pay relationship begs the question whether other theories perform any better. Somewhat surprisingly, very few alternative theories exist. Holmström and Milgrom (1994) show that monitoring and pay/performance sensitivity are complements in their linear incentive model, a point which is also found in Milgrom and Roberts (1992, p226-7).[13] However, they do not study the covariance between monitoring and average pay, which is the central issue in the empirical literature that we address here. An alternative theory of the observed negative relationship between monitoring and average pay, which is put forward informally by Groshen and Krueger (1990), is that the return to monitoring is higher when workers have low skills. If this theory is correct, the negative relationship might disappear with better controls for worker skills.

For the moment, the jury is still out, but if it turns out that Edgeworth complementarity of monitoring and pay generalizes even further, then we would be forced to seriously question the empirical relevance of the shirking model.

[13] This model differs from the efficiency wage model in a number of respects. In the linear incentive model, agents are assumed to be risk averse, but to have unlimited liability. Thus, the mechanism behind the complementarity between monitoring and incentive intensity is quite different; more monitoring reduces the income risk, and thereby reduces the cost of strong incentives.

References

Arai, Mahmood (1994a): An Empirical Analysis of Efficiency Wages and Wage Dispersion, *Scandinavian Journal of Economics* 96, 31–50.

Arai, Mahmood (1994b): Compensating Wage Differentials versus Efficiency Wages: An Empirical Study of Job Autonomy and Wages, *Industrial Relations* 33, 249–262.

Athey, Susan and Scott Stern (1997): An Empirical Framework for Testing Theories About Complementarity in Organizational Design, Mimeo. MIT.

Bowles, Samuel (1985): The Production Process in a Competitive Economy: Walrasian, Neo–Hobbesian, and Marxian Models, *American Economic Review* 75, 16–35.

Braverman, Harry (1974): *Labor and Monopoly Capital,* New York: Monthly Review Press.

Bulow, Jeremy I. and Lawrence H. Summers (1986): A Theory of Dual Labor Markets with Applications to Industrial Policy, Discrimination, and Keynesian Unemployment, *Journal of Labor Economics* 4, 376–414.

Demougin, Dominique and Claude Fluet (1997): Monitoring versus Incentives: Substitutes or Complements? Mimeo. Université du Québec à Montréal.

Gordon, David (1990): Who Bosses Whom? The Intensity of Supervision and the Discipline of Labor, *American Economic Review, Papers and Proceedings* 80, 28–32.

Gordon, David (1994): Bosses of Different Stripes: A Cross–National Perspective on Monitoring and Supervision, *American Economic Review, Papers and Proceedings* 84, 375–379.

Groshen, Erica L. and Alan B. Krueger (1990): The Structure of Supervision and Pay in Hospitals, *Industrial and Labor Relations Review* 43, S134-S146.

Holmström, Bengt and Paul Milgrom (1994): The Firm as an Incentive System, *American Economic Review* 84, 972–991.

Krueger, Alan (1991): Ownership, Agency, and Wages, *Quarterly Journal of Economics* 106, 75–101.

Kruse, Douglas (1992): Supervision, Working–Conditions, and the Employer Size–Wage Effect, *Industrial Relations* 31, 229–249.

Leonard, Jonathan S. (1987): Carrots and Sticks: Pay, Supervision, and Turnover, *Journal of Labor Economics* 8, S135–S152.

Milgrom, Paul and John Roberts (1992): *Economics, Organization and Management,* (Englewood Cliffs, NJ: Prentice–Hall).

Neal, Derek (1993): Supervision and Wages across Industries, *Review of Economics and Statistics* 75, 409–417.

Prendergast, Canice (1999): The Provision of Incentives in Firms, *Journal of Economic Literature* Vol. XXXVII, 7–63.

Rebitzer, James B. (1993): Radical Political Economy and the Economics of Labor Markets, *Journal of Economic Literature* 31, 1394–1434.

Rebitzer, James B. (1995): Is There a Trade–Off between Supervision and Wages? An Empirical Test of Efficiency Wage Theory, *Journal of Economic Behavior and Organization* 28, 107–129.

Shapiro, Carl and Joseph E. Stiglitz (1984): Involuntary Unemployment as a Worker Discipline Device, *American Economic Review* 74, 433–444.

Appendix 1

Suppose the effort level is restricted to take only the values $e = 0$ and $e = \bar{e}$. Provided that the principal wants to implement $\bar{e}$, his maximization problem is

$$\underset{p,w}{Max}\ \beta B(\bar{e}) - w - \mu M(p)$$

subject to the incentive constraint

$$p(w - w_0) = \gamma C(\bar{e}).$$

Using the constraint to substitute for w reduces the maximization problem to

$$\underset{p}{Max}\ \beta B(\bar{e}) - w_0 - \frac{\gamma C(\bar{e})}{p} - \mu M(p).$$

Assuming an interior solution, the first order condition is

$$\frac{\gamma C(\bar{e})}{p^2} - \mu M'(p) = 0, \qquad (2.26)$$

which can be rewritten in terms of w as

$$\frac{w - w_0}{\gamma C(\bar{e})} - \mu M'\left(\frac{\gamma C(\bar{e})}{w - w_0}\right) = 0. \qquad (2.27)$$

Total differentiation of (2.26) and (2.27) gives us the formulas

$$C(\bar{e})d\gamma - p^2 M'(p)d\mu - \mu p\left[2M'(p) + pM''(p)\right]dp = 0 \quad (2.28)$$

and

$$\begin{aligned}\mu C(\bar{e})\left[M'\left(\frac{\gamma C(\bar{e})}{w - w_0}\right) + \frac{\gamma C(\bar{e})}{w - w_0}M''\left(\frac{\gamma C(\bar{e})}{w - w_0}\right)\right]d\gamma \\ +\gamma C(\bar{e})M'\left(\frac{\gamma C(\bar{e})}{w - w_0}\right)d\mu \\ -\left[2(w - w_0) + \mu\left(\frac{\gamma C(\bar{e})}{w - w_0}\right)^2 M''\left(\frac{\gamma C(\bar{e})}{w - w_0}\right)\right]dw = 0.\end{aligned} \qquad (2.29)$$

Using (2.28), it is easily seen that the effect of a change in μ on the monitoring accuracy, p, is

$$\frac{dp}{d\mu}\Big|_{e=\bar{e}} = -\frac{pM'(p)}{\mu\left[2M'(p) + pM''(p)\right]}$$

and using (2.29) we get the analogous effect on the wage, w

$$\frac{dw}{d\mu}\Big|_{e=\bar{e}} = \frac{(w - w_0)M'(p)}{\mu\left[2M'(p) + pM''(p)\right]} = \frac{w - w_0}{p}\frac{dp}{d\mu}.$$

Thus, this effect is always a substitution effect. And under the reasonable assumption that the denominator is positive in these expressions, we have the result that a decrease in μ will make the principal substitute wage for monitoring.

From the same formulas, (2.28) and (2.29), we find that the effect of a change in the parameter γ on p and w is

$$\frac{dp}{d\gamma}|_{e=\bar{e}} = \frac{C(\bar{e})}{\mu p\left[2M'(p) + pM''(p)\right]}$$

and

$$\frac{dw}{d\gamma}|_{e=\bar{e}} = \frac{C(\bar{e})\left[M'(p) + pM''(p)\right]}{p\left[2M'(p) + pM''(p)\right]} = \mu\left[M'(p) + pM''(p)\right]\frac{dp}{d\gamma}.$$

Thus here, given that $-pM''(p)/M'(p) < 1$ (inequality (2.22)), the effect is a complementary effect. Notice finally that there is no effect of a change in the parameter β on the equilibrium values of p and w, since β does not appear in either (2.26) or (2.27).

3 MONITORING AND PAY: GENERAL RESULTS

Abstract This paper considers the optimal incentives for motivating a risk neutral wealth constrained agent. In particular, monitoring and pay are shown to be complementary instruments under very general conditions, extending earlier results by Allgulin and Ellingsen (1998). The paper also proves that linear incentive schemes are strictly sub-optimal in this setting.

JEL CLASSIFICATION: J31, J41.
KEYWORDS: Monitoring, efficiency wages, incentive pay.

Standard analysis of the efficiency wage model argues that monitoring and pay are substitute instruments for motivating workers.[1] However, Allgulin and Ellingsen have recently demonstrated that this result hinges on unduly restrictive assumptions regarding workers' choice of effort - for example that there are only two possible effort levels. They extend the simple shirking model to a continuum of effort levels and characterize the profit maximizing levels of monitoring and pay. The equilibrium levels of monitoring and pay are characterized by two first-order conditions containing three general functions: the principal's benefit function of the agent's effort, the agent's cost of effort and the principal's cost of monitoring. In the analysis it is calculated how monitoring and pay vary according to multiplicative shifts to three functions: the profitability $B(e)$ (i.e. the principal's benefit from the agent's effort), the agent's cost of effort $C(e)$, and the principal's cost of monitoring, $M(p)$. It is shown that under reasonable assumptions, monitoring and pay are complementary instruments, i.e. a parameter change which causes an increase in the accuracy of monitoring also causes an increase in the level of pay and vice versa. This paper will demonstrate that the result is true for any monotonic shift, not just the multiplicative.

Two extensions of the model are also investigated. The first is the introduction of an ex ante individual rationality constraint for the agent in addition to the previous ex post limited liability constraint. This will typically be of interest if unemployed workers bid over each other with entrance fee offers for a job opportunity. The implication is that a binding ex ante individual rationality constraint generalizes the result even further.

The second extension allows the principal to credibly use a mixed strategy when monitoring the agent. In that case, any concave monitoring cost function can be replaced by another monitoring cost function that is cheaper and linear. Hence, the monitoring technology used by the principal is always weakly convex, implying that even this extension will make the result more general.

Finally, the optimal non-linear incentive scheme solution in the paper is compared to the best linear incentive scheme solution. The finding is that, for the principal, there is always a non-linear wage contract which strictly dominates any linear wage contract. This result may not be surprising, but the main

[1] See e.g. Shapiro and Stiglitz (1984) and Milgrom and Roberts (1992).

contribution of this section is to provide a common framework for efficiency wages and linear incentive wages, in which their different implications can be analyzed.

The paper is organized as follows. The next section sets up a more general version of the model proposed by Allgulin and Ellingsen (1998). The third section contains a complete analysis of this model. The fourth section introduces an ex ante individual rationality constraint into the model and discusses its implications when it is binding. In the section after that, the principal is allowed to use mixed strategies, and this is shown to strengthen the previous results. The following shows that the optimal incentive scheme used in the paper strictly dominates a linear incentive scheme. The last section concludes.

Model

A risk neutral principal employs a single risk neutral agent. The agent can exert effort $e \in \mathbf{R}_+$ which affects the principal's benefit, $B(e)$, at some cost to the agent, $C(e)$. The principal motivates the agent through a compensation contract $w(e)$, where w is the wage that the agent receives if the principal observes e.

The principal can observe and verify the agent's effort with probability p. This probability is affected by the principal's choice of monitoring technology. In order to attain probability p of observing the agent's effort, the principal has to pay $M(p)$. The following assumption regarding functional forms is made.

Assumption 1
(i) $B'(e) > 0,\ B''(e) \leq 0,$
(ii) $C(0) = 0,\ C'(e) > 0,\ C''(e) > 0,$
(iii) $M'(p) > 0.$

The ex post utility of the principal can now be written

$$U = B(e) - w - M(p), \qquad (3.1)$$

and the ex post utility of the agent is

$$V = w - C(e). \qquad (3.2)$$

A key assumption of the efficiency wage model is that there is a lower limit, $w_0 \in \mathbf{R}_+$, to the payment. The limit may be due either to legal rules or to a wealth constraint.

Since effort is not always observed, the compensation contract also needs to specify some payment $\bar{w} \in \mathbf{R}_+$ that the agent is to receive in this case. The agent is assumed to maximize the expected utility

$$\mathrm{E}[V] = pw(e) + (1-p)\bar{w} - C(e). \tag{3.3}$$

Suppose now that the principal will induce the agent to take the level of effort $\hat{e} \in \mathbf{R}_+$. Then the following incentive compatibility constraint must be satisfied for all e:

$$pw(\hat{e}) + (1-p)\bar{w} - C(\hat{e}) \geq pw(e) + (1-p)\bar{w} - C(e).$$

We see that any incentive compatible contract that implements $\hat{e}$ can, without loss to the principal, be replicated by a step function of the form $w(e) = w_0$ for $e < \hat{e}$ and $w(e) = w$ for $e \geq \hat{e}$. In other words, the principal sets an effort target, $\hat{e}$. The agent gets w if he meets or exceeds the target and the minimum payment w_0 otherwise.[2] Later, it is shown that this kind of contract strictly dominates a linear incentive scheme. Given the above contract, if an agent ever wants to deviate, he will deviate to $e = 0$. Thus, the incentive compatibility constraint becomes

$$p(w - w_0) \geq C(\hat{e}). \tag{3.4}$$

Finally, the assumption that an indifferent agent takes the action that the principal favors enables the principal to lower the wage down to the level where the incentive compatibility constraint becomes an equality. Inverting this equality, we obtain an expression for the actual effort which the agent will exert,

$$e(p,w) := C^{-1}((w - w_0)p). \tag{3.5}$$

Note that $\bar{w}$, the wage in the case that the effort level is not observed, is irrelevant for the agent's incentives. It is common to assume that $\bar{w} = w(\hat{e})$, and we follow this practice. One reason why the assumption is plausible is that if $\bar{w} < w(\hat{e})$, the principal would have had an incentive not to monitor (or to falsely claim that he has not monitored).

The principal's problem is to find a probability p and a wage w to maximize

$$U(p,w) = B(e(p,w)) - w - M(p) \tag{3.6}$$

subject to the constraints $w \geq w_0$ and $p \in [0,1]$. This is a straightforward maximization problem in two variables. Let

[2] As shown by Demougin and Fluet (1997), the optimality of effort targets is quite general. It is not an artefact of the simple monitoring technology.

a solution to this problem be denoted (p^*, w^*), and let $e^* := e(p^*, w^*)$ denote the associated effort level. The first–order conditions for the solution can then be written

$$p^* \frac{B'(e^*)}{C'(e^*)} - 1 \leq 0, \tag{3.7}$$

with equality if $w^* > w_0$, and

$$(w^* - w_0)\frac{B'(e^*)}{C'(e^*)} - M'(p^*) \geq 0, \tag{3.8}$$

with equality if $p^* < 1$.

Equation (3.7) tells us that the marginal benefit from increased effort will be larger than the marginal cost whenever the principal chooses to monitor imperfectly. The reason for the distortion away from the socially optimal effort level is that the agent must be paid a rent in order not to shirk. Also, equation (3.7) confirms that there must be a positive level of monitoring in order to induce any effort. (If $p^* = 0$, then $w^* = w_0$.)

Let U_{ij} denote the second derivative of U and let U^*_{ij} denote the second derivative of U evaluated at a solution (p^*, w^*). The second–order conditions are then $U^*_{ww} < 0, U^*_{pp} < 0$ and $U^*_{pp}U^*_{ww} - (U^*_{pw})^2 > 0$. To state the conditions in full, note that

$$U_{ww} = p^2 h(e), \tag{3.9}$$

$$U_{pp} = (w - w_0)^2 h(e) - M''(p), \tag{3.10}$$

$$U_{pw} = p(w - w_0)h(e) + \frac{B'(e)}{C'(e)}, \tag{3.11}$$

where

$$h(e) := \frac{B''(e)C'(e) - C''(e)B'(e)}{(C'(e))^3}.$$

Analysis

Method

The method to analyze the effects of a shift in one of the general functions is as follows. The first-order conditions (3.7) and (3.8) are used to create two equations characterizing the equilibrium values of w and p where, A) equation 1 is lacking one of the control variables and, B) equation 2 is lacking any form of the general function that is about to be shifted.

The trick is that we can now look at the shift as though it created a chain reaction. The equilibrium value of the remaining control variable in equation 1 is solely determined by this equation. A shift in the general function in equation 1 may change the equilibrium value of this control variable. Then, since equation 2 is lacking any form of the shifted function, the entire effect of the shift on the other control variable is forwarded by the first control variable according to equation 2. Hence we only have to differentiate equation 2 in order to find out the condition for the control variables to be positively related.

