SNF REPORT NO. 58/01

Essays on Public Policy and Private Incentives

by

Egil Kjerstad

Financial support from the Research Council of Norway, The Ministry of Labour, the Ministry of Health, Programme for Health Economics in Bergen (HEB) and Foundation for Research in Economics and Business Administration (SNF) is gratefully acknowledged.

FOUNDATION FOR RESEARCH IN ECONOMICS AND BUSINESS ADMINISTRATION
BERGEN, DECEMBER 2001

ISBN 82-491-0184-7
ISSN 0803-4036

Essays
on
Public Policy
and
Private Incentives

by

Egil Kjerstad

Thesis submitted for the degree of dr. polit
University of Bergen

July 2001

Foundation for Research in Economics and Business Administration (SNF)
and Programme for Health Economics in Bergen (HEB)

Acknowledgements

I would like to thank a number of people, who in various ways have contributed to the works presented here.

First of all, I would like to thank my supervisors Jan Erik Askildsen and Kjell Vaage. It has been a true pleasure to work with them. They have throughout the last two years showed continuous interest in the research projects that lie behind the essays presented here. They have challenged me to come back with more convincing arguments and to produce new results, steadily making the essays better. Their 'open door policy' has contributed to the projects advancing fairly quickly. Needless to say, I have learned a lot from them both. Equally much, I would like to thank Jan Erik and Kjell for their constantly good spirits and fabulous sense of humour. I feel that we have managed to mix 'business and pleasure' in an optimal way. Thanks to both for many good laughs, on all kinds of topics, many of which do not deserve publication!

My co-authors, Steinar Vagstad and Arild Aakvik, certainly do deserve tons of thanks, too. I have known Steinar for years and it was inspiring to write an article with him early on in the work with the thesis. Our collaboration ended up in my first publication and I picked up a lot through that process. Steinar is certainly also a person that can throw a joke. Thanks for your easygoing manners.

I have learned a lot from Arild, too. He had the patience to explain methodological issues to me and although our collaboration is manifested in one essay only, he inspired me to throw myself into fairly deep water resulting in two other empirical essays. And yes, we have had our laughs, too.

During the Summer and Autumn of 2000, I spent an inspiring time at University of York and had the pleasure of getting to know many good colleagues at Department of Economics and Related Studies and Centre for Health Economics. In particular I would like to mention Andrew Jones, to whom I am very grateful. Our many discussions over a cup of coffee at the Coffee Shop at Alcuin College gave me the opportunity to test my understanding of methodological issues (and to discuss the latest movies on City Screen, my favourite cinema).

I would also like to thank Kåre P. Hagen for encouraging me to get back into research again and for having taught me a thing or two about economics over the years. Many thanks to Per Heum who offered me a research post at SNF dedicated to health economic research. Thanks to Alf Erling Risa and Espen Bratberg for pieces of advice and encouragement. Kjell Erik Lommerud and Trond Olsen also deserve thanks for support and advice in an earlier phase of this work. A collective thanks to the staff at Programme for Health Economics in Bergen (HEB) and to all my colleagues at SNF for support in different ways.

Financial support from HEB and the Research Council of Norway is gratefully acknowledged.

Last, but not least, I would like to thank Trude for all her love and encouragement over the years and finally, many thanks to my son, Ask, for pushing me to finalise the thesis just in time of his birth.

Egil Kjerstad
Bergen, July 2001

TABLE OF CONTENT:

Chapter 1

"It Must Be Some Kind of Asymmetric Information" - An Introduction to the Essays

"It Must Be Some Kind of Asymmetric Information" - An Introduction to the Essays

by
Egil Kjerstad*

* Foundation for Research in Economics and Business Administration (SNF) and Programme for Health Economics in Bergen (HEB).

1. INTRODUCTION

For the last ten to fifteen years there has been a steady international move towards changing the way health care is delivered financed and regulated.[1] Ham (2000) makes an interesting observation commenting on the development. Countries which have traditionally relied on a market in health care are making greater use of regulation and planning. In contrast, those countries which traditionally have relied on regulation and planning are moving towards a more competitive approach. In no country is there complete satisfaction with existing methods of finance and delivery, and there is a search for new policy instruments.

The main question asked in this thesis is how market-based governance structures influence the resource utilisation in public sector in general and the health care sector in particular.

Why is this question an interesting research topic? Public sector in general and the health care sector in particular have been under a 'siege' for the last twenty years. In most OECD countries the size of and the composition of services provided by the Welfare State have been questioned. The extent of public ownership to firms, organisations and real resources (oil, gas, land etc) has been questioned, too. In particular, public monopolies in the telecom, energy and transport sectors have been the focus of much criticism.

The criticism against the Welfare State and public ownership has centred on three main themes. Firstly, the Welfare State changes individual citizen's incentive in a negative way, particularly the incentives to find work if unemployed. Secondly, public ownership in general is not conducted in a professional and market like manner causing inefficiencies (organisational slack, lack of focus on consumer satisfaction etc.). Thirdly, lack of competition or market-like governance structures also causes the public

[1] See Smith (ed) (2000) for a comprehensive discussion of the introduction of 'internal markets' in the National Health Service (NHS) in UK. Abbott (ed) (1995) gives insight into fairly recent changes of health care policy and regulations in USA. For a somewhat outdated but still illustrative comparative study of health reforms and health care systems in OECD countries, see OECD (1992) and OECD (1993a, b).

sector to under-perform (efficiency, quality of services etc). See for instance Bishop, Kay and Mayer (1995) for a discussion of the UK system of regulations introduced as part of the structural changes in 1980s and Clarke and Pitelis (1993) discussion of privatisation programs/projects world wide.

The main arguments in favour of the Welfare State and the public sector at large have been basically founded on distributional issues. The welfare state is a social insurance system giving for instance security in case of ill health; financial support in case of redundancy[2] and primary and secondary education free of charge. Although some citizens may 'misuse' the services, one argues that generally this is not the case. Furthermore, public ownership has been used to achieve higher level of employment and a fairer distribution of jobs and services.

Many of the discussions of the Welfare State, including its size and scope, have centred on efficiency issues versus distributional issues. Those who advocate market like mechanisms like outsourcing, use of auctions, performance pay and privatisation of some services and enterprises have been criticised for wanting to destroy the social insurance system it has taken years to establish, basically leaving distributional issues to the market. For instance, Esping-Andersen (1999) argues that a strategy based purely on deregulation and privatisation cannot be welfare and efficiency optimising: "..., if unregulated labour markets create people who are poor, then it cannot be a welfare-optimizing model if those very same people find themselves excluded from social protection (p 178)."

On the other hand, those who advocate in favour of status quo have been criticised for only having the interest of public sector employees in mind.

However, in the last decade is has been more common to argue in favour of and in fact implement market-like mechanisms in the public sector in an attempt to achieve a better resource allocation and utilisation. Rather than letting the Welfare State wither away,

[2] In Esping-Andresen and Regini (2000) the empirical connections between unemployment and worker protection are discussed. Evidence suggests that a radical strategy of deregulation of labour markets probably cause more harms than benefits for European economic performance.

one argues that higher efficiency is the only way to keep the Welfare State 'alive' in the future. Giddens (1998), for instance, argues for a third way, guiding the renewal of social democracy. Prime Minister Tony Blair in the UK has used this slogan to promote the Labour Party's policies. Basically, many of the policies rely quite explicitly on market-like mechanisms or market-based governance structures.

Thus, studying the effect of market-like governance structures in the health care sector is an important and interesting exercise both because of its day-to-day relevance and because of the theoretical and methodological challenges facing such research.

2. THE THEORETICAL BASIS

The discussions in this thesis are based on economic theory developed in the last twenty years. The title of this essay, or rather the title of the comments on the essays to follow, stems from Rasmusen (1994). Rasmusen starts out his description of categories of asymmetric information models pointing out that the economist's generic answer to someone who brought up peculiar behaviour which seemed to contradict basic theory used to be "It must be some kind of price discrimination". Today, writes Rasmusen, we have a new answer: "It must be some kind of asymmetric information". And surely, due to the effort of Jean Tirole, Jean-Jacques Laffont, Bengt Holmstrom, Oliver Hart, Robert Wilson and Paul Milgrom to mention but a few of the researchers in the field of contract theory[3], it is difficult not to confront questions ranging from political economy to internal organization issues with that 'opening' statement.