Main results

(3.7) and (3.8) together tell us that

$$(w^* - w_0) = pM'(p^*). \tag{3.12}$$

This equation can be used to substitute for w^* in (3.7) to get the equation

$$p^* \frac{B'(C^{-1}((p^*)^2 M(p^*)))}{C'(C^{-1}((p^*)^2 M(p^*)))} = 1, \tag{3.13}$$

which solely determines p^*. The equations (3.12) and (3.13) together characterize the equilibrium values of w and p. We see that neither the principal's benefit function, $B(e)$, nor the agent's cost function, $C(e)$, appears in any form in equation (3.12). Therefore, any effect that any change in these functions will have on w^* will be forwarded by p^* according to equation (3.12). By differentiating (3.12) we obtain

$$\frac{dw^*}{dp^*} = p^* M''(p^*) + M'(p^*),$$

which is positive if and only if

$$-p^* \frac{M''(p^*)}{M'(p^*)} < 1.$$

Hence, we can conclude that monitoring and pay are complementary instruments according to *any* shift in the principal's benefit function, $B(e)$, or the agent's cost of effort function, $C(e)$, as long as the condition above is fulfilled.

To analyze the effects of a shift in the principal's monitoring cost function, first define

$$f(p^*) = p^* M'(p^*).$$

Using this and inverting (3.12) now yields

$$p^* = f^{-1}(w^* - w_0),$$

which we can use to substitute for p^* back in equation (3.7) to get the equation

$$f^{-1}(w^* - w_0) \frac{B'(C^{-1}((w^* - w_0) f^{-1}(w^* - w_0)))}{C'(C^{-1}((w^* - w_0) f^{-1}(w^* - w_0)))} = 1 \tag{3.14}$$

which solely determines w^*. This equation together with (3.7) characterizes the equilibrium values of w and p and (3.7) is lacking any form of the monitoring cost function, $M(p)$. Differentiation of (3.7) yields

$$\frac{dw^*}{dp^*} = -\frac{U^*_{pw}}{U^*_{ww}}$$

which is positive if and only if

$$U^*_{pw} > 0.$$

Thus we can conclude that monitoring and pay are complementary instruments according to *any* shift in the principal's monitoring cost function, $M(p)$, as long as monitoring and pay are Edgeworth complements.

The above analysis can be summarized in the following two propositions,

Proposition 1 *According to any shift in the principal's benefit function, $B(e)$, or the agent's cost of effort function, $C(e)$, monitoring and pay are complementary instruments if and only if $-pM''(p^*)/M'(p^*) < 1$.*

Proposition 2 *According to any shift in the principal's monitoring cost function, $M(p)$, monitoring and pay are complementary instruments if and only if $U^*_{pw} > 0$.*

Binding individual rationality constraint

So far it has implicitly been assumed that the agent is willing to be hired by the principal, given the equilibrium solution. The agent will receive a higher wage than w_0 in any solution, and the interpretation of w_0 as the market price of labor has justified the previous ignorance of a participation constraint.

However, there are two important objections against this reasoning. Firstly, one of the foundations of efficiency wages is that there is incomplete competition in the labor market, due to for example the non-homogeneity of firms, or the transaction costs of being fired for the agent. Hence, w_0 may be lower than the agent's outside option before the relationship starts. Secondly, it has frequently been argued that the agent's (ex ante) individual rationality constraint *must* be binding because otherwise the agent would just offer to pay an entrance fee for the job opportunity, such that the individual rationality constraint is binding anyway.[3]

[3] See for example Carmichael (1985) and (1990).

Below, in addition to the limited liability constraint, w_0, on the maximum punishment that the principal can impose on the agent ex post, an ex ante individual rationality constraint is introduced into the model,

$$w - C(e) \geq \bar{V}, \tag{3.15}$$

where the ex ante reservation utility, $\bar{V}$, is assumed to be higher than the ex post limited liability constraint, w_0. Clearly, there are two distinct cases. If the solution in the above analysis yields utility for the agent that is higher than the reservation utility, i.e. if

$$w^* - C(e^*) \geq \bar{V}, \tag{3.16}$$

then the participation constraint is not binding and the analysis remains intact. On the other hand, if this is not the case, the following analysis will apply.

Rearranging and assuming that the participation constraint, (3.15), is binding yields the expression for the lowest possible monitoring accuracy,

$$p = \frac{w - \bar{V}}{w - w_0}. \tag{3.17}$$

Using this to substitute for the monitoring accuracy and the incentive constraint, (3.5), to substitute for the effort level, yields the following maximization problem for the principal.

$$\underset{w}{Max} \;\; U(w) = B(C^{-1}(w - \bar{V})) - w - M(\frac{w - \bar{V}}{w - w_0}) \tag{3.18}$$

subject to the constraints $w \geq w_0$ and $\bar{V} > w_0$. Let the solution to this problem be denoted w^*, and let $e^* := e(w^*)$, $p^* := p(w^*)$ denote the associated effort level and monitoring accuracy respectively. Assuming an interior solution, the first–order condition for the solution can be written

$$\frac{B'(e^*)}{C'(e^*)} - 1 - M'(p^*)\frac{\bar{V} - w_0}{(w^* - w_0)^2} = 0. \tag{3.19}$$

The optimal wage is solely determined by this condition and neither the principal's benefit function, $B(e)$, the principal's monitoring cost function, $M(p)$, or the agent's cost function, $C(e)$, appears in any form in the expression (3.17). Hence, differentiation of (3.17) in equilibrium will reveal how the wage and the monitoring accuracy vary together according to any shift in these functions,

$$\frac{dp^*}{dw^*} = \frac{\bar{V} - w_0}{(w^* - w_0)^2}.$$

This is clearly always positive.

Proposition 3 *If the agent's individual rationality constraint is binding, then according to any shift in the principal's benefit function, $B(e)$, the principal's monitoring cost function, $M(p)$, or the agent's cost of effort function, $C(e)$, monitoring and pay are complementary instruments.*

Mixed strategies

With a non-binding individual rationality constraint, the analysis section demonstrated that monitoring and pay are complements according to *any* shift in the principal's benefit function, $B(e)$, or the agent's cost function, $C(e)$, as long as the condition

$$-p^* \frac{M''(p^*)}{M'(p^*)} < 1 \tag{3.20}$$

is fulfilled. Thus a sufficient, but not necessary, condition is that $M(p)$ is convex. Generally, some increasing returns to monitoring may be allowed for, but when the returns to monitoring increase sufficiently fast the result may be overturned.

However, if the principal can use mixed strategies there will never be increased returns to monitoring. More precisely, if the principal is able to randomize between the monitoring accuracy $p = 0$ and $p = 1$, he will optimally do so whenever the monitoring cost function, $M(p)$, is concave. As illustrated in Figure 5, by this randomization he will create a new and cheaper monitoring cost function,

$$\dot{M}(q) = (1 - q)M(0) + qM(1), \tag{3.21}$$

where $q \in [0, 1]$, i.e. is linear in q.

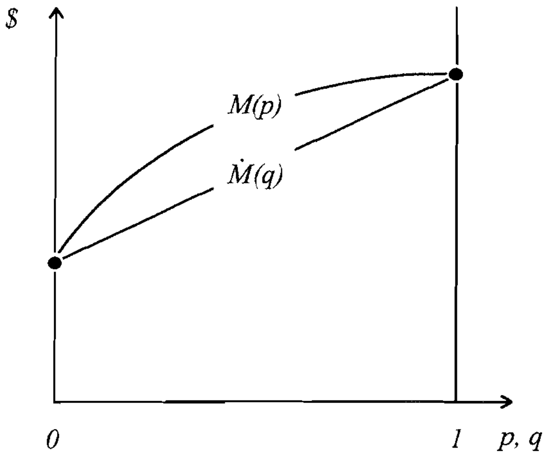


Figure 5

The principal is able to do this because the agent's expected utility (3.3)

$$\mathrm{E}[V] = pw(e) + (1-p)\bar{w} - C(e) \qquad (3.22)$$

is linear in the detection probability p. The agent's expected utility when the principal randomizes in this manner is

$$\mathrm{E}[V] = q(1*w(e)+0*\bar{w})+(1-q)(0*w(e)+1*\bar{w})-C(e), \quad (3.23)$$

which can be simplified to

$$\mathrm{E}[V] = qw(e) + (1-q)\bar{w} - C(e). \qquad (3.24)$$

Hence, the agent will react in exactly the same way towards the randomization variable q as he did towards the detection probability p.

Of course, one could argue that there may be credibility problems for the principal: it could be tempting for the principal to pretend that the randomization variable q is high when it is not. Technically, this problem should arise even with the original monitoring function, but this can be solved by e.g. the assumption that the monitoring technology able to detect at probability p^* is observable and installed before the agent makes his effort. However, this objection against the use of mixed strategies for the principal falls if he is able to create a mechanism that commits him to a certain q. Thus in a final proposition it can be concluded that:

Proposition 4 *According to any shift in the principal's benefit function, $B(e)$, or the agent's cost of effort function, $C(e)$, monitoring and pay are complementary instruments if the principal can credibly use mixed strategies.*

Assume that the principal is interested in implementing a certain effort level, $\hat{e}$. Below, we shall now see that implementation of $\hat{e}$ is (strictly) more costly using a linear incentive scheme than using the optimal non-linear scheme.[4]

If the principal is free to use a non-linear incentive scheme, he will simply solve the minimization problem

$$\underset{w,p}{Min}\ \ w + M(p) \tag{3.25}$$

subject to the incentive compatibility constraint

$$p(w - w_0) = C(\hat{e}). \tag{3.26}$$

Substitution of the constraint into the minimization problem yields a problem with only one variable, for example

$$\underset{p}{Min}\ \ w_0 + \frac{C(\hat{e})}{p} + M(p) \tag{3.27}$$

with the first order condition

$$(p^*)^2 M'(p^*) = C(\hat{e}) \tag{3.28}$$

if an interior solution is assumed.

If the principal is restricted to using a linear incentive scheme, the new restriction is introduced that the wage must be of the form $\alpha_0 + \alpha_1 e$ if the agent's effort is observed. If the agent's effort is not observed, the agent receives the wage $\alpha_0 + \alpha\hat{e}$. Furthermore, the limited liability constraint on the agent restricts the principal to setting α_0 to at least w_0: the wage which the agent gets if he makes no effort and this is detected. The principal does not give anything away for free so he will optimally set $\alpha_0 = w_0$. Hence, the agent's ex ante wage will now be

$$w(e) = w_0 + \alpha(pe + (1 - p)\hat{e}). \tag{3.29}$$

Thus the agent will face the following maximization problem:

$$\underset{e}{Max}\ \ V(e) = w_0 + \alpha(pe + (1 - p)\hat{e}) - C(e) \tag{3.30}$$

with the first-order condition

$$\alpha p = C'(e). \tag{3.31}$$

[4] This result contrasts with the conventional result in many principal-agent models, where linear incentive schemes are optimal. The difference is accounted for by limited liability.

Thus, the principal solves the minimization problem (3.25) subject to the constraints

$$w = w_0 + \alpha \hat{e} \tag{3.32}$$

and

$$\alpha p = C'(\hat{e}). \tag{3.33}$$

Substitution of the constraints into the minimization problem yields a problem with only one variable, for example

$$\underset{p}{Min} \; w_0 + \frac{C'(\hat{e})\hat{e}}{p} + M(p) \tag{3.34}$$

with the first order condition

$$(p^{**})^2 M'(p^{**}) = C'(\hat{e})\hat{e} \tag{3.35}$$

for an interior solution.

By the assumptions $C(0) = 0$ and $C''(e) > 0$ we know that $C(e) < eC'(e)$. This can easily be seen in Figure 6. There, $C(\hat{e}) = \int_0^{\hat{e}} C'(e)de$ is represented by the area B, and $\hat{e}C'(\hat{e})$ is represented by the strictly larger area $A + B$.

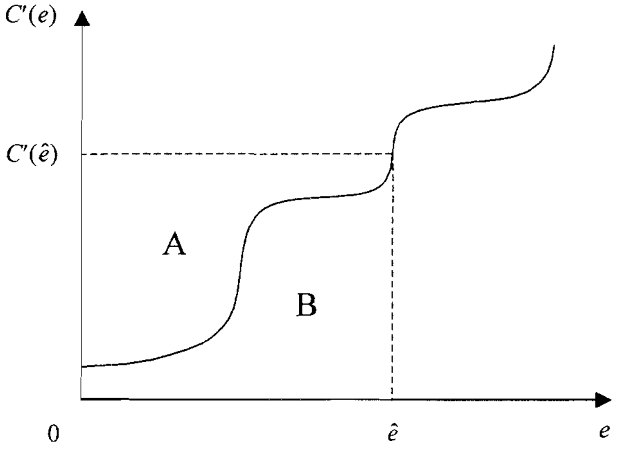


Figure 6

Thus (3.28) and (3.35) together tell us that

$$(p^*)^2 M'(p^*) < (p^{**})^2 M'(p^{**}), \tag{3.36}$$

implying that for any level of effort the principal wants to implement he will monitor more accurately under a linear incentive scheme if $M''(p) \geq 0$.

It is now easy to compare the least resources required to implement the effort level, $\hat{e}$, using a non-linear contract and a linear contract respectively. If one substitutes for the equilibrium wage, it is clear that

$$w_0 + p^* M'(p^*) + M(p^*) < w_0 + p^{**} M'(p^{**}) + M(p^{**}). \quad (3.37)$$

Thus, the principal can implement any level of effort under the best non-linear incentive scheme using less resources on incentives for the agent than under the best linear incentive scheme.

Proposition 5 *For any level of effort the principal wants to implement, he will monitor more accurately if he is restricted to using linear incentive schemes than if he is not, if $M''(p) \geq 0$ or if the principal can credibly use mixed strategies.*

If $M''(p) < 0$, the optimal monitoring accuracy under a non-linear incentive scheme, p^*, may be higher than the optimal monitoring accuracy under a linear incentive scheme, p^{**}. But even if this is the case, it is more costly for the principal to implement $\hat{e}$ if the wage contract is restricted to be linear. To see this, first note that $p^* > p^{**}$ together with (3.36) imply that

$$p^* M'(p^*) < p^{**} M'(p^{**}), \quad (3.38)$$

or alternatively expressed,

$$\int_{p^{**}}^{p^*} (M'(p) + pM''(p))\, dp < 0. \quad (3.39)$$

Assume now that the inequality (3.37) does not hold, i.e. that

$$\int_{p^{**}}^{p^*} (2M'(p) + pM''(p))\, dp \geq 0. \quad (3.40)$$

It is easily seen that the following must then be true:

$$\int_{p^{**}}^{p^*} M'(p) > 0. \quad (3.41)$$

But this is impossible since it contradicts the assumption that $M''(p) < 0$. Evidently, the principal is always harmed by restricting the wage contract to be linear. Thus a final proposition can be stated:

Proposition 6 *The set of optimal linear contracts and the set of optimal non-linear contracts are disjunct.*

Concluding remarks

Allgulin and Ellingsen (1998) demonstrate that a natural generalization of the conventional shirking model of efficiency wages completely overturns previous intuitions. The current paper has demonstrated the robustness of this result to a variety of generalizations. In addition, it has shown that linear incentive schemes are suboptimal.

The most obvious argument against the latter result is that it hinges on the assumption of risk neutral workers. In practice, linear incentive schemes may be preferable. But the contribution here is not merely the result; it is rather the framework. Previously, efficiency wages and linear wage contracts have been described in different settings. Standard shirking models of efficiency wages, such as Shapiro and Stiglitz (1984), assume imperfect monitoring and limited liability but no risk aversion. On the other hand, models of linear wage contracts, such as Holmström (1994) (or Milgrom and Roberts (1992, p226-7)), assume imperfect monitoring, risk aversion and unlimited liability. To be able to compare the two different types of wage contracts, the underlying assumptions for them must of course be the same. A first step towards a common framework is taken here, as both the conventional efficiency wage contract and the linear wage contract are modelled in a world of imperfect monitoring, limited liability and no risk aversion. The next step, which clearly merits an investigation, is a common framework with risk averse workers. A conjecture is that, when both limited liability and risk aversion are present, the optimal wage contract is neither an efficiency wage step function nor a linear wage, but rather a smooth non-linear wage contract.

References

Allgulin, Magnus and Tore Ellingsen (1998): Monitoring and Pay, *Working Paper Series in Economics and Finance* No. 245, Stockholm School of Economics.

Carmichael, H. Lorne (1985): Can Unemployment be Involuntary?: Comment, *American Economic Review* December, 75, 1213-1214.

Carmichael, H. Lorne (1990): Efficiency Wage Models of Unemployment - One View, *Economic Inquiry* April, 28, 269-295.

Demougin, Dominique and Claude Fluet (1997): Monitoring versus Incentives: Substitutes or Complements?, Mimeo, Université du Québec à Montréal.

Holmström, Bengt and Paul Milgrom (1994): The Firm as an Incentive System, *American Economic Review* 84, 972–991.

Milgrom, Paul and John Roberts (1992): *Economics, Organization and Management,* (Englewood Cliffs, NJ: Prentice–Hall).

Shapiro, Carl and Joseph E. Stiglitz (1984): Involuntary Unemployment as a Worker Discipline Device, *American Economic Review* 74, 433–444.

4 LIMITED LIABILITY AND DYNAMIC INCENTIVES

Abstract If efficiency wages really exist, as proposed by Shapiro and Stiglitz (1984), why do we not see more job purchases? A conventional answer is that with multiple periods, low pay in initial periods serves as an implicit payment (Lazear (1981)). This paper presents a formal analysis of this issue. A major result is that the per period worker rents associated with efficiency wages are inversely related to the number of periods, but are never zero. The paper also discusses how remaining worker rents can be eliminated by implicit bonds, such as firm-specific human capital investments.

JEL CLASSIFICATION: J31, J41.
KEYWORDS: Monitoring, efficiency wages, implicit bond, entrance fee.

Introduction

The objection against the shirking version of the efficiency wage model that most frequently recurs in the literature is that unemployed workers should offer to pay entrance fees for job opportunities, and such fees are very seldom observed. The best known paper which presents a shirking efficiency wage model is Shapiro and Stiglitz (1984). Indeed, Shapiro and Stiglitz were aware of this weakness and had already defended their model against the entrance fee criticism in the original paper. They pointed out the two most obvious arguments against job purchases. The first is that workers may simply be unable to pay the entrance fee. The idea is that workers are wealth constrained due to imperfect capital markets. The second is that the resulting employment contract is not self enforcing. The firm would have an incentive to claim that a worker shirks, fire him once the fee is paid, and sell the employment opportunity again.

However, many economists have not been convinced by these explanations of the lack of entrance fee observations. Carmichael (1985), in his reply to Shapiro and Stiglitz and Carmichael (1990), argues that the presence of an imperfect capital market, or the moral hazard problem, is not sufficient to prevent the emergence of entrance fees. For example, the moral hazard problem can be solved by the introduction of a third party. Moreover, in the Shapiro Stiglitz model, the firm and the worker do not know when the relationship is to end. Bull (1985, 1987) argues that this is unrealistic and continues to show that unemployment will not be sufficient to generate the surplus required for a self-enforcing contract in a model with a last period. As a consequence, the efficiency wage theory has lost advocates to a related theory, referred to as delayed payment/bonding contracts. In these contracts, firms initially pay wages below the alternative wage and later pay wages above the alternative wage to discourage shirking when monitoring is imperfect. The important point is that they are protected against the entrance fee criticism since they do not alter the present value of compensation from the first-best, full-information level.

The main purpose of this paper is to show how a firm should adjust the level and timing of compensation in the best possible way. Incentives can be constructed by both efficiency wages and back-loaded compensation, and it turns out that a rational firm will use both of them. It will be demonstrated that in a self-enforcing efficiency wage model with a finite number of periods and wealth constrained workers, the shape of the optimal wage path may look very much like the one derived from a delayed payment/bonding contract, such as for example Lazear (1981) and Becker and Stigler (1974). It will typically be constituted by a high wage in the last period and a lower and stationary wage for all preceding periods. The worker rents associated with efficiency wages still exist, but they are strongly diminished by the length of the relationship. This result is perhaps most remarkable for not fully eliminating worker rents; Lazear (1995, page 71) and others have claimed that delayed payment is a perfect substitute for entrance fees.

Bonding contracts and entrance fees might in principle eliminate the remaining worker rents. However, both bonds and entrance fees are inferior means of extracting worker rents. Investment in firm-specific human capital solves the moral hazard problem more efficiently. (With capital market imperfection, the firm might also be able to extract a greater portion of worker rents in this way.) Shapiro and Stiglitz briefly mention that when the cost of losing job-specific human capital is substantial, workers may have an incentive to exert effort even under conditions of full employment. That is true even in this paper. The point is that a firm-specific human capital investment increases the worker's liability and hence increases the firm's profit at the investing worker's expense. Unemployed workers or workers employed by other firms bid over each other for a job opportunity with higher education levels. This may go on until the hired worker is totally extracted. The reason why bids are in education levels and not in money (bond or entrance fee) is simply that it solves the moral hazard problem in the simplest possible way: if the employer fires the worker he will lose the firm-specific human capital investment together with the worker.

The model is based on the static efficiency wage model first presented in Allgulin and Ellingsen (1998), and later revised in Allgulin (1999), which contrary to Shapiro and Stiglitz's model allows the workers' effort levels to be adjusted in a continuous fashion. A secondary purpose of the paper has been to investigate whether the result from this model, which will be referred to as the static model throughout the paper, carries over to the dynamic case. The finding is that it does; the result is even strengthened by an increase in the number of periods that the relationship between the worker and the firm will last.

Two extensions are also discussed. The first makes the firm able to commit itself to future control variables. The solution could be interpreted as the optimal explicit bonding contract, and the wage path is simply as much payment as possible postponed to the last period and, as a consequence, the lowest possible wage in all preceding periods. The second briefly discusses how a worker's commitment to saving or educating himself at work can be an alternative to bonds or pre-work education, if he is too wealth-constrained for the latter.

The paper is organized as follows. The next section sets up the general model, derives the optimal contract and analyzes how the results from the static model carry over to a dynamic extension. Welfare effects are discussed in the third section. The fourth section, explains how bonding contracts are captured by the model and how firm-specific human capital investments can act as implicit bonds. The following section briefly discusses some extensions and finally, the last section concludes.

Model

A risk-neutral employer employs a single risk-neutral employee. The employee can exert effort $e_t \in \mathbf{R}_+$ yielding the benefit $B(e_t)$ to the employer at some cost $C(e_t)$ to the employee, where $t \in \{1, 2, .., T-1, T\}$ and T is the potential number of periods the relationship between the employer and the employee can last. The employer motivates the employee through a compensation contract $w(e_t)$. The employer has the ability to observe and verify the exerted level of effort e_t with probability p_t, and he can indirectly choose this detection probability through his choice of resources devoted to monitoring, $M(p_t)$. The benefit, cost and monitoring cost functions are assumed to have the plausible properties: $B'(e_t) > 0$, $B''(e_t) < 0$, $C(0) = 0$, $C'(e_t) > 0$, $C''(e_t) > 0$ and $M'(p_t) > 0$. Both the employer and the employee discount future utilities using a common discount factor δ.

Define the employer's ex post utility as

$$U = \sum_{t=0}^{T-1} \delta^t (B(e_t) - w_t - M(p_t)). \tag{4.1}$$

The employee's ex post utility in the last period T is

$$V_T = w_T - C(e_T), \tag{4.2}$$

and the total value of the present and future periods for him at any other time is

$$V_t = w_t - C(e_t) + \delta V_{t+1} \quad \text{for } t \in \{1, .., T-1\}. \tag{4.3}$$

There is a lower limit $w_0 \in \mathbf{R}_+$, to the payment. The limit may be due either to legal rules or to a wealth constraint. Legal rules also make it impossible for the employer to fire the employee if he fulfills his task, and to exchange him for another employee. Since effort is not always observed, the compensation contract also needs to specify some payment $\bar{w}_t \in \mathbf{R}_+$ that the employee is to receive in this case. Furthermore, the employer cannot commit himself not to change his control variables in future periods. On the other hand, he can commit not to rehire the employee once he has been fired. This is a crucial assumption and it implies that the employee's decisions, and hence the equilibrium values of the employer's control variables in different periods, are different. Without this assumption the result would just be the static model's outcome repeated T times.

Suppose the employer wants to induce some level of effort $\hat{e}_t$. The employee's only interest is to maximize his expected utility; therefore for all $t \in \{1, ..T-1\}$ the following incentive compatibility (IC) constraints must be satisfied for all e_t:

$$\begin{aligned} p_t(w_t(\hat{e}_t) + \delta V^*_{t+1}) + (1-p_t)(\bar{w}_t + \delta V^*_{t+1}) - C(\hat{e}_t) \geq \\ p_t(w_t(e_t) + \sum_{i=1}^{T-t} \delta^i w_0) + (1-p_t)(\bar{w}_t + \delta V^*_{t+1}) - C(e_t), \end{aligned} \quad (4.4)$$

where V^*_{t+1} denotes the value of future periods for the worker given the (optimal) incentive scheme. For $t = T$, the incentive compatibility constraint becomes

$$\begin{aligned} p_T w_T(\hat{e}_T) + (1-p_T)\bar{w}_T - C(\hat{e}_T) \geq \\ p_T w_T(e_T) + (1-p_T)\bar{w}_T - C(e_T). \end{aligned} \quad (4.5)$$

As shown in Demougin and Fluet (1997), any incentive-compatible contract that implements $\hat{e}_t$ can, without loss to the employer, be replicated by a step function of the form $w_t(e_t) = w_0$ for $e_t < \hat{e}_t$ and $w_t(e_t) = w_t$ for $e_t \geq \hat{e}_t$. In other words, the employer sets an effort target $\hat{e}_t$. The employee gets w_t if he meets or exceeds the target and the minimum payment w_0 otherwise. If the employee deviates under this contract he will choose the effort level $e_t = 0$.

Finally, we impose the standard assumption that an indifferent employee takes the action that the employer favors. Thus, the employer can lower the wage down to the level where the incentive compatibility constraints become equalities,

$$p_t(w_t - w_0 + \delta V^*_{t+i} - \sum_{i=1}^{T-t} \delta^i w_0) = C(\hat{e}_t) \quad (4.6)$$

for $t \in \{1, .., T-1\}$ and

$$p_T(w_T - w_0) = C(\hat{e}_T). \quad (4.7)$$

Note that $\bar{w}_t$, is irrelevant for the employee's incentives. It is common to assume that $\bar{w}_t = w_t(\hat{e}_t)$, and this practice is followed. One reason why the assumption is plausible is that if $\bar{w}_t < w(\hat{e}_t)$, the employer would have had an incentive not to monitor (or to falsely claim that he has not monitored).

Using the incentive compatibility constraints (4.6) and (4.7), starting at period T, the effort level in each period can now be computed. The highest possible effort level the employer is able to implement as a function of monitoring and pay in the last period is naturally the same as in the static case,

$$e_T = C^{-1}\left(p_T(w_T - w_0)\right). \tag{4.8}$$

But for all other periods it is cheaper for the employer to implement a given effort level, since the employee is not only threatened by being fired but also by never being rehired by the employer. Thus we have the following expressions for the highest possible effort levels in each of these periods as a function of monitoring and pay:

$$e_t = C^{-1}\left(p_t(w_t - w_0 + \delta V^*_{t+1} - \sum_{i=1}^{T-t} \delta^i w_0)\right). \tag{4.9}$$

These expressions will be used to substitute for e_t in the employer's utility function. Hence he will face T maximization problems to choose probabilities p_t and transfers w_t that solve

$$\underset{p_t, w_t}{Max} \;\; B(e_t) - w_t - M(p_t) + \sum_{i=1}^{T-t} \delta^i (B(e_{t+i}) - w_{t+i} - M(p_{t+i}))$$

subject to the constraints $w_t \geq w_0$ and $p_t \in [0, 1]$ for all $t \in \{1, .., T\}$. Denote the solutions to these problems (p^*_t, w^*_t), and let $e^*_t := e_t(p^*_t, w^*_t)$ be the corresponding effort.

Assuming an interior solution, first order conditions are

$$\frac{B'(e^*_t)}{C'(e^*_t)} p^*_t - 1 = 0 \tag{4.10}$$

for $t \in \{1, .., T\}$,

$$\frac{B'(e^*_T)}{C'(e^*_T)}(w^*_T - w_0) - M'(p^*_T) = 0, \tag{4.11}$$

and

$$\frac{B'(e^*_t)}{C'(e^*_t)}(w^*_t - w_0 + \delta V^*_{t+1} - \sum_{i=1}^{T-t} \delta^i w_0) - M'(p^*_t) = 0 \tag{4.12}$$

for $t \in \{1, .., T-1\}$.

Let U_{tij} denote the second derivative of U_t and let U^*_{tij} denote the second derivative of U_t evaluated at a solution (p^*_t, w^*_t). The second-order conditions are then $U^*_{tw_tw_t} < 0, U^*_{tp_tp_t} < 0$ and

$U^*_{tp_tp_t}U^*_{tw_tw_t} - (U^*_{tp_tw_t})^2 > 0$. To state the conditions in full, note that

$$U_{tw_tw_t} = (p_t)^2 h(e_t) \quad \text{for } t \in \{1,..,T\}, \tag{4.13}$$

$$U_{Tp_Tp_T} = (w_T - w_0)^2 h(e_T) - M''(p_T), \tag{4.14}$$

$$U_{Tp_Tw_T} = p_T(w_T - w_0)h(e_T) + \frac{B'(e_T)}{C'(e_T)}, \tag{4.15}$$

$$U_{tp_tp_t} = (w_t - \sum_{i=0}^{T-t} \delta^i w_0 + \delta V^*_{t+1})^2 h(e_t) - M''(p_t) \tag{4.16}$$

for $t \in \{1,..,T-1\}$, and

$$U_{tp_tw_t} = p_t(w_t - \sum_{i=0}^{T-t} \delta^i w_0 + \delta V^*_{t+1}) h(e_t) + \frac{B'(e_t)}{C'(e_t)} \tag{4.17}$$

for $t \in \{1,..,T-1\}$, where

$$h(e_t) := \frac{B''(e_t)C'(e_t) - C''(e_t)B'(e_t)}{(C'(e_t))^3}. \tag{4.18}$$

Determination of the optimal contract

The first order-conditions (4.10), (4.11) and (4.12) together tell us that

$$w^*_T = w_0 + p^*_T M'(p^*_T) \tag{4.19}$$

and

$$w^*_t = w_0 + p^*_t M'(p^*_t) - \delta V^*_{t+1} + \sum_{i=1}^{T-t} \delta^i w_0 \tag{4.20}$$

for $t \in \{1,..,T-1\}$. Using these expressions to substitute for w^*_t in the incentive constraints (4.8) and (4.9) yields

$$e^*_t = C^{-1}\left((p^*_t)^2 M'(p^*_t)\right) \tag{4.21}$$

for $t \in \{1,..,T\}$. With these expressions, it is now possible to substitute for e^*_t in the first-order condition (4.10) to get an equation that solely determines p^*_t for $t \in \{1,..,T\}$,

$$\frac{B'\left(C^{-1}\left((p^*_t)^2 M'(p^*_t)\right)\right)}{C'\left(C^{-1}\left((p^*_t)^2 M'(p^*_t)\right)\right)} p^*_t - 1 = 0. \tag{4.22}$$

Evidently, the equation characterizing the optimal monitoring accuracy is time independent and we have

$$p^*_t = p^*_{t+1} = p^*, \tag{4.23}$$

where p^* is the optimal monitoring accuracy in the static model. This in turn leads directly to another result. If we drop the time index on the monitoring accuracy (which we are now allowed to do), and look back at the expression (4.21), we can easily conclude that even the optimal effort level is time independent. Thus

$$e_t^* = e_{t+1}^* = e^*, \tag{4.24}$$

where e^* is the optimal effort level in the static model.

The time independency of the monitoring accuracy (4.23), together with the expression for the last period's wage (4.19), implies that the optimal wage in the last period is the same as the optimal wage in the static model, w^*, i.e. that

$$w_T^* = w^*. \tag{4.25}$$

To investigate the wage levels in other periods, first note that from the employee's ex post utility and the incentive constraint respectively in the last period, (4.2) and (4.7), together with the time independency of the monitoring accuracy, (4.23), we have:

$$V_T^* = (1 - p^*)(w_T^* - w_0) + w_0. \tag{4.26}$$

This utility can be used together with the employee's ex post utilities and incentive-compatibility constraints for other periods, (4.3) and (4.9), to solve backwards for an expression of V_{t+1}:

$$V_{t+1}^* = (1 - p^*)(w_T^* - w_0) + \sum_{i=0}^{T-t-1} \delta^i w_0. \tag{4.27}$$

This in turn substituted into the expression for the wage in other periods than the last period, (4.20), together with the expression for the wage in the last period, (4.19), finally yield the following expression for the employee's wage in all periods but the last one:

$$w_t^* = (1 - \delta(1 - p^*))w_T^* + \delta(1 - p^*)w_0 \tag{4.28}$$

for $t \in \{1, .., T - 1\}$. The above results can be summarized in the following proposition:

Proposition 1 *The optimal wage path is constituted by the equilibrium wage of the static model in the last period and a lower and stationary wage for all preceding periods. Both wage levels are independent of the length of the relationship. The optimal effort level and the optimal monitoring accuracy are stationary and have the same values as in the static model.*

The wage path is illustrated in Figure 7.