Going from the "It must be some kind of asymmetric information" to formulating the basic asymmetric information problem at hand, is not necessarily a trivial task, though. Contract theory covers a multitude of possible models and games and research on contracts has progressed along several different lines, each with its own particular interest. Four such lines of research are regulation theory (Laffont and Tirole (1993)),

[3] See for instance Hart and Holmstrom (1987), Milgrom and Roberts (1992), Tirole (1992), Laffont and Tirole (1993), Laffont (1994) and Laffont (2000).

the internal organization of the firm (see Tirole (1992) and Gibbons (1997)), the workings of the labour market (see for instance Hillier (1997) for a comparatively short and straightforward discussion) and the workings of financial markets (see Harris and Raviv (1992) for an overview of the literature). The essays presented here draw extensively on research represented in the first three lines of research while the last, financial markets and what determines how firms finance their investments and operations, is not, at least not directly, incorporated in the essays.

Contract theory is concerned with how to motivate individuals, although in many cases what is modelled is how to motivate a firm or an organization, implicitly assuming that the management on their part is able to motivate their employees accordingly. The motivation problem arises because individuals follow their own private interests, which are rarely perfectly aligned with the interests of say, society as a whole. Milgrom and Roberts (1992) write ".. this is a caricature of actual human motivation and behaviour. Yet it is a powerful analytic simplification. Given this assumption, the motivation problem becomes one of arranging affairs so that, as far as possible, selfish individual actions take proper account, not only of how decision makers are affected by a decision, but of how others are affected as well (pp 126-27)".

The way to "arrange affairs" is through a contract, which may or may not have a legal status.[4] Obviously, the basic aim of a contract is to align incentives, i.e. to protect the contracting parties' interests. On the other hand, the alignment of interests is not necessarily a simple exercise, neither in theory nor in practice. When a central planner or a principal is aiming at maximising social surplus from an intervention or a regulatory effort, he is most likely faced with three major forms of constraints: (i) the existence of asymmetric information, (ii) lack of commitment and/or (iii) imperfect regulators or 'middlemen'. Depending on the sector, market or industry, these constraints may play different roles or have different impact on the formulation of the optimal contract or the specific governance structure chosen in practice.

[4] In fact, as Milgrom and Roberts (1992) point out, contracts may be completely unarticulated and implicit, with no power of law behind them.

A discussion of the first type of constraints brings us back to the quote from Rasmusen (1994) and we obviously have to push things further and ask ourselves what kind of asymmetric information problem with which we are faced. Are we studying a problem of moral hazard, i.e. a situation where the agent's action is not verifiable, or to put it differently, a situation where the agent receives private information after the relationship has been initiated? Or is the problem of the adverse selection kind, i.e. the agent holds private information before the relationship has begun? Laffont and Tirole (1993) refer to moral hazard as a situation in which there exist endogenous variables that are not observed by the principal. Adverse selection arises when the agent (e.g. the firm, the organization, the individual) has more information than the principal about some exogenous variables.

Studying a specific principal-agent relationship, it is important to clarify the information structure, i.e. the sequence of the decisions or the order of moves. To illustrate the importance of the information structure, an adverse selection game can in many instances more properly be looked upon as a game of signalling, i.e. the agent can send a signal that is observed by the principal before the principal offers a contract. Or it may be the other way round, that the principal first offers a contract and the agent responds, i.e. a screening game.[5]

Turning to the second type of constraints – lack of commitment - many contributions are based on the condition that perfect, complete contracts are possible to make.[6] Implementation of the optimal contract requires that a principal is able to commit himself/herself to the terms of the agreement before the agent makes the participation decisions (e.g. entry, investment). In practice, contracts are often short-term, so a particular long-term relationship between contracting parties is often run by a series of short-term contacts. Why is this so? Not all plans can be described in a complete, enforceable contract, and this leads to problems of imperfect commitment. Even if a contingency can be foreseen and planned for and contractual commitments can be

[5] Some authors prefer the use of the term 'sorting' models and divide these types of models into 'screening' models and 'signalling' models.