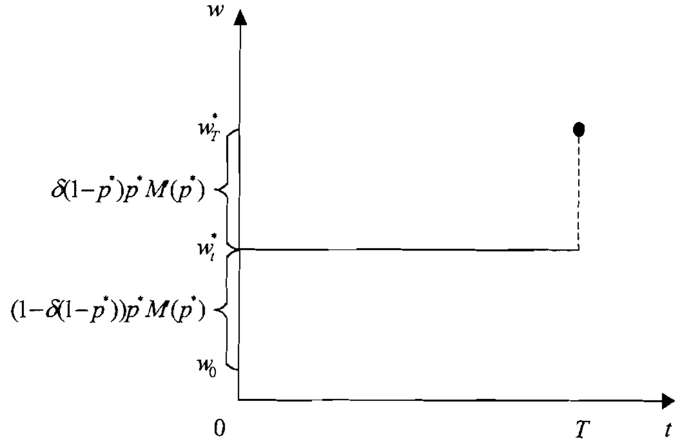


Figure 7

Here one can clearly see that, with no discounting of future periods ($\delta = 1$), all monetary incentives lie in the last period. The intuition is that worker rents in a preceding period exist only to compensate for the depreciation of future rents during that period.

The result implies that all second-order conditions are fulfilled if and only if the second-order conditions for the last period are fulfilled. Thus it is sufficient that

$$U^*_{Tw_Tw_T} < 0, U^*_{Tp_Tp_T} < 0$$

and

$$U^*_{Tp_Tp_T}U^*_{Tw_Tw_T} - (U^*_{Tp_Tw_T})^2 > 0,$$

where

$$U_{Tw_Tw_T} = (p_T)^2 h(e_T), \tag{4.29}$$

$$U_{Tp_Tp_T} = (w_T - w_0)^2 h(e_T) - M''(p_T) \tag{4.30}$$

and

$$U_{Tp_Tw_T} = p_T(w_T - w_0)h(e_T) + \frac{B'(e_T)}{C'(e_T)}, \tag{4.31}$$

to ensure that the interior solution is a max point. The second-order conditions for the last period are the same as are required in the static model; hence the second-order conditions will not put further restrictions on the solution by the above T-period extension of the static model.

The correlation between monitoring and pay

A quick glimpse at equation (4.28) also reveals that the main results from the static model remain intact. Clearly the monitoring accuracy and the wage in the last period are complementary instruments for motivating workers in this T-period extension if and only if the monitoring accuracy and the wage are complements in the static model; they are characterized by the same equations and thus have the same values. Thus, the following propositions can be stated:

Proposition 2 *According to any shift in the employer's benefit function, $B(e)$, or the employee's cost of effort function, $C(e)$, monitoring and the last period's wage are complementary instruments if and only if $-p^*M''(p^*)/M'(p^*) < 1$.*

Proposition 3 *According to any shift in the employer's monitoring cost function, $M(p)$, monitoring and the last period's wage are complementary instruments if and only if $U^*_{pw} > 0$.*

Equation (4.28) tells us that the optimal wage in other periods than the last period is a linear combination between the optimal wage in the last period and the minimum payment. The weight on the minimum payment decreases as the monitoring accuracy increases and vice versa, if the above conditions for complementarity hold. Hence, if monitoring accuracy and the wage in the last period are complementary instruments for motivating workers, then monitoring accuracy and the wage in other periods are as well. More formally, by differentiating the expression for the optimal wage in other periods than the last period, (4.28), with respect to the optimal monitoring accuracy,

$$\frac{dw_t^*}{dp^*} = (w_T^* - w_0) + (1 - \delta(1 - p^*))\frac{dw_T^*}{dp^*},$$

it is easily seen that if monitoring and the last period's wage are positively correlated, monitoring and the wage in other periods must also be positively correlated. However, the opposite is not true! The condition for complementarity between w_t^* and p^* is weaker than the condition for complementarity between w_T^* and p^*.

Proposition 4 *According to any shift in the employer's benefit function, $B(e)$, or the employee's cost of effort function, $C(e)$, monitoring and pay are complementary instruments if $-p^*M''(p^*)/M'(p^*) < 1$.*

Proposition 5 *According to any shift in the employer's monitoring cost function, $M(p)$, monitoring and the last period's wage are complementary instruments if $U^*_{pw} > 0$.*

The equation (4.28) also gives an insight into how the employer optimally should respond to changes in the discount factor. Clearly, the optimal monitoring accuracy, the optimal effort level and the optimal wage in the last period are all independent of the discount factor since they are the same as in the static case. On the other hand, the weight on the minimum payment in the linear combination determining the optimal wage increases as the discount factor increases and vice versa. Hence, we have that for all periods but the last one, the optimal wage is negatively correlated with the discount factor.

Proposition 6 *The optimal wage in all periods but the last period is negatively correlated with the discount factor, while the optimal wage in the last period, the optimal monitoring accuracy and the optimal effort level are independent of the discount factor.*

Welfare implications

Since both the exerted effort level e_t and the monitoring accuracy p_t are stationary, the length of the relationship does not affect the degree of social efficiency of the contract as long as the principal earns enough to start the project. The contract length will merely have distributional implications. It is easy to rewrite the expression (4.27) to get the following expression for the present value of getting hired by the firm:

$$V_1^*(T) = p^*(1-p^*)M'(p^*) + \sum_{i=0}^{T-1} \delta^i w_0. \qquad (4.32)$$

Thus, the employee's per period compensation is decreasing with the length of the relationship. Conversely, the employer benefits from a longer relationship and will only offer the longest contract which is allowed for by the nature of the project or other exogenous circumstances.

In the absence of the legal rules making it impossible for the employee to fire a non-shirking employee, the employer is forced to repeat the static outcome T times. In this case, the present value for the employee becomes

$$\sum_{i=0}^{T-1} \delta^i V_1(1) = \sum_{i=0}^{T-1} \delta^i (p^*(1-p^*)M'(p^*) + w_0). \qquad (4.33)$$

This is larger than the present value from the long-term relationship (4.32). Hence the following propositions can be stated:

Proposition 7 *The degree of social efficiency of the contract is not affected by the length of the relationship.*

Proposition 8 *The length of the relationship is a matter for dispute: the employer prefers longer and the employee prefers shorter relationships.*

Proposition 9 *Legal rules protecting employees from being fired lower the employees' present value of compensation.*

The welfare implications from changes in the employee's liability are almost trivial. It is easy to check that the equations (4.21) and (4.22) imply that the equilibrium monitoring accuracy, p^*, and the equilibrium level of effort, e^*, are both independent of the limited liability constraint, w_0. Thus the degree of social efficiency is independent of the employee's liability. This should be said with one important reservation though. The equations (4.19) and (4.28) reveal that

$$\frac{dw_t^*}{dw_0} = \frac{dw_T^*}{dw_0} = 1. \tag{4.34}$$

In other words, an increase in the limited liability parameter redistributes wealth from the employer to the employee. If the limited liability parameter becomes high enough, it is no longer profitable for the employer to hire the employee, and the degree of social efficiency will of course be affected. Hence, it can be stated that:

Proposition 10 *So long as the employer can afford to hire the employee, the degree of social efficiency is not affected by changes in the employee's liability.*

Implicit bonding contracts

The above version of the efficiency wage model still implies rents to employees, even if they diminish with the length of the relationship. This common feature of the shirking model of efficiency wages has been criticized in many articles. Carmichael (1985), for example, argues that these rents should be eliminated by entrance fees or bonds. The idea of a bond is found in for example Lazear (1981) and Becker and Stigler (1974). There the employee posts a bond initially, is paid interest on it, and gets it back when the relationship is over. The bond transactions are included in the employee's wage path in their papers. One result is a negative compensation in the first period, and a constant compensation in intermediate periods which is lower than the higher compensation in the last period.

To understand this result in the framework of the model in this paper, an important clarification is called for. A factor closely related to the limited liability constraint is the employee's initial wealth. If contractible, i.e. if it can be taken away from the employee in the case where he does not meet or exceed his effort target, the wealth will add liability and hence lower the liability constraint. Wealthier employees would then be more attractive to hire since, as we learned from the previous section, all periods' equilibrium wages will be lower. On the other hand, if the employee cannot commit to keeping his wealth contractible, as soon as he is hired, he will try to make his wealth non-contractible or simply get rid of it by for example consuming it. The commitment is something that can be accomplished by a bond. This paper will make a clear distinction between the wage and the bond transaction by simply assuming that the wealth that is used for a bond never leaves the employee. A bond is just a contract which says that some of the employee's initial wealth, that can be taken away from him if he does not meet or exceed his effort target, must remain intact during the employment relationship. This is entirely captured by a decrease in the limited liability parameter equivalent to the size of the bond.

The idea of employees purchasing job opportunities, which is sometimes mentioned in this context, is something completely different if the relationship between the employer and the employee lasts longer than one period. If the employer sells a job and the employee uses hi contractible wealth to pay for it, the employer will regret selling the job as soon as he realizes that the equilibrium wage path has shifted upwards and, according to equation (4.34), he will indirectly repay the employee the whole price during each period of their relationship. It is easy to check that if the employee instead uses non-contractible wealth to pay for the job opportunity, it is always better for the employer to let him keep the money if he can commit, for example via a bond, not to spend it (i.e. if the relationship lasts longer than one period). Thus, for an employer to charge an entrance fee is always a bad idea.

Firm-specific human capital investments

Both bonds and entrance fees have the unappealing property of introducing moral hazard into the model. That is, the employee will have an incentive to claim that the employee shirks even if he does not, so that he can appropriate the bond or the entrance fee. Maybe that is the reason why we do not see bonds more frequently. However, a bond can appear in other shapes. Under reasonable assumptions one can interpret an employee's investment in firm-specific human capital as an implicit bond. The only feature of wealth that is of interest when using it for a bond is that the employee will be hurt by losing it, and this is true for firm-specific human capital as well.

Below, it will be demonstrated how a firm-specific human capital investment can be used as a substitute for an entrance fee and at the same time solve the moral hazard problem in a self-enforcing static efficiency wage contract. The moral hazard problem disappears because if the employer fires the employee, he will lose the firm-specific human capital investment at the same time. The analogous reasoning applies for a multi-period model, but then it is for a bond instead of an entrance fee.

The timing is as follows. First, many risk-neutral employees who compete for a job opportunity can invest in firm-specific human capital at a cost $h \in \left[0, \bar{h}\right]$. The investment is sunk and gives an employee a private benefit of exactly h, if he is engaged by the firm; thus there are no *direct* gains from the investment. Then, a risk-neutral employer observes the education levels and employs a single employee. After the employee is hired there are only two possible outcomes of his private benefit from education: his chosen level of education if he remains hired and zero if he gets fired. It is easily seen that it is best for the employer to give the employee $w + h$ if he meets or exceeds the target and the minimum payment w_0 otherwise. Hence, now the incentive compatibility constraint becomes

$$p(w - w_0 + h) \geq C(\hat{e}) \tag{4.35}$$

and the expression for the actual effort which the employee will exert is

$$e(p, w) := C^{-1}((w - w_0 + h)p). \tag{4.36}$$

The two equations characterizing the solution, (4.19) and (4.22), now become

$$w^* = w_0 - h + pM'(p^*) \tag{4.37}$$

and

$$p^* \frac{B'(C^{-1}((p^*)^2 M(p^*)))}{C'(C^{-1}((p^*)^2 M(p^*)))} - 1 = 0. \tag{4.38}$$

To investigate the implications of the employee's firm-specific human capital investment, the employer's and the employee's utilities in equilibrium are differentiated with respect to the education level, h, chosen by the employee. The employer's utility in equilibrium is

$$U(p^*, w^*) = B(C^{-1}(p^*(w^* - w_0 + h))) - w^* - M(p^*) \quad (4.39)$$

and the employee's utility in equilibrium is

$$\begin{aligned} V(p^*, w^*) &= w^* - C(C^{-1}(p^*(w^* - w_0 + h))) \\ &= w^* - p^*(w^* - w_0 + h). \end{aligned} \quad (4.40)$$

Equation (4.38) reveals that the optimal monitoring accuracy, p^*, is independent of h. Knowing this, it is easy to differentiate equation (4.37) to get

$$\frac{dw^*}{dh} = -1. \quad (4.41)$$

This will be of help when differentiating (4.39) and (4.40) with respect to the education level chosen by the employee:

$$\frac{dU(p^*, w^*)}{dh} = \frac{B'(e^*)}{C'(e^*)} p^* \left(\frac{dw^*}{dh} + 1 \right) - \frac{dw^*}{dh} = 1 \quad (4.42)$$

and

$$\frac{dV(p^*, w^*)}{dh} = \frac{dw^*}{dh} - p^* \left(\frac{dw^*}{dh} + 1 \right) = -1. \quad (4.43)$$

It can now be seen that the firm-specific human capital investment indirectly redistributes wealth from the employee to the employer. The amount that is taken away from the employee and given to the employer is exactly the same as the former's private benefit from his education level.

Now, go back in time to when many employees compete for the job opportunity. The goal is to find a Nash equilibrium in education level choices. Firstly, any situation in which more than one employee chooses a positive level of education cannot be a Nash equilibrium. This is because all the employees who chose a positive level of education and were not hired would then have preferred to choose another education level (zero, or higher than the hired employee). Secondly, any situation in which the hired employee receives rents, i.e. any situation in which

$$V(p^*, w^*) = w^* - p^*(w^* - w_0 + h) > w_0 \quad (4.44)$$

or more simply, in which

$$h < (1 - p^*)p^* M'(p^*), \quad (4.45)$$

cannot be a Nash equilibrium, since then another employee would have wanted to undercut him by choosing a higher education level. And finally, any situation in which the hired employee receives negative rents, i.e. any situation in which

$$h > (1 - p^*)p^*M'(p^*), \tag{4.46}$$

cannot be a Nash equilibrium, since then the hired employee would have been better off by choosing a zero education level.

Thus, the only remaining candidate for a Nash equilibrium is that one employee chooses the education level

$$h = (1 - p^*)p^*M'(p^*), \tag{4.47}$$

and all the others choose the education level zero. And this is indeed a Nash equilibrium. Hence, the following proposition can be stated:

Proposition 11 *If employees can freely choose to invest in firm-specific human capital or if their potential education level, $\bar{h}$, is higher than $(1 - p^*)p^*M'(p^*)$, the hired employee will receive no rents, i.e. his individual rationality constraint will be binding.*

And if the upper bound of the education choice is binding for the employee, it can be stated that:

Proposition 12 *If the employee's potential education level, $\bar{h}$, is less than $(1 - p^*)p^*M'(p^*)$, the hired employee will receive rents, i.e. his individual rationality constraint will not be binding.*

Extensions and further research

Firm commitment, or explicit bonding contracts

A natural extension of the model is to allow the employer to commit himself to future control variables. So far, no restriction on the equilibrium wage path has been imposed on the model. In a setting where the employer cannot commit to future control variables, such a restriction is obviously not needed since equilibrium wages are always higher than the limited liability constraint. In the first example below, a minimum wage restriction that coincides with the employee's limited liability is imposed. Then it is discussed what happens in the absence of that restriction. Assume for the sake of simplicity that $T = 2$. Using the expression for the employee's ex post utility in the last period (4.26) to substitute for V^*_{t+1}, the incentive compatibility constraints (4.9) and (4.8) now become

$$e_1 = C^{-1}\left(p_1(w_1 - w_0 + \delta(1-p_2)(w_2 - w_0)\right) \qquad (4.48)$$

and

$$e_2 = C^{-1}\left(p_2(w_2 - w_0)\right). \qquad (4.49)$$

The employer will use these expressions to substitute for e_1 and e_2 in his utility function. Hence he will face the following maximization problem:

$$\underset{p_1,p_2,w_1,w_2}{Max} \quad B(e_1) - w_1 - M(p_1) + \delta(B(e_2) - w_2 - M(p_2))$$

subject to the constraints $w_1, w_2 \geq w_0$ and $p_1, p_2 \in [0,1]$. Form the Lagrangian

$$\begin{aligned} L &= B(e_1) - w_1 - M(p_1) + \delta(B(e_2) - w_2 - M(p_2)) \\ &\quad +\lambda_1(w_1 - w_0) + \lambda_2(w_2 - w_0) \end{aligned}$$

and denote the solution to this problem $(p_1^*, p_2^*, w_1^*, w_2^*)$. Furthermore, let the corresponding amount of executed effort, $e_1(p_1^*, p_2^*, w_1^*, w_2^*)$ and $e_2(p_2^*, w_2^*)$ be denoted e_1^* and e_2^* respectively. The Kuhn-Tucker conditions are then:

$$\frac{B'(e_1^*)}{C'(e_1^*)}(w_1^* - w_0 + \delta(1-p_2^*)(w_2^* - w_0)) - M'(p_1^*) = 0, \quad (4.50)$$

$$\frac{B'(e_1^*)}{C'(e_1^*)}p_1^* - 1 + \lambda_1 = 0, \qquad (4.51)$$

$$\lambda_1(w_1^* - w_0) = 0, \qquad (4.52)$$

$$\frac{B'(e_2^*)}{C'(e_2^*)}\delta(w_2^* - w_0) - \frac{B'(e_1^*)}{C'(e_1^*)}\delta p_1^*(w_2^* - w_0) - \delta M'(p_2^*) = 0, \quad (4.53)$$

$$\frac{B'(e_1^*)}{C'(e_1^*)}\delta p_1^*(1-p_2^*) + \frac{B'(e_2^*)}{C'(e_2^*)}\delta p_2^* - \delta + \lambda_2 = 0, \qquad (4.54)$$

and

$$\lambda_2(w_2^* - w_0) = 0. \qquad (4.55)$$

It is easily seen that λ_2 must be zero, because otherwise the condition (4.55) tells us that w_2^* must be equal to w_0, which substituted into condition (4.53) together with the incentive constraint (4.49) implies zero effort in the second period. Then the employer would rather skip the last period. λ_1 on the other hand cannot be zero, because if it is, the conditions (4.51) and (4.54) require that $B'(e_2^*)/C'(e_2^*) = 1$, which again substituted into condition (4.53) together with the incentive constraint (4.49) implies zero effort in the second period. This is an example of the general result that is stated in the following proposition:

Proposition 13 *If the employer can commit himself to future control variables, the optimal wage path will be constituted by the minimum wage in all periods but the last period.*

The wage path is illustrated in Figure 8.

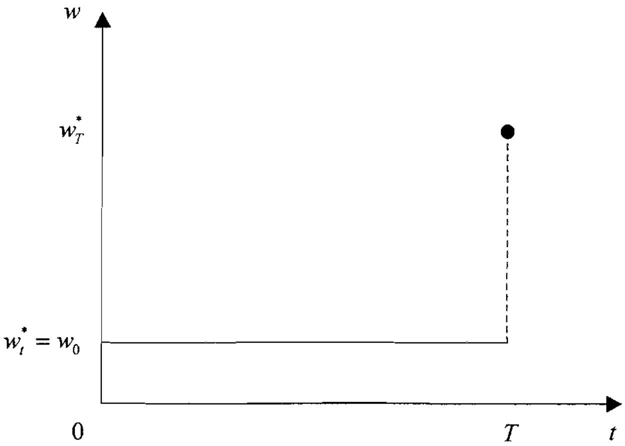


Figure 8

This result is not surprising. If the employee believes the employer's promises of future rewards, the existence of a wage premium in earlier periods than the last period is a conspicuous inefficiency. Without changing the incentives in one period and adding extra incentives in subsequent periods, a wage premium in that period can be replaced by the present value of this wage premium in the last period at no cost to the employer.

The minimum wage restriction is binding in the first period, and without it the employer would choose a lower wage in the first period together with a promise of a higher second period wage. In fact, he would minimize monitoring costs and approach the first best solution in the second period by an infinitely large fee in the first period together with an infinitely large second-period wage. In practice, even if there is no minimum wage restriction, the equilibrium wage may at least be restricted to be non-negative. But the intuition is straightforward: if the employer is able to make commitments he should save as much as possible of the payment for the last period.

Savings

If the limited liability constraint is affected by the employee's earlier wages, the resulting optimal wage path is affected as well. Say for example that the employer's wage is used entirely to add up liability. Then every dollar that is given to the employee in the first period makes the employer able to subtract one dollar from the employee's paycheck in all the remaining periods. But a rational employee who foresees this reaction of the employer would of course spend all his money on consumption, leaving nothing to save. In a more general setting one could include the mechanism that employees compete for a job opportunity with commitments to save, thus gradually lowering their limited liability constraints. This may be done to the extent that all the hired employee's surplus is given up to the employer. It is noteworthy that a commitment to save can be made even if the employee cannot afford to buy a bond. The analogical reasoning applies for firm specific-human capital. This may be the reason why we so often see employers letting their employees educate themselves during working hours.

Concluding remarks

The unique optimal wage path derived in this paper shows similarities with one of the possible optimal wage paths described in Lazear (1981). There, the employee receives a wage lower than the value of his marginal product over his lifetime, but is compensated by a large lump sum in the last period to set present values of payment and marginal products equal. There is an important difference though; here the employee still receives some rents inherited from the efficiency wage mechanism, even if they are strongly diminished by the length of the relationship.

Another possibility that Lazear presents is the same as that first found in Becker and Stigler (1974). The employee posts a bond initially, is paid interest on it, and gets it back when the relationship is over. The result is a negative compensation in the first period, a constant compensation in intermediate periods lower than the higher compensation in the last period and in total, no rents to the employee. There are two obvious differences compared to the result in this paper: the first period compensation and the employee's rents. The reason for the first is only technical. In this paper, the idea of a bond is captured by the limited liability constraint. If the employee has initial wealth that could be used for a bond, it is simply assumed to lower the limited liability constraint, w_0. The employee doe not have to post the bond as long as the wealth that was meant for the bond is contractible. The reason for the second difference is more important. One of the goals with this paper has been to explain an upward sloping wage profile with a self enforcing contract, without introducing a reputation mechanism, a third part or the possibility to make commitments. If this is left out, the moral hazard problem of employer default is not a problem, as shown in for example Bhattacharya (1987), Carmichael

(1983a, 1983b) and Malcomson (1984, 1986). And if the moral hazard problem is not present, it is not hard to eliminate the employee's rents.

A third path proposed by Lazear is that the employee is paid less than the value of his marginal product initially, and receives a pension for some periods after the actual relationship is over. If one relaxes the assumption that the employer cannot commit to future control variables, and assumes that he can commit to pay a pension, then there is an infinite number of optimal contracts. The employer could for example distribute the last period's wage premium plus compensation for the resulting depreciation evenly over a number of following periods. The employer's gain from postponing some of the payment is exactly the same as the employee's loss, since they have identical discounting factors. But since the employer has to compensate for the employee's loss, both the employer and the employee will be indifferent between no pension and all kinds of such pension arrangements.

Lazear also finds that wage profiles which are flatter and more smoothly increasing than step functions suggested above can be optimal. He argues that in addition to the employee's incentive gains from a steep wage profile, there is a loss represented by a higher temptation for the employer to breach the contract. This is a trade-off that is excluded in the model analyzed in this paper. It simply assumes that the employer cannot fire the employee if he is not caught shirking. However, the employer is not tempted to breach the contract because he makes a positive profit in the last period. It is optimal to pay the higher wage in the last period, resulting from efficiency wages contrary to the delayed payment/bonding contract wage, not because it compensates the employee for previous periods, but because it yields the correct incentives in the last period!

Influential papers presenting efficiency wage models, such as Shapiro and Stiglitz (1984) and Bulow and Summers (1986), have ruled out the possibility of delayed payment/bonding contracts. This is mainly because they have not been satisfied with the treatment in earlier versions of the employer's moral hazard problem and the employee's wealth constraint. Maybe the most important contribution of this paper is that it builds a bridge between efficiency wages and delayed payment/bonding contracts; it demonstrates how a postponed payment can emerge in a self-enforcing contract, even if the employee has no initial wealth. Thus, in opposition to previous intuitions, there should be no antagonism between efficiency wages and delayed payment/bonding contracts. On the contrary, they are two compatible mechanisms which the profit maximizing firm should use in concord.[1]

[1] For other hybrid models of efficiency wages and delayed payment/bonding contracts, see Eaton and White (1982) and Dickens, Katz, Lang and Summers (1989).

References

Allgulin, Magnus (1999): Monitoring and Pay: General results, *SSE/EFI Working Paper Series in Economics and Finance* No. 340, Stockholm School of Economics.

Allgulin, Magnus and Tore Ellingsen (1998): Monitoring and Pay, *Working Paper Series in Economics and Finance* No. 245, Stockholm School of Economics.

Becker, Gary and George Stigler (1974): Law Enforcement, Malfeasance, and the Compensation of Enforcers, *Journal of Legal Studies* III, 1-18.

Bhattacharya, Sudipto (1987): Tournaments, Termination Schemes, and Forcing Contracts, Mimeo, School of Business, University of California, Berkeley.

Bull, Clive (1987): The Existence of Self-Enforcing Implicit Contracts, *Quarterly Journal of Economics* February, 147–159.

Bull, Clive (1985): Equilibrium Unemployment as a Worker Discipline Device: Comment, *American Economic Review* September, 890–891.

Bulow, Jeremy I. and Lawrence H. Summers (1986): A Theory of Dual Labor Markets with Applications to Industrial Policy, Discrimination, and Keynesian Unemployment, *Journal of Labor Economics* 4, 376–414.

Carmichael, H. Lorne (1990): Efficiency Wage Models of Unemployment - One View, *Economic Inquiry* April, 28, 269-295.

Carmichael, H. Lorne (1985): Can Unemployment be Involuntary?: Comment, *American Economic Review* December, 75, 1213-1214.

Carmichael, H. Lorne (1983b): The Agent-Agents Problem: Payment by Relative Output, *Journal of Labor Economics* January, 37-50.

Carmichael, H. Lorne (1983a): Firm Specific Human Capital and Promotion Ladders, *Bell Journal* Spring, 241-258.

Demougin, Dominique and Claude Fluet (1997): Monitoring versus Incentives: Substitutes or Complements?, Mimeo, Université du Québec à Montréal.

Dickens, W. T., L. F. Katz, K. Lang and L. H. Summers (1989): Employee Crime and the Monitoring Puzzle, *Journal of Labor Economics* vol. 7, no. 3, 331–347.

Eaton, C. and W. D. White (1982): Agent Compensation and the Limits of Bonding, *Economic Inquiry* XX, 330–343.

Lazear, Edward (1995): *Personnel Economics,* (Cambridge Mass.: MIT Press).

Lazear, Edward (1981): Agency, Earnings Profiles, Productivity and Hours Restrictions, *American Economic Review* LXXI, 606-620.

Malcomson, James (1986): Rank Order Contracts for a Principal with Many Agents, *Review of Economic Studies* October, 807-818.

Malcomson, James (1984): Work Incentives, Hierarchy, and Internal Labor Markets, *Journal of Political Economy* June, 486-507.

Shapiro, Carl and Joseph E. Stiglitz (1984): Involuntary Unemployment as a Worker Discipline Device, *American Economic Review* 74, 433–444.

5 DO MARKET-BASED INCENTIVES LOWER THE COST OF COMPLIANCE?

Abstract This paper analyzes policies for regulating polluting firms under imperfect monitoring. The main finding is that a certain emission target is always more costly to enforce if there exists a market for emission permits than if there is none. The intuition is that a market restricts the regulatory agency to imposing incentive schemes which are linear in firms' emission levels, and these are less powerful than the best non-linear incentive schemes. Another result is that monitoring and monetary incentives are complementary policy instruments.

JEL CLASSIFICATION: Q28, Q38.
KEYWORDS: Monitoring, quota, market, emission permits.

Introduction

Ignoring monitoring and enforcement costs while designing a regulatory standard will most certainly lead to the implementation of an inferior policy. Either the regulatory agency will impose a too expensive policy, or it will not succeed in making firms comply with the standard. In fact, the consequences may be disastrous.[1] This is a lesson both academics and policy makers have learned over the last few years. Consequently, monitoring and enforcement have been the focus in many recent theoretical and empirical studies, and regulatory agencies have read the proposed prescriptions carefully. The most obvious question that needs to be answered is of course: What is the socially most desirable regulation? Or more precisely, which is preferable, a command and control policy or more market-based incentives?

The purpose of this paper is to answer that question, primarily with reference to any negative external effect on the environment, but the reasoning throughout the paper can easily be extended to embrace most kinds of negative external effects. The regulatory agency is modelled as the principal and a firm that produces a good with a negative external effect, e.g. pollution, is modelled as the agent in a standard principal agent setting.

The paper starts out by deriving the optimal command and control policy. This will consist of an emission quota, a penalty if the emission level exceeds the quota, and a compliance fee otherwise. The firm has to be given correct incentives in order to obey the quota. The mechanism for this is first described in Becker's (1968) economic analysis of crime.[2] The basic insight of that article is that potential criminals respond to both the probability of detection and the severity of punishment if detected and convicted. Thus, deterrence from criminal activity may be enhanced either by raising the penalty, or by increasing monitoring activities to raise the likelihood that the offender will be caught. That is true even here. On the other hand, here it will be assumed that there is an exogenous limit to how high a penalty is feasible, which will be binding. The reason may be political or due to a wealth constraint of the regulated firm. Instead the net penalty, i.e. the difference between the penalty and a compliance fee, can be varied.

The first noteworthy result is that the optimal compliance

[1] See the section *Empirical evidence* and Sigman (1998).

[2] Downing and Watson (1974) and Harford (1978) were the first variants of the Becker (1968) model that appeared in the environmental economics literature.

fee may be negative, i.e. some polluters should not carry any tax burden at all but should rather be subsidized.[3] This follows from the fact that a smaller fee is optimally used as a complement to increased monitoring to create incentives; consequently certain firms, which need strong incentives, are monitored very accurately and are awarded a subsidy in equilibrium.

The idea of a compliance subsidy is found earlier in the environmental economics literature: Downing and Kimball (1982) note that a cost subsidy will make a regulated firm more inclined to comply with a regulatory policy, because the firm has more to lose by not complying. The same message is put forward in for example Sullivan (1987), Fullerton and Kinnaman (1995) and Sigman (1998). A somewhat different cause for a subsidy is presented in Nowell and Shogren (1994), who study its effects on liability evasive activities.

The paper continues by finding the optimal emission-fee regulatory policy, and this policy is shown to be at least as good as the best market-based policy. Finally, the command and control and the emission fee policies are compared. This leads directly to a second result. Despite the fact that the major part of the environmental economics literature touts the use of market-based incentives, the finding here is the opposite: a certain emission target is always more costly to enforce if there exists a market for emission permits than if there does not.[4] The intuition behind the result is simply that a market restricts the regulatory agency to use an incentive scheme for the polluting firms that is linear in their emission levels, and that such a restriction makes the incentive scheme less powerful.

The paper is organized as follows. The next section sets up the model. In the third section, the best command and control policy is derived. The fourth section derives the best emission-fee policy, which is also shown to be weakly better than any market-based incentive scheme. A comparison between the different policies is performed in the section after that. The following section discusses some empirical evidence and finally, the last section concludes.

[3] The analogy with Becker's model is straightforward. If it is impossible to raise the penalty, deterrence from criminal activity may for example be enhanced by giving a subsidy to the poorest people in the community, since they will be the most tempted to engage in criminal activities.

[4] Sandmo (1998) compares a uniform tax rate to the use of quotas. He focuses on the case when there is imperfect compliance in equilibrium. Hence, in the case of quotas, with some probability, a violation is discovered and a penalty will be paid. This turns out to have some of the characteristics of a uniform tax, even though it is imposed only with a certain probability. Throughout this chapter, there is perfect compliance in equilibrium, but a quota is accompanied by a *compliance* tax that is not present in Sandmo (1998). The different implications of the two models is essentially accounted for by this difference.

A risk neutral principal imposes a policy on a single risk neutral agent. There is perfect knowledge on the part of both the principal and the agent. The agent can produce an amount of a good $x \in \mathbf{R}_+$ yielding the profit $\beta\Pi(x)$ to both him and the principal, but at the same time the negative external effect $\gamma E(x)$ is only to the principal. The parameters β and γ are both positive. The policy specifies the agent's right to produce a certain amount of the good x and a transfer t that the agent has to pay for the right. If the principal cannot find a policy yielding a positive utility, he can costlessly close down the agent's activity.

The principal has the ability to observe and verify the amount of the good x that is produced with probability p. The principal can indirectly choose this detection probability through his choice of resources devoted to monitoring, $\mu M(p)$, where $\mu > 0$. The profit, externality and monitoring cost functions are assumed to have the plausible properties: $\Pi''(x) < 0$, $E(0) = 0$, $E'(x) > 0$, $E''(x) \geq 0$[5] and $M'(p) > 0$.

Define the principal's ex post utility as

$$U = \beta\Pi(x) + \lambda t - \gamma E(x) - (1 + \lambda)\mu M(p), \qquad (5.1)$$

where $\lambda > 0$ denotes the shadow price of raising public funds. Note that $\gamma E(x)$ is not multiplied by $(1 + \lambda)$. It is assumed that the burden of the negative external effects is carried by the economy rather than the authorities. However, this assumption will not affect the results in any major respects.