[6] One important qualification, making the constraint somewhat less 'suffocating', is that commitment can be thought of as a situation in which the parties are free to agree to modify the contract to their mutual advantage. In other words, the contract will be implemented in the future if one them wishes to do so.

enforced, one of the parties may have relevant private information before the contract is signed. Another possibility is that the information available ex post is not adequate to tell whether the terms of the contract are fulfilled. Thus, the basic challenge with contracting is that contracts are bound to be incomplete and contract incompleteness is a source of the motivation problem. First, one of the parties may renege on the deal. Second, representing a more subtle point, it may be Pareto improving to renegotiate contracts ex post. Knowing this, though, it might be difficult to agree on terms in the first place.

Laffont and Tirole (1993) argue that there are two basic reasons why governments do not commit. The first part of their argument is, as already pointed out, difficulty of "signing complete state-contingent contracts in an uncertain environment". The difficulties of signing complete contracts are due to transaction costs of one kind or the other. Although transaction costs are an important explanation of short-run commitment, transaction costs cannot by themselves account for the array of constitutional and administrative rules that prohibit long-term commitment. Laffont and Tirole (1993) therefore argue in favour of a second foundation for noncommitment. Their main idea is that "nonbenevolent governments can do more harm if they are allowed to commit. Short commitments allow wrong policies to be corrected by future administrations". Although the basic point made by Laffont and Tirole (1993) in the quotation above is easy to agree with, it is of course possible to question it: From whose point of view is the present policy harmful or wrong? And as regards corrections to policies, to whom are the changes harmful? It is not evident that changes in policies based on short-term commitment are 'good' changes for all, even if this is the connotation in Laffont and Tirole's second foundation of non-commitment.

The last major type of constraints is what can be termed the politics of regulation or to narrow it down, regulatory capture. The basic idea is that interest groups can play a strong part in the formation of public policy and, one could add, the ability to *change* public policy already implemented. Stigler (1971), the 'founder' of the Chicago school, and later, among others Becker (1983), developed a positive theory on how interest groups could shape public policy. The Chicago school also showed the lack of empirical relevance of the public interest view of public policy, i.e. that politicians and public bureaucrats act as benevolent administrators of laws and rules. For instance, politicians and regulators can further their own private interests and be captured by firms or organizations to amend laws and rules in favour of some parties on the expense of others. The Chicago school approach concludes that ex post efficiency is achieved since there is no reason why politicians would not maximize the sum to be distributed.

Laffont (2000) is a recent treatment of political institutions. Laffont explores in depth the use of principal-agent models and their developments for understanding constitutional design. For instance, Laffont shows that the Chicago approach is limited by the fact that the supply side of political favors, i.e. why and how politicians and voters offer these favors, is left unmodeled. Laffont quotes Becker (1983), who acknowledges this limitation: "I shall not try to model how different political systems translate the activities of pressure groups into political influence [..] an explicit modeling of coalition formation would surely add to the power of the approach". Laffont (2000) first models the collusion process itself based on Tirole's (1986) supervision model. Information is verifiable (hard), i.e. a complete contracting approach, and the optimal organizational responses to collusion are characterized. Laffont extends this analysis by proposing a new methodology for characterizing the optimal constitutional response to the activities of interest groups when their private information is soft and cannot be verified by a third party, i.e. the constitutional design is endogenous.

What the theory on regulatory capture teaches us is that 'gaming' between levels of government and the institution to be governed is likely. As such, the 'capture' hypothesis, as the related theme of non commitment, should have bearing on the optimal regulatory scheme. In practice, both regulatory capture and non commitment may have consequences for the functioning of 'real life' regulatory efforts (broadly defined), like the effects of changing the way hospitals are financed or the strategies of suppliers participating in an auction, as discussed in two of the essays to follow.

What follows next is a short presentation of each essay with the aim of fitting them into the broader research program described above.