The agent's ex post net profit is

$$V = \beta\Pi(x) - t. \qquad (5.2)$$

There is an upper limit $T \in \mathbf{R}_+$ to what the principal can extract from the agent if the latter does not comply with the policy. For example, the legislature might impose a fixed penalty or firm insolvency constraints might make higher penalties inappropriate. The agent's only interest is to maximize his

[5] The assumption of a weakly convex externality function is plausible for a large set of environmental externalities. However, the model allows for some concavity as long as the second order conditions for the principal's maximization problem hold. Thus, the assumption may be relaxed to

$$E''(x) > \frac{E'(x)\Pi''(x)}{\Pi'(x)}.$$

expected net profit; therefore the following incentive compatibility (IC) constraint must be satisfied for all x:

$$\beta\Pi(\widehat{x}) - t \geq \beta\Pi(x) - (1-p)t - pT. \qquad (5.3)$$

Command and control policy

We see that any incentive compatible policy that implements $\widehat{x}$ can, without loss to the principal, be replicated by a step function of the form $t(x) = T$ for $x > \widehat{x}$ and $t(x) = t$ for $x \leq \widehat{x}$.[6] In other words, the principal sets a quota $\widehat{x}$. The agent pays t if he meets or produces less than the quota and he pays T if he exceeds the quota, as illustrated in Figure 9.

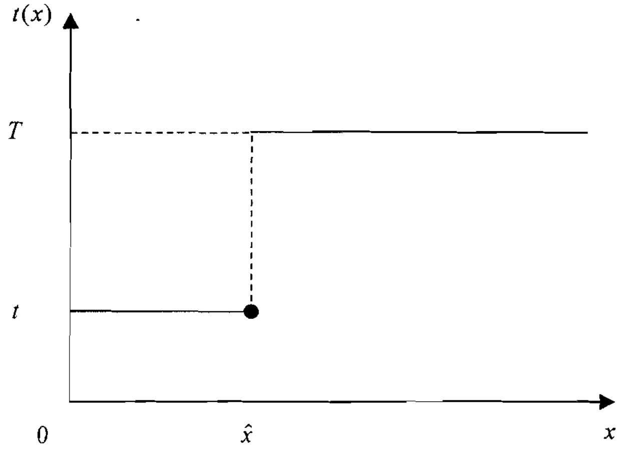


Figure 9

In the absence of a policy, and if the agent were free to produce any desired amount of the good x, he would choose to produce $x = X$, where $X = \arg\max(\beta\Pi(x))$. Since any correctly designed quota, $\hat{x}$, will be less than or equal to X, it follows directly from the initial assumptions that $\Pi'(\hat{x}) \geq 0$. If the agent does not want to comply with this policy, he will choose to produce $x = X$. Therefore the IC constraint simplifies to

$$p(T-t) \geq \beta(\Pi(X) - \Pi(\widehat{x})). \qquad (5.4)$$

[6] The optimality of a step function incentive scheme is quite general. Demougin and Fluet (1997) show it for an effort target in an efficiency wage model, but this kind of incentive scheme also belongs to the set of optimal regulatory policies. What is interesting is that, generally, regulatory agencies have a strong predilection for a special case of this policy: a quota and a penalty, but no compliance fee.

The monitoring accuracy, p, is costly and the transfer, t, is beneficial for the principal; thus, in order to implement a certain $\hat{x}$ he will choose them such that inequality (5.4) is binding. Rearranging and inverting this equality yields an expression for the amount of the good the agent will produce:

$$x(p,t) := \Pi^{-1}(\Pi(X) - p(T-t)/\beta). \tag{5.5}$$

It is now possible to use the expression (5.5) for $x(p,t)$ to substitute for x in the principal's utility function (5.1). It follows that the principal has a problem where he has to choose a probability p and a transfer t that solve

$$\underset{p,t}{Max}\ \ U(p,t) = \beta\Pi(x(p,t)) + \lambda t - \gamma E(x(p,t)) - (1+\lambda)\mu M(p) \tag{5.6}$$

subject to the constraints $t \leq T$ and $p \in [0,1]$. Denote the solution to this problem (p^*, t^*), and let $x^* := x(p^*, t^*)$ be the corresponding amount of the good produced. The first-order conditions characterizing the solution are then

$$\lambda + p^* - r(x^*)p^* \leq 0 \tag{5.7}$$

and

$$-(T-t^*) + r(x^*)(T-t^*) - (1+\lambda)\mu M'(p^*) \leq 0, \tag{5.8}$$

with equality whenever the solution is interior and $r(x) = \gamma E'(x)/\beta\Pi'(x)$ denotes the ratio of the marginal externality to the marginal profit.

Equation (5.7) tells us that the marginal externality from a production increase will always be larger than the marginal profit; thus the principal will allow for some inefficient overproduction in any solution. In addition, according to equation (5.8) there must be a positive level of monitoring in order to induce a production level lower than X. The first-order conditions (5.7) and (5.8) together tell us that in any interior solution, we will have the relation

$$\lambda(T-t^*) = (1+\lambda)p^*\mu M'(p^*). \tag{5.9}$$

Using (5.9) to substitute for t^* in the principal's utility function gives us the following expression for the principal's utility of x^* and p^*:

$$U(x^*,p^*) = \beta\Pi(e^*) - \lambda T - \gamma E(x^*) - (1+\lambda)(p^*\mu M'(p^*) - \mu M(p^*)). \tag{5.10}$$

This expression will be used later when it is assumed that the principal is interested in implementing a certain production level $\hat{x}$, and the above utility is compared to the highest utility that the principal can get by using an emission fee policy.

Let U_{ij} denote the second derivative of U and let U^*_{ij} denote the second derivative of U evaluated at a solution (p^*, t^*). The second-order conditions are then $U^*_{tt} < 0, U^*_{pp} < 0$ and $U^*_{pp}U^*_{tt} - (U^*_{pt})^2 > 0$, where

$$U_{tt} = \frac{\gamma p^2 h(x)}{\beta^2}, \tag{5.11}$$

$$U_{pp} = \frac{\gamma (T-t)^2 h(x)}{\beta^2} - \mu M''(p)(1+\lambda), \tag{5.12}$$

$$U_{pt} = \frac{\gamma p (T-t) h(x)}{\beta^2} + 1 - r(x), \tag{5.13}$$

and

$$h(x) := \frac{\Pi''(x)E'(x) - E''(x)\Pi'(x)}{(\Pi'(x))^3}. \tag{5.14}$$

Note that, since $h(x)$ is always negative and $r(x^*)$ is positive and larger than 1, this will imply that $U^*_{pt} < 0$.

Comparative statics

The objective here is to characterize how the level of the transfer t, the accuracy of monitoring p and the production level x vary with the parameters β, γ and μ. To examine the effect of a change in the external effect from the good, γ, we differentiate the two first-order conditions (5.7) and (5.8) and get the two equations:

$$U^*_{tt}dt^* + U^*_{pt}dp^* = \frac{p^*}{\gamma} r(x^*) d\gamma \tag{5.15}$$

$$U^*_{pt}dt^* + U^*_{pp}dp^* = -\frac{T-t^*}{\gamma} d\gamma. \tag{5.16}$$

By Cramer's rule, we get

$$\frac{dt^*}{d\gamma} = \frac{\left[p^*U^*_{pp} + (T-t^*)U^*_{pt}\right] r(x^*)}{\left[U^*_{tt}U^*_{pp} - (U^*_{pt})^2\right] \gamma} \tag{5.17}$$

and

$$\frac{dp^*}{d\gamma} = \frac{\left[-(T-t^*)U^*_{tt} - p^*U^*_{pt}\right] r(x^*)}{\left[U^*_{tt}U^*_{pp} - (U^*_{pt})^2\right] \gamma}. \tag{5.18}$$

The signs of both $dt^*/d\gamma$ and $dp^*/d\gamma$ are determined by the second-order conditions and the fact that $U^*_{pt} < 0$. Clearly it is always true that

$$\frac{dt^*}{d\gamma} < 0 \tag{5.19}$$

and

$$\frac{dp^*}{d\gamma} > 0. \tag{5.20}$$

To understand this we differentiate (5.5) in equilibrium to get

$$dx^* = \frac{[p^* dt^* - (T - t^*)dp^*]}{\beta \Pi'(x^*)}. \tag{5.21}$$

Together with (5.17) and (5.18) this tells us that

$$\frac{dx^*}{dt^*} > 0 \tag{5.22}$$

and

$$\frac{dx^*}{dp^*} < 0. \tag{5.23}$$

The intuition is now straightforward: the more externality the principal gets from a given production level, x, the lower production level, x, he wants. To implement a lower x the principal has to adjust the monitoring accuracy, p, and the transfer, t, in order to fulfill the IC constraint. It is cheaper for the principal to give up a little of the transfer, t, at the same time as increasing the monitoring accuracy, p, than to enforce the lower production level, x, solely by increasing the monitoring accuracy, p, or by lowering the transfer, t. The principal will optimally use p and t to lower the level of x such that the costs of using them are equal at the margin.

Secondly, we investigate the effects of an increase in the cost of monitoring. In this case we have:

$$\frac{dt^*}{d\mu} = \frac{-(1+\lambda)M'(p^*)U^*_{pt}}{\left[U^*_{tt}U^*_{pp} - (U^*_{pt})^2\right]} \tag{5.24}$$

and

$$\frac{dp^*}{d\mu} = \frac{(1+\lambda)M'(p^*)U^*_{tt}}{\left[U^*_{tt}U^*_{pp} - (U^*_{pt})^2\right]}. \tag{5.25}$$

Again, by the second-order conditions and the fact that $U_{pt} < 0$ we have

$$\frac{dt^*}{d\mu} > 0 \tag{5.26}$$

and

$$\frac{dp^*}{d\mu} < 0. \tag{5.27}$$

The intuition for this is that the more expensive or difficult it gets to monitor, the less monitoring the principal will choose to engage in.[7] Less monitoring will in turn make it more

[7] Allgulin and Ellingsen (1998) argue that there is an important difference between the amount of resources devoted to monitoring and the monitoring accuracy. The effect of an increase in the cost of monitoring on the amount of resources devoted to monitoring is

$$\frac{d\mu M(p^*)}{d\mu} = M(p^*) + \mu M'(p^*)\frac{dp^*}{d\mu},$$

expensive to use a lax transfer as an incentive. Thus the principal will increase the transfer and instead allow for a higher production level in order to fulfill the IC constraint.

The analysis of a change in the profit function is unfortunately not as trivial. Using Cramer's rule one more time gives us:

$$\frac{dt^*}{d\beta} = \frac{-p^* U^*_{pp} - (T - t^*) U^*_{pt}}{\left[U^*_{tt} U^*_{pp} - (U^*_{pt})^2\right] \beta} \left[\frac{\gamma p^* (T - t^*) h(x^*)}{\beta^2} + r(x^*) \right] \tag{5.28}$$

and

$$\frac{dp^*}{d\beta} = \frac{(T - t^*) U^*_{tt} + p^* U^*_{pt}}{\left[U^*_{tt} U^*_{pp} - (U^*_{pt})^2\right] \beta} \left[\frac{\gamma p^* (T - t^*) h(x^*)}{\beta^2} + r(x^*) \right]. \tag{5.29}$$

There is no clear rule for whether the sign of the expression in the last parenthesis in these equations is positive or negative. However, by the second-order conditions and the fact that $U^*_{pt} < 0$, it is at least possible to conclude that the sign of $(dt^*/d\beta)/(dp^*/d\beta)$ is always negative. Thus, monitoring and the transfer will be negatively correlated, independently of the source of variation.

It is noteworthy that it is easier to get complementarity between monitoring and monetary incentives in this model than in Allgulin and Ellingsen's (1998) model of efficiency wages. The reason for this is that here, the principal (the regulatory agency) incorporates the agent's benefit (the polluting firm's profit) in its objective.

Finally, using the first order conditions to solve for t^* gives us the expression

$$t^* = T - p^* \mu M'(p^*)(1 + \lambda)/\lambda. \tag{5.30}$$

Hence, it can be claimed that the transfer, t^*, optimally may be chosen to be negative. That is, the principal could optimally be willing to subsidize the agent. To see that this is possible, just assume that $T = 0$ and remember that in any interior solution there must be a positive level of monitoring. According to the results above, the transfer becomes a subsidy for industries with a high γ and a low μ, i.e. for industries that are associated with low monitoring costs and "costly" negative external effects. Moreover, the lower the upper limit, T, to what the principal can extract from the agent in case he does

and the sign is ambiguous in general. Allgulin and Ellingsen (1998) continue to note that a major drawback of most empirical work to date in the efficiency wage literature, is that it uses the amount of resources devoted to monitoring as a measure of monitoring accuracy. Thus, here is a warning for investigators of the correlation between monitoring and a compliance fee.

not comply with the policy and the lower the shadow price of raising public funds, λ, the more likely it is that the transfer is a subsidy.

To sum up the results so far, the following propositions are stated,

Proposition 1 *The accuracy of monitoring, p^*, is increasing in the externality parameter γ and decreasing in the monitoring cost parameter μ.*

Proposition 2 *The transfer, t^*, and the production level, x^*, are decreasing in the externality parameter γ and increasing in the monitoring cost parameter μ.*

Proposition 3 *A decrease in the transfer, t^*, and an increased monitoring accuracy, p^*, are complementary instruments for implementing a lower production level of the polluting good, x^*.*

Proposition 4 *The transfer, t^*, becomes negative (a subsidy) if and only if $T < p^* \mu M'(p^*)(1+\lambda)/\lambda$.*

Welfare implications

Shavell (1979) and Cohen (1987) showed the startling result that if agents are risk averse, they may actually like monitoring. The reasoning is the following: an agent facing an uncertain negative payoff would prefer to pay the expected dollar value of the payoff as an insurance premium rather than face the uncertain situation. Hence, potential violators would be better off if they were monitored more frequently and received a lower penalty (which they do in their models), if found to be in violation of environmental laws, than if they were seldom detected and paid a high price for the rare finding of non-compliance.

Below, it will be demonstrated that this counter-intuitive result may be true even in this model, but for entirely different reasons. From the incentive compatibility constraint, (5.4),we have

$$\beta\Pi(x^*) = \beta\Pi(X) - p^*(T - t^*).$$

Knowing the expression for the equilibrium transfer, (5.30), we can express the agent's net profit in equilibrium in the following way:

$$V = \beta\Pi(X) - T + p^*(1-p^*)\mu M'(p^*)(1+\lambda)/\lambda.$$

Differentiation of this expression reveals that the agent's profit is at its maximum when

$$(1-2p^*)M'(p^*) + p^*(1-p^*)M''(p^*) = 0,$$

i.e. the agent prefers some intermediate level of monitoring. Thus for low levels of monitoring, the agent's profit is increasing in the monitoring accuracy. The intuition is that more monitoring always makes it beneficial for the principal to lower the compliance fee (and eventually pay a transfer), but also to demand a lower production level of the polluting good. For low levels of monitoring accuracy, the compliance fee effect dominates and the agent's net profit is increasing.

Emission-fee policy

The emission-fee policy is that the regulatory agency lets the polluting firm pay a fee per unit of the produced good. To make the policy a little stronger, the principal is allowed to ask for an entrance fee.[8] That is, the policy is of the form $\alpha_0 + \alpha x$ if the agent's production level is observed. As depicted in Figure 10, the strongest punishment is only carried out for the maximum deviation.

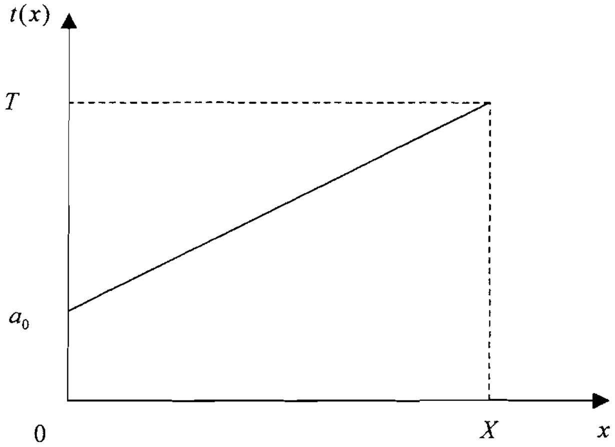


Figure 10

If the agent's production level is not observed, the rational expectations assumption is maintained, that the agent follows the incentive scheme even if the principal does not see it. Thus

[8] Even though this might not reflect real life conditions, this freedom is given to the principal. In the next section it is shown that with this freedom, the emission fee policy is strictly dominated by the command and control policy. With no intercept, the emission fee policy would be even worse. Thus, the reason for the assumption is to eliminate any suspicion that a restriction of no intercept is resposible for the result.

in this case the agent pays the transfer $\alpha_0 + \alpha x^{**}$, where x^{**} denotes the production level in equilibrium. Furthermore the limited liability constraint on the agent restricts the principal to set α_0 to at most $T - \alpha X$. The principal does not give anything away for free so he will optimally set $\alpha_0 = T - \alpha X$.

Hence, the ex ante transfer will now be

$$t(x) = T - \alpha X + \alpha(px + (1-p)x^*). \tag{5.31}$$

Thus the agent will solve the problem:

$$\underset{x}{Max}\ \ V(x) = \beta\Pi(x) - T + \alpha X - \alpha(px + (1-p)x^*) \tag{5.32}$$

with the first-order condition

$$\beta\Pi'(x) = \alpha p. \tag{5.33}$$

The principal takes this maximization behavior as given when maximizing his utility. Inverting the agent's first-order condition and substituting for x, the principal faces the maximization problem

$$\begin{aligned}\underset{\alpha,p}{Max}\, U(\alpha,p) &= \beta\Pi((\Pi')^{-1}(\frac{\alpha p}{\beta})) + \lambda(T - \alpha X + \alpha(\Pi')^{-1}(\frac{\alpha p}{\beta})) \\ &\quad -\gamma E((\Pi')^{-1}(\frac{\alpha p}{\beta})) - (1+\lambda)\mu M(p). \qquad (5.34)\end{aligned}$$

Assuming an interior solution, the first-order conditions are

$$p^{**}\left[\frac{\Pi'(x^{**})}{\Pi''(x^{**})} + \frac{\lambda\alpha^{**}}{\beta\Pi'(x^{**})} - \frac{\gamma E'(x^{**})}{\beta\Pi''(x^{**})}\right] - \lambda(X - x^{**}) = 0 \tag{5.35}$$

and

$$\alpha^{**}\left[\frac{\Pi'(x^{**})}{\Pi''(x^{**})} + \frac{\lambda\alpha^{**}}{\beta\Pi'(x^{**})} - \frac{\gamma E'(x^{**})}{\beta\Pi''(x^{**})}\right] - (1+\lambda)\mu M'(p^{**}) = 0. \tag{5.36}$$

(5.35) and (5.36) together tell us that in any interior solution, we will have the relation

$$\lambda\alpha^{**}(X - x^{**}) = (1+\lambda)p^{**}\mu M'(p^{**}). \tag{5.37}$$

Using (5.37) to substitute for $\alpha^{**}x^{**}$ in the principal's utility function yields an identical expression to (5.10), but now for the principal's utility of x^{**} and p^{**} under a linear incentive scheme:

$$\begin{aligned}U(x^{**},p^{**}) &= \beta\Pi(x^{**}) - \lambda T - \gamma E(x^{**}) \\ &\quad -(1+\lambda)(p^{**}\mu M'(p^{**}) - \mu M(p^{**})). (5.38)\end{aligned}$$

Tradable emission permits

The above policy is very similar to what implicitly emerges if there is a market for emission permits with many agents, in which each agent is a price taker. An agent faces a cost per unit of the produced good, i.e. the market price, even if the regulatory agency tries to impose a policy in the form of a step function. The difference is that he will get the offer from other agents instead of the regulatory agency itself. Figure 11 illustrates an example of an arbitrary non-linear regulation policy and the implicit linear policy that emerges if there exists a market in which agents can trade emission permits freely.

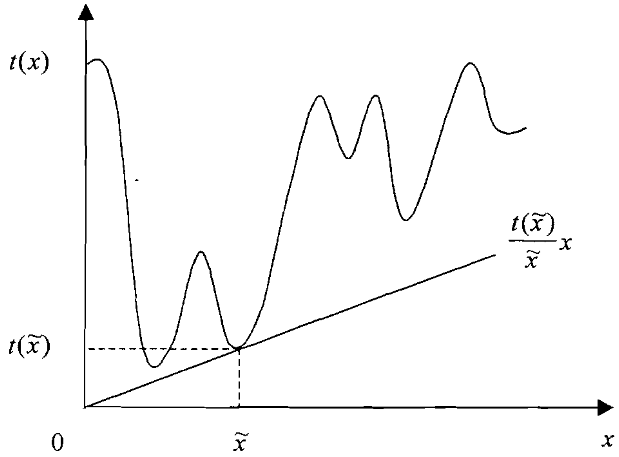


Figure 11

Here, the market buys emission permits where the per unit price, $t(x)/x$, is the lowest, i.e. at the price $t(\tilde{x})/\tilde{x}$. Then it offers each agent the implicit linear policy $t(x) = xt(\tilde{x})/\tilde{x}$. Note that a market also buys at the cheapest price across agents if the regulatory agency, instead of the second-degree price discrimination described above, engages in third-degree price discrimination.

Accordingly, the principal controls the initial allocation of emission permits if he is free to discriminate when selling the emission permits in the first place. But this is only an apparent freedom; the agents are free to trade and they will reallocate the emission rights such that the marginal value of an emission permit is equal for all firms. However, because of the shadow price of raising public funds, the principal will allocate the emission permits from the beginning such that the marginal value of an emission permit is equal for all firms, leaving no gains from trade to the agents.

The principal controls the market price, which of course is the same for all agents, implicitly through the total amount of emission permits sold. This is a more restricted policy, compared to the emission-fee policy where the principal can vary the per unit fee across agents. For example, if there are two different types of agent, the principal will choose a market price lower than the optimal emission fee for one type and higher than the optimal emission fee for the other type. Thus, tradable emission permits will result in an implicit policy that is weakly dominated by the emission-fee policy.

Furthermore, it may be hard for the regulatory agency to ask the agents for an entrance fee for the right to trade in the emission permits market. Thus, in addition to the above disadvantage compared to the emission-fee policy, the intercept, a_0, may be restricted to be zero.[9]

The suboptimality of the emission-fee policy

Assume that the principal is interested in implementing a certain production level, $\hat{x}$. To be able to compare the principal's utility in the two cases (5.10) and (5.38), an investigation of the relationship between p^* and p^{**} is needed. From the incentive compatibility constraint (5.5) and the principal's first order conditions together (5.9) in the non-linear case, it is seen that

$$(p^*)^2 \mu M'(p^*) = \frac{\lambda}{1+\lambda}\beta(\Pi(X) - \Pi(\hat{x})). \tag{5.39}$$

The corresponding expression for the linear case, given by the agent's first order condition (5.33) and the principal's first order conditions together (5.37), is

$$(p^{**})^2 \mu M'(p^{**}) = \frac{\lambda}{1+\lambda}\beta\Pi'(\hat{x})(X - \hat{x}). \tag{5.40}$$

By the assumption $\Pi''(x) < 0$ we know that

$$\Pi(X) - \Pi(\hat{x}) < \Pi'(\hat{x})(X - \hat{x}) \text{ for any } \hat{x} < X. \tag{5.41}$$

This can easily be seen in Figure 12. Here,

$$\Pi(X) - \Pi(\hat{x}) = \int_{\hat{x}}^{X} \Pi'(x)dx$$

is represented by the area B, and

$$\Pi'(\hat{x})(X - \hat{x})$$

[9] This may be a more significant drawback of an emission permits market than the one principally studied in this paper.

is represented by the strictly larger area $A + B$.

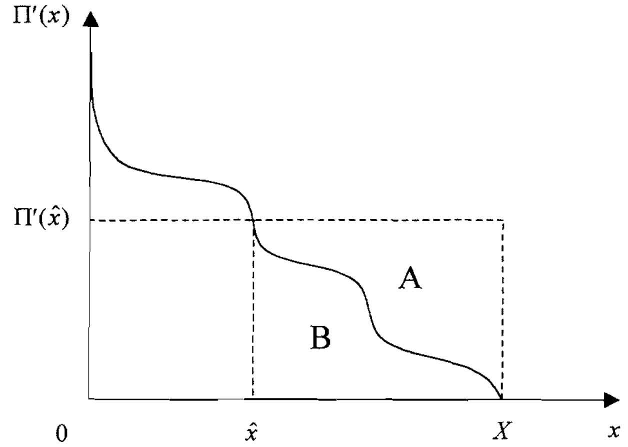


Figure 12

Thus (5.39) and (5.40) together tell us that

$$(p^*)^2 M'(p^*) < (p^{**})^2 M'(p^{**}), \tag{5.42}$$

implying that for any production level the principal wants to implement he will monitor more accurately under a linear incentive scheme if $M''(p) \geq 0$.[10] The least resources required to implement the effort level, $\hat{e}$, using a non-linear policy or a linear policy respectively are now easy to compare. If one substitutes for the equilibrium transfer, it is clear that

$$\beta\Pi(\hat{x}) - \lambda T - \gamma E(\hat{x}) - (1+\lambda)(p^*\mu M'(p^*) - \mu M(p^*)) >$$

$$\beta\Pi(\hat{x}) - \lambda T - \gamma E(\hat{x}) - (1+\lambda)(p^{**}\mu M'(p^{**}) - \mu M(p^{**})), \tag{5.43}$$

or equivalently,

$$(1+\lambda)(p^*\mu M'(p^*) - \mu M(p^*)) < (1+\lambda)(p^{**}\mu M'(p^{**}) - \mu M(p^{**})). \tag{5.44}$$

Thus, the principal can implement any production level under the best non-linear incentive scheme using less resources on incentives for the agent than under the best linear incentive scheme.

If $M''(p) < 0$, the optimal monitoring accuracy under a non-linear incentive scheme, p^*, may be higher than the optimal

[10] Allgulin (1999) argues that the monitoring function will always be weakly concave if an employer can credibly use mixed strategies when monitoring an employee, i.e. if the employer can randomize between different monitoring accuracies. The analogous reasoning applies when a regulatory agency monitors a polluting firm; the regulated agent's reactions to the probability of detection and expected probability of detection are identical.

monitoring accuracy under a linear incentive scheme, p^{**}. But even if this is the case, it is more costly for the principal to implement $\hat{x}$ if the contract is restricted to be linear. To see this, first note that $p^* > p^{**}$ together with (5.42) implies that

$$p^* M'(p^*) < p^{**} M'(p^{**}). \tag{5.45}$$

Assume now that the inequality (5.44) does not hold, i.e. that,

$$p^* M'(p^*) - M(p^*) \geq p^{**} M'(p^{**}) - M(p^{**}). \tag{5.46}$$

It is then easily seen that the following must be true:

$$M(p^{**}) > M(p^*).$$

But this contradicts the assumption that $M'(p) > 0$. Evidently, the principal is always harmed by restricting the incentive scheme to be linear.

Proposition 5 *The principal can enforce any emission level under a non-linear incentive scheme using less resources on incentives for the agent than under linear incentive schemes.*

Proposition 6 *For any emission level the principal will implement, he wants to monitor more accurately if he is restricted to using linear incentive schemes than if he is not, if $M''(p) \geq 0$.*

Empirical evidence

Most empirical studies of enforcement have focused on the elementary issue of establishing that monitoring actually deters regulatory violation. Not surprisingly, they are unanimous in their affirmative answers. Epple and Visscher (1984), for example, examine the US Coast Guard's enforcement of oil spill regulations. They have a theoretical model in which they assume that the enforcement agency's policies are fixed and that firms react to that policy. Empirically, they estimate the volume of oil spilled in US waters as a function of Coast Guard monitoring activities, which vary by port and over time. Their main finding is that increased monitoring activity results in lower oil spill volume.

Another example is Magat and Viscusi (1990), who study the impact of government inspections on water pollution levels and compliance with standards in the pulp and paper industry in the US. They differ from Epple and Visscher (1984) in that they know the name of the company being inspected. They can therefore use actual inspections of the plant instead of aggregate inspections in the region as an explanatory variable.[11] Consistently with Epple and Visscher (1984), they find that higher levels of monitoring activity result in lower levels of pollution.

Monitoring of pollution at pulp and paper mills has also been studied by Laplante and Rilstone (1996). Like Magat and Viscusi (1990), they use actual inspection rates. In addition to previous results, they also find that the expected inspection rate (or the threat of inspections) induces compliance.[12]

Having established this basic relation, the existing empirical studies that are most relevant here are the ones which examine the relationship between subsidies and compliance. Downing and Kimball (1982) have documented the low penalties for non-compliance rates in US. They find that firms receive cost subsidies in the form of tax breaks and special financing. They note that a cost subsidy will affect the cost-benefit calculus which a firm must undergo when determining the expected cost of compliance in the presence of non-compliance penalties. The idea is very similar to the one presented in this paper: the loss of a subsidy is as severe for a firm as a non-compliance penalty.

Disposal of wastes is interesting to study because unlike many areas of environmental regulation, it has often been subject to a regulatory policy consistent with the optimal command and control policy such as the one presented in this paper: not only does the regulatory agency decide how to price non-compliance and how much resources are devoted to monitoring, but the agency also prices compliance by setting the price of legal disposal. Sullivan (1987) shows that, if the price of legal disposal is too high, the regulatory agency actually encourages illegal disposal. Conversely, one way to encourage legal disposal is to subsidize it. He provides an initial estimate of the optimal subsidy and enforcement budget for hazardous waste disposal and determines the condition under which a subsidy is preferable to increased enforcement and vice versa.

Similar analyses are conducted by Fullerton and Kinnaman (1995) in the context of household garbage, and Sigman (1998) for used oil disposal. These studies identify a positive correlation between the cost of legal disposal and the amount of illegal disposal that is observed. For example, Sigman (1998) estimates that a ban on used oil disposal (requiring that used oil be recycled or reused) will result in 34% of the waste previously disposed of legally being illegally dumped. Since illegal dumping is likely to be worse than the previous method of legal disposal, one cannot say a priori whether a ban on used oil

[11] See footnote 7 for why this is important.

[12] See footnote 10.

disposal is socially beneficial.

Although non-compliance is not present in equilibrium in the model in this paper, the above findings are interesting, because they clearly highlight the trade-off between raising the cost of legal disposal and compliance: a subsidy could be used to reach compliance. In addition to this logic, the theory in this paper presents a new very simple empirical implication. Monitoring and monetary incentives are used as complementary instruments to persuade firms to comply with a regulatory standard. Thus, whenever a subsidy is present, one should expect close monitoring as well.

Final remarks

The answer to the title is a remarkably clear *no!* That is, if the model presented in this paper is an accurate description of the world, market-based incentives do not lower the cost of compliance. The most intuitive explanation for the result is that market-based incentives are closely related to linear incentives. And linear incentives do not use all accessible power; they only carry out the strongest punishment for the maximum deviation.

It is noteworthy that this result is likely to withstand a loosening of the risk-neutrality assumption of the agent, since non-compliance[13] is not present in the equilibrium solution.[14] But if pollution is random, i.e. if it is not perfectly correlated with an agent's action, risk-neutrality would matter. An agent facing an uncertain negative payoff would prefer to pay the expected dollar value of the payoff as an insurance premium rather than face the uncertain situation. Consequently, there is a trade-off between the agent's desire for insurance and his need for incentives, and perhaps a linear regulatory policy may be close to the optimal mix.

Perhaps a better way of dealing with both risk-neutrality and the random nature of pollution is to regulate the agent's action (i.e. take precautions to avoid pollution) instead.[15] This way, the regulatory agency can go back to a situation in which non-compliance is not present in equilibrium, by bearing all

[13] In for example Sandmo (1998), where non compliance is present in the equilibrium solution, risk-neutrality matters.

[14] However, risk-neutrality of firms is a very natural assumption; investors use portfolio methods to spread their risks and do not want firms themselves to take risk into consideration.

[15] Cohen (1987) performs such an extension on the Becker (1968) model.

the risk itself. Hence, the agent will not need any insurance and the regulatory agency can impose a non-linear regulatory policy. This policy may be very similar to the one when the level of pollution is regulated: the regulatory agency sets a target of how much effort the polluting firm must devote to prevent pollution, and the polluting firm pays the compliance fee if it meets or exceeds the target and the maximum penalty[16] otherwise.[17]

Of course, the result must not be misunderstood; one should not jump to the conclusion that nothing good comes out of a market. The main contribution of this paper is to provide a framework in which the *incentives* from the best non-linear regulatory policy and the best linear regulatory policy can be compared. Even though a market for emission permits provides inferior (to expensive) incentives, there are other qualities associated with markets which may be desirable. For example, it is likely that a market is more dynamic, i.e. reacts and reallocates faster.than a regulatory agency. One may also argue that the equal treatment of firms in a market is cheaper to handle.[18] Equal treatment of firms may also be considered more fair. The list may be made long, but it is important to understand that there is a cost associated with the construction of a marketplace for a regulated good if the good was not possible to trade in the first place.