3. A BRIEF OVERVIEW OF THE ESSAYS

The four essays to follow are predominately occupied with policy issues. For instance, all essays are inspired by public sector reforms implemented across OECD member states during the last ten to fifteen years. All essays could have had the subtitle 'public policy versus private incentives' which is basically the quintessential conflict in a world of asymmetric information *and* non-aligned objectives. Only one essay has been given that subtitle, though.

Three out of four essays are applied econometric work concerned with analysing and describing effects of reforms made in the health care and social care sectors. The fourth paper is a theoretical work comparing different types of auction designs.

3.1 *"Prospective Funding of Somatic Hospitals in Norway – Incentives for Higher Production?"*

In the first paper - *"Prospective Funding of Somatic Hospitals in Norway – Incentives for Higher Production?"* - the aim is to evaluate the effect of introducing a prospective payment system (PPS) at hospitals in Norway. The Government shifted from a traditional block grant system to a Diagnosis Related Group (DRG) based prospective payment system in 1997. Dranova and Satterthwaite (2000) draw attention to Shleifer's (1985) model of "yardstick competition" when explaining the perceived incentive effects of PPS. Assuming that all hospitals are local monopolists, the PPS forces each hospital to compete against average prices based on average treatment costs across all hospitals. Thus, when treating a patient with a given diagnosis (or a patient classified in one Diagnosis Related Group), the hospital is faced with a fixed-price contract. We know from theory that fixed-price contracts give strong incentives to use the most cost efficient technology, or to frame it differently, hospitals are given incentive to exert first best effort. Cost-plus contracts, on the other hand, only give weak incentives to increase cost efficiency. In the essay a theoretical model is developed that shows that a PPS system gives incentives to cut costs by for example reducing length of stay; by reducing quality; by reducing intensity of nursing or by reducing capacity. It may be the case that a hospital will treat more patients, but not necessarily so. Rather it may be more 'profitable' to reduce the number of treated patients depending on the cost structure and the type of patients available to the hospital. In a theoretical model it is shown that a PPS system can influence differently the incentives to produce DRG points and number of patients. The explicit goal of the reform though, as stated by the Government, was to create incentives to increase the number of treated patients. Testing whether this is the case is the basic aim of the paper.

One important feature of the reform is that it merely guides the transfers of funds from central government to the local authorities that own the hospitals, not to the hospitals directly. Some local authorities have mimicked the reform locally while others have not, which makes it possible to divide the hospitals into an experiment group and a comparison group. It is argued that fixed-effects models are suitable specifications of this evaluation study, handling selection bias and the influence of unobservable explanatory variables in a consistent manner.

Fixed-price type of contracts might lower hospitals' incentive to deliver quality while strengthening the incentive to curb costs. The Government's goal on the other hand was to achieve higher production without renouncing on quality. In the analysis, average length of stay and number of readmissions enter as explanatory variables. Interpreting these variables as quality indicators makes it possible to discuss whether the reform has had a negative effect on quality as predicted in the literature and in the model developed in the essay.

Commitment, or rather the consequences of not being able to commit, is also discussed as one of the possible reasons why such a funding system may be less successful than hoped for by the Ministry of Health. Under the Norwegian PPS the Ministry's lack of commitment to stick to the overall budgets combined with 'political gaming' involving local politicians and Parliament pressing the Government to be less restrictive is a well-known feature. Incentives to curb costs may be severely weakened if it is possible to influence policy makers to give additional funding during the budget year.

The analysis shows that there are differences in performance between the two groups of hospitals. The experiment group appears to produce a greater change in number of treated patients and DRG points produced compared to the comparison group of hospitals.