In this paper, I have neglected the issue of illegal arbitrage between regulated agents. It is well known that a non-linear regulatory policy generates a demand for a market. If it is very costly to prevent an illegal market from emerging, which is the case if the good is easy to trade, the regulatory agency may yield and instead form a legal market. Gasoline is an excellent example of such a good, and attempts from regulatory agencies to deviate from a linear tax policy are almost always followed

[16] In the parlance of the law and economics literature, instead of strict liability standard, this would be a negligence-based penalty.

[17] Of course, this is a very stylized example. Generally, if no measure is perfectly correlated with the regulated agent's action, the regulatory policy should be made contingent on all measures (See Holmström; 1979).

[18] A standard claim is that a tax levied at a uniform rate on all firms is optimal if it is adjusted so that the aggregate emission level is optimal. Quotas, on the other hand, have to be set for each individual firm. Therefore, it is argued that a uniform tax is cheaper to control than a quota: the regulatory agency does not have to collect information about individual production functions (See for example Sandmo(1998)).

However, this reasoning does not apply in the setting in this chapter. The optimal uniform emission fee would not be equal across firms if firms have different production functions or differ in how much the regulatory agency can extract from them in case they don't comply with the policy. Thus, even for the case with a uniform fee, there is a motive for the regulatory agency to collect information about individual production functions.

by emerging "black" markets. Ironically, sometimes the authorities seem to act exactly in opposition to the prescription in this paper: they try to impose non-linear regulatory policies on tradable externalities and they fabricate marketplaces for non-tradable externalities.

References

Allgulin, Magnus (1999): Monitoring and Pay: General results, *SSE/EFI Working Paper Series in Economics and Finance* No. 340, Stockholm School of Economics.

Allgulin, Magnus and Tore Ellingsen (1998): Monitoring and Pay, *Working Paper Series in Economics and Finance* No. 245, Stockholm School of Economics.

Becker, Gary (1968): Crime and Punishment: An Economic Approach, *Journal of Political Economy* 76, 169-217.

Cohen, M. A. (1987): Optimal Enforcement Strategy to Prevent Oil Spills: An Application of Principal-Agent Model with 'Moral Hazard', *Journal of Law and Economics* 30(1), 23-51.

Demougin, Dominique and Claude Fluet (1997): Monitoring versus Incentives: Substitutes or Complements?, Mimeo, Université du Québec à Montréal.

Downing, P. and J. Kimball (1982): Enforcing Pollution Control Laws in the United States, *Policy Studies Journal* 11, no. 1.

Downing, P. and W. D. Watson (1974): The Economics of Enforcing Air Pollution Controls, *Journal of Environmental Economics and Management* 1, 219-236.

Epple, D and M. Visscher (1984): Environmental Pollution: Modeling Occurrence, Detection and Deterrence, *Journal of Law and Economics* 27, 29-60.

Fullerton, D. and T. C. Kinnaman (1995): Garbage, Recycling, and Illicit Burning or Dumping, *Journal of Environmental Economics and Management* 29, 78-91.

Harford, J. D. (1978): Firm Behavior under Imperfectly Enforceable Pollution Standards and Taxes, *Journal of Environmental Economics and Management* 5, 26-43.

Holmström, Bengt (1979): Moral Hazard and Observability, *Bell Journal of Economics* 10, 74-91.

Laplante, B. and P. Rilstone (1996): Environmental Inspections and Emissions of the Pulp and Paper Industry in Quebec, *Journal of Environmental Economics and Management* 31(1), 19-36.

Magat, W. and W. K. Viscusi (1990): Effectiveness of the EPA's Regulatory Enforcement: The Case of Industrial Effluent Standards, *Journal of Law and Economics* 33, 331-360.

Nowell, C. and J. Shogren (1994): Challenging the Enforcement of Environmental Regulation, *Journal of Regulatory Economics* 6, 265-282.

Sandmo, Agnar (1998): Efficient Environmental Policy with Imperfect Compliance, *Discussion paper* No. 8, Norwegian School of Economics and Business Administration.

Shavell, S. (1979): Risk Sharing and Incentives in the Principal and Agent Relationship, *Bell Journal of Economics* 10, 55-73.

Sigman, H. (1998): Midnight Dumping: Public Policies and Illegal Disposal of Used Oil, *RAND Journal of Economics* 29(1), 155-178.

Sullivan, A. M. (1987): Policy Options for Toxics Disposal: Laissez-Faire, Subsidization, and Enforcement, *Journal of Environmental Economics and Management* 14, 58-71.

Index

EFI
The Economic Research Institute

Reports since 1995

Published in the language indicated by the title

1999

Andersson, P., Experto Credite: Three Papers on Experienced Decision Makers.
Ekman, G., Från text till batong – Om poliser, busar och svennar.
Eliasson, A-C., Smooth Transitions in Macroeconomic Relationships.
Flink, H., Gunnarsson, J., Wahlund, R., Svenska hushållens sparande och skuldsättning– ett konsumentbeteende-perspektiv
Gunnarsson, J., Portfolio-Based Segmentation and Consumer Behavior: Empirical Evidence and Methodological Issues.
Hamrefors, S., Spontaneous Environmental Scanning.
Helgesson, C-F., Making a Natural Monopoly: The Configuration of a Techno-Economic Order in Swedish Telecommunications.
Japanese Production Management in Sunrise or Sunset. Christer Karlsson (editor).
Jönsson, B., Jönsson, L., Kobelt, G., Modelling Disease Progression And the Effect of Treatment in Secondary Progressive MS. Research Report.
Lindé, J., Essays on the Effects of Fiscal and Monetary Policy.
Ljunggren, U., Indikatorer i grundskolan i Stockholms stad före stadsdelsnämndsreformen – en kartläggning.
Lundbergh, S., Modelling Economic High-Frequency Time Series.
Mägi, A., Store Loyalty? An Empirical Study of Grocery Shopping.
Mölleryd, B.G., Entrepreneurship in Technological Systems – the Development of Mobile Telephony in Sweden.
Nilsson, K., Ledtider för ledningsinformation.
Rognes, J., Telecommuting – Organisational Impact of Home Based – Telecommuting.
Sandström, M., Evaluating the Benefits and Effectiveness of Public Policy.
Sjöstrand, S-E., Sandberg, J., Tyrstrup, M., (red) Osynlig Företagsledning.
Skalin, J., Modelling Macroeconomic Time Series with Smooth Transition Autoregressions.
Spagnolo, G., Essays on Managerial Incentives and Product-Market Competition.
Strauss, T., Governance and Structural Adjustment Programs: Effects on Investment, Growth and Income Distribution.
Svedberg Nilsson, K., Effektiva företag? En studie av hur privatiserade organisationer konstrueras.
Söderström, U., Monetary Policy under Uncertainty.
Werr, A., The Language of Change The Roles of Methods in the Work of Management Consultants.
Wijkström, F., Svenskt organisationsliv – framväxten av en ideell sektor.

1998

Andersson, M., On Testing and Forecasting in Fractionally Integrated Time Series Models.
Berg-Suurwee, U., Styrning av kultur- och fritidsförvaltning innan stadsdelsnämndsreformen

Berg-Suurwee, U., Nyckeltal avseende kultur- och fritidsförvaltning innan stadsdelsnämndsreformen.
Bergström, F., Essays on the Political Economy of Industrial Policy.
Bild, M., Valuation of Takeovers.
Charpentier, C., Samuelson, L.A., Effekter av en sjukvårdsreform – en analys av Stockholmsmodellen.
Eriksson-Skoog, G., The Soft Budget Constraint: The Emergence, Persistence and Logic of an Institution. The Case of Tanzania 1967-1992.
Gredenhoff, M., Bootstrap Inference in Time Series Econometrics.
Ioannidis, D., I nationens tjänst? Strategisk handling i politisk miljö – en nationell teleoperatörs interorganisatoriska, strategiska utveckling.
Johansson, S., Savings Investment, and Economic Reforms in Developing Countries.
Levin, J., Essays in Company Valuation.
Ljunggren, U., Styrning av grundskolan i Stockholms stad innan stadsdelsnämndsreformen.
Mattsson, S., Från stat till marknad – effekter på nätverksrelationer vid en bolagiseringsreform.
Nyberg, A., Innovation in Distribution Channels – An Evolutionary Approach.
Olsson, P., Studies in Company Valuation.
Reneby, J., Pricing Corporate Securities.
Roszbach, K., Essays on Banking Credit and Interest Rates.
Runsten, M., The Association Between Accounting Information and Stock Prices. Model development and empirical tests based on Swedish Data.
Segendorff, B., Essays on Bargaining and Delegation.
Sjöberg, L., Bagozzi, R., Ingvar, D.H., Will and Economic Behavior.
Sjögren, A., Perspectives on Human Capital: Economic Growth, Occupational Choice and Intergenerational Mobility.
Studier i kostnadsintäktsanalys, red Jennergren, P.
Söderholm, J., Målstyrning av decentraliserade organisationer. Styrning mot finansiella och icke-finansiella mål.
Thorburn, K., Cash Auction Bankruptcy and Corporate Restructuring
Wijkström, F., Different Faces of Civil Society.
Zethraeus, N., Essays on Economic Evaluation in Health Care. Evaluation of Hormone Replacement Therapy and Uncertainty in Economic Evaluations.

1997

Alexius, A., Essays on Exchange Rates, Prices and Interest Rates.
Andersson, B., Essays on the Swedish Electricity Market.
Berggren, N., Essays in Constitutional Economics.
Ericsson, J., Credit Risk in Corporate Securities and Derivatives. Valuation and Optimal Capital Structure Choice.
Charpentier, C., Budgeteringens roller, aktörer och effekter. En studie av budgetprocesserna i en offentlig organisation.
De Geer, H., Silfverberg, G., Citizens´Trust and Authorities´ Choices, A Report from The Fourth International Conference on Ethics in the Public Service, Stockholm June 15-18, 1994.
Friberg, R., Prices, Profits and Exchange Rates.
Från optionsprissättning till konkurslagstiftning. red. Bergström, C., Björk, T.
Hagerud, G.E., A New Non-Linear GARCH Model.
Haksar, A., Environmental Effects of Economywide Policies: Case Studies of Costa Rica and Sri Lanka.
He, C., Statistical Properties of Garch Processes.

Holmgren, M., Datorbaserat kontrollrum inom processindustrin; erfarenheter i ett tidsperspektiv.
Jennergren, P., Tutorial on the McKinsey Model for Valuation of Companies.
Lagerlöf, J., Essays on Political Economy, Information, and Welfare.
Lange, F., Wahlund, R., Planerade och oplanerade köp - Konsumenternas planering och köp av dagligvaror.
Löthgren, M., Essays on Efficiency and Productivity; Contributions on Bootstrap, DEA and Stochastic Frontier Models.
Nilsson, H.E., Using Conceptual Data Modelling in Management Accounting: A Case Study.
Sjöberg, L., Ramsberg, J., En analys av en samhällsekonomisk bedömning av ändrade säkerhetsföreskrifter rörande heta arbeten.
Säfvenblad, P., Price Formation in Multi-Asset Securities Markets.
Sällström, S., On the Dynamics of Price Quality.
Södergren, B., På väg mot en horisontell organisation? Erfarenheter från näringslivet av decentralisering och därefter.
Tambour, M., Essays on Performance Measurement in Health Care.
Thorén, B., Berg-Surwee, U., Områdesarbete i Östra Hökarängen - ett försök att studera effekter av decentralisering.
Zhang Gang., Chinese Rural Enterprises Between Plan and Market.
Åhlström, P., Sequences in the Profess of Adopting Lean Production.
Åkesson, G., Företagsledning i strategiskt vakuum. Om aktörer och förändringsprocesser.
Åsbrink, S., Nonlinearities and Regime Shifts in Financial Time Series.

1996

Advancing your Business. People and Information Systems in Concert.
red. Lundeberg, M., Sundgren, B.
Att föra verksamheten framåt. Människor och informationssystem i samverkan.
red. Lundeberg, M., Sundgren, B.
Andersson, P., Concurrence, Transition and Evolution - Perspectives of Industrial Marketing Change Processes.
Andersson, P., The Emergence and Change of Pharmacia Biotech 1959-1995. The Power of the Slow Flow and the Drama of Great Events.
Asplund, M., Essays in Industrial Economics.
Delmar, F., Entrepreneurial Behavior & Business Performance.
Edlund, L., The Marriage Market: How Do You Compare?
Gunnarsson, J., Three Studies of Financial Behavior. Research Report.
Hedborg, A., Studies of Framing, Judgment and Choice.
Holgersson, C., Höök, P., Ledarutveckling för kvinnor - Uppföljning av en satsning på Volvo - Research Report.
Holmberg, C., Stores and Consumers - Two Perspectives on Food Purchasing.
Håkansson, P., Wahlund, R., Varumärken. Från teori till praktik.
Karlsson, A., The Family Business as an Heirloom. Research Report.
Linghag, S., Man är handelsstudent. Research Report.
Molin, J., Essays on Corporate Finance and Governance.
Mägi, A., The French Food Retailing Industry - A Descriptive Study.
Mölleryd, B., Så byggdes en världsindustri - Entreprenörskapets betydelse för svensk mobiltelefoni. Research Report.
Nielsen, S., Omkostningskalkulation for avancerede produktions-omgivelser - en sammenligning af stokastiske og deterministiske omkost-ningskalkulationsmodeller.
Normark, P., Danielsson, L., Larsson, A., Lundblad, P., Kooperativa nyckeltal. Research Report.

Sandin, R., Heterogeneity in Oligopoly: Theories and Tests.

Sandén-Håkansson, U., Från kampanjmål till mediemix - en studie av samarbete mellan annonsörer, reklambyråer och mediebyråer. Research Report.

Stein, J., Söderlund, M., Framgång i arbetet, strategier för att utföra arbetsuppgifterna, arbetsuppgifternas karaktär och utbildningskvalitet. En empirisk studie av civilekonomer. Research Report.

Strömberg, P., Thorburn, K., An Empirical Investigation of Swedish Corporations in Liquidation Bankruptcy. Research Report.

Söderlund, M., Och ge oss den nöjda kunden. En studie av kundtillfredsställelse och dess orsaker och effekter. Research Report.

Thodenius, B., Användningen av ledningsinformationssystem i Sverige: Lägesbild 1995. Research Report.

Ulfsdotter, U., Internationalisering för expansion eller hemmamarknadsförsvar? De nordiska marknaderna för fruktyoghurt 1982-1994.

Westelius, A., A Study of Patterns of Communication in Management Accounting and Control Projects.

Wijkström, F., Den svenska ideella sektorn och pengarna. Research Report.

Örtendahl, M., Health and Time - A Problem of Discounting.

1995

Becker, T., Essays on Stochastic Fiscal Policy, Public Debt and Private Consumption.

Blomberg, J., Ordning och kaos i projektsamarbete - en socialfenomenologisk upplösning av en organisationsteoretisk paradox.

Brodin, B., Lundkvist, L., Sjöstrand, S-E., Östman, L., Styrelsearbete i koncerner

Brännström, T., Bias Approximation and Reduction in Vector Autoregressive Models. Research Report.

Ekonomisk politik i omvandling. red. Jonung, L.

Gunnarsson, J., Wahlund, R., Hushållens finansiella strategier. En explorativ studie.

Höök, P., Chefsutveckling ur könsperspektiv - Mentorskap och nätverk på Vattenfall - Research Report.

Levin, J., Olsson, P., Looking Beyond the Horizon and Other Issues in Company Valuation. Research Report.

Mägi, A., Customer Statisfaction in a Store Performance Framework. Research Report.

Persson, P-G., Modeling the Impact of Sales Promotion on Store Profits.

Roman, L., Institutions in Transition. A Study of Vietnamese Banking.

Sandberg, J., Statistisk metod - dess vetenskapliga hemvist, grundläggande principer och möjligheter inom samhällsvetenskapen. Research Report.

Sandberg, J., How Do We Justify Knowledge Produced by Interpretative Approaches? Research Report.

Schuster, W., Redovisning av konvertibla skuldebrev och konvertibla vinstandelsbevis - klassificering och värdering.

Söderberg, K., Farmartjänsten - Ny kooperation inom lantbruket. Research Report.

Söderqvist, T., Benefit Estimation in the Case of Nonmarket Goods. Four Essays on Reductions of Health Risks Due to Residential Radon Radiation.

Thorén, B., Användning av information vid ekonomisk styrning - månadsrapporter och andra informationskällor.