3.2 *"Skill Formation among Vocational Rehabilitation Clients – Public Policy vs Private Incentives"*

In the second paper - *"Skill Formation among VR Clients – Public Policy vs Private Incentives"* – co-authored with Arild Aakvik, the aim is to study closer the self-selection process behind enrolment in vocational rehabilitation. From the Government's point of view the chain of thought is that training will increase VR clients' probability of entering the labour market, reducing the spending of social security funds in the future. To create incentives for the individual to invest in human capital, the Government bears the direct costs of training. In other words, the client does not pay the tuition fee. On the other hand, it is up to the client to decide whether to participate or not and in the case of participation, what kind of training to opt for. Although the Government has all the good intentions offering training for free, tuition fees are only one of the cost components that may form the private incentives to invest in human capital. Other cost components are monetary opportunity costs and non-monetary (disutility of effort) costs of training. It is argued that both these cost components are important in an individual's decision making process. While the first component is based on the human capital investment models, the latter type of cost is defined in the spirit of signalling models. The assumption made in labour market signalling models is that persons of low ability find signalling through training more costly than do high-ability persons. In that sense the disutility cost is private information and represents an adverse selection variable.

The cost components are not directly observable but we argue that some of the background characteristics of the individual clients can be interpreted as indicators of both monetary opportunity cost and disutility costs. We test whether there are significant differences between three groups of clients: non-participation, work related training programs and educational training programs. Given that non-participation is an available option, i.e. to continue receiving social security benefits or to end up with a disability pension, public policy in this sector is confronted head on with private incentives.

Several individual characteristics appear to have an impact on the choice of whether to participate in active rehabilitation or not, and in case of participation, in which type of program to participate. We find that the non-participants differ from participants and that clients in educational programs differ from participants in work related programs.

Participants in educational training have comparatively low disutility of training as we define it. We also find that the monetary opportunity costs of training is comparatively low for this group of clients making signalling an even more attractive option. A narrow interpretation of signalling theory is that we should expect to find that clients with relatively high educational levels dominate among participants in educational training and that active clients separate themselves into the two different programs depending on factors such as former educational level and age. Another interpretation is that while participation in educational training can be interpreted as a signalling decision, participation in work related training is the same as screening.

Participants in work related training do differ in terms of relevant background characteristics and although signalling theory does not rule out pooling of types, clients taking active part in training seem to take separate actions. Thus, participating in work related training could be interpreted as an investment decision rather than a signalling decision since the client reveals private information through his/her conduct at work.

Non-participation could be termed a signal, perhaps the strongest negative signal of ability seen from prospective employers' point of view.

3.3 *"Procurement Auctions with Entry of Bidders"*

The third essay *"Procurement Auctions with Entry of Bidders"* is a theoretical one and co-authored with Steinar Vagstad. The paper considers procurement of fixed quantities of well-defined goods or services. In such situations, most procurers seem to agree on using some variant of what can be called a 'plain' auction: Invite potential suppliers to submit bids, and choose the one with the lowest bid. This practice corresponds to what theory prescribed before 1980, but does not meet the standards set by more recent theory.

What does recent theory prescribe? The central planner is faced with both moral hazard and adverse selection having to decide how to organize a procurement auction. The starting point of the essay is that a procurer can auction a carefully designed incentive contract in such a way that the firm self-selects and elicits optimal effort, i.e. the firm chooses the contract that is designed for its type. In the essay the necessity of auctioning incentive contracts is challenged.

An incentive contract is basically a rent-extracting mechanism. Reducing the expected rent to potential bidders reduces the incentives to enter the auction in the first place. The reduction in the number of participants may possibly result in higher expected project costs. The analysis here combines two related fields of literature on procurement and auctions. One is the 'auctioning incentive contracts' literature, focusing on the desirability of the rent extracting mechanism in a setting with a fixed number of potential suppliers. The other field is the theory of auctions with entry in which it is assumed that each potential supplier must sink a relation-specific entry investment before he can participate in the auction.

Equilibrium entry requires each potential firm to enter if and only if the expected profit is large enough to cover the entry cost. The problem with a rent-extracting mechanism in a model with entry is that reducing the firms' expected rent also reduces their incentives to enter, possibly resulting in higher expected project costs if the number of bidders is reduced. This raises the question of the nature of the optimal mechanism. Is the optimal mechanism a modified incentive contract mechanism or radically different? It is shown that depending on the information structure, fixed-price contracts or plain auctions may be optimal. In particular, if potential suppliers have no private information at the time they take their entry decisions, distortive mechanisms do more harm than good, while plain auction mechanisms perform surprisingly well. In contrast, if potential suppliers know their costs at the time they make their entry decisions, distortive mechanisms are back in business.

In both cases, implementation of the optimal mechanism requires commitment to a mechanism before the entry decisions are made. It is crucial that the mechanism is easy to describe in advance, and that ex post violations are easy to detect. It is argued that neither should be a problem in the two cases discussed in the paper. The most general policy conclusion is that procurers should think twice before trying to improve upon a plain auction, making sure that the benefits of being 'smarter' exceeds the costs, paying particular attention to the effect on entry.

3.4 *"Auctions vs Negotiations – a Study of Price Differentials"*

Although labour costs are the major cost component at hospitals, procurement of equipment, medicine and medical and surgical articles are important components, too. While wages often are fixed, the prices for equipment, medicine and medical and surgical articles are not but at least in part on how procurement is organised. The forth and final essay *"Auctions vs Negotiations – a Study of Price Differentials"* is an econometric study of whether the number of suppliers do depend on how procurement at hospital level is organised and whether number of suppliers matters for the price to be paid. In a sense the essays represent the econometric follow up of *"Procurement*

Auctions with Entry of Bidders". The trading procedures compared are fixed-price auctions and negotiations, though, not auctioning of incentives contracts and pure auctions. Still, the essay represents a step in the direction of testing different market designs. Market design is meant to be understood as the trading rules chosen (here what kind of auction or bargaining game to play), the influence of the rules on the market structure (number of suppliers) and finally the prices to be paid.

Recent contributions in auction and bargaining theory suggest that a procurer should place more faith in the power of competition among alternative suppliers than in his or her own negotiating skill. In this study, based on data collected by the project, the procurement units at publicly owned hospitals or units making procurement on behalf of the hospitals, are compared in terms of prices the hospitals pay for medical and surgical articles. Based on data from approximately 230 contracts between procurers and suppliers of fourteen different medical and surgical articles, we test whether auctions and bargaining result in significantly different prices. We also test whether the market structure, i.e. the number of potential suppliers, depends on the particular trading procedure chosen.

Differences in trading procedures can often be explained by differences in value of the tender. Central governments are obliged to use the international market place and auction mechanisms because orders are generally above the EU thresholds. Local governments with purchases below the relevant EU thresholds can use the national or local market place and choose from a larger set of trading procedures. The essence of the European Union procurement regime is the insistence that major contracts be awarded according to specific procedures designed to ensure openness: active advertising and preference for open tendering; equal treatment: neutral specifications and objective award criteria; and transparency: a clearly defined set of rules, applied in a predictable manner and subject to public inspection.

The EU scheme does not explicitly state economy and efficiency in public procurement as an overall objective. Still, it is clear that the underlying premises are that the enforcement of openness, equal treatment and transparency lead to economic and efficient results compared to other trading procedures not in compliance with the underlying premises.

Bargaining represents an important alternative class of trading procedures to auctions. Given that procurers with orders below the EU threshold can choose from a wider set of procedures, should they opt for negotiations or auctions? And for that matter, should procurers with orders exceeding the thresholds rather split the purchases into smaller orders substituting negotiations for auctions?

The main result of this study is that auctions do not appear to give lower prices compared to negotiations.

REFERENCES

Abbott, T.A. III (ed) (1995), *Health Care Policy and Regulation*, Kluwer Academic Publishers: Boston.

Becker, G. (1983), "A Theory of Competition Among Pressure Groups for Political Influence", *Quarterly Journal of Economics*, 98, 371-400.

Bishop, M., J. Kay and C. Mayer (1995), *The Regulatory Challenge*, Oxford University Press: Oxford.

Clarke, T. and C. Pitelis (1993), *The Political Economy of Privatization*, Routledge: London.

Dranova, D. and M.A. Satterthwaite (2000), "The industrial organization of health care markets", in Culyer, A. and J. Newhouse (eds.): *Handbook of Health Economics, Vol. 1B*. North-Holland: Amsterdam.

Esping-Andersen, G. (1999), *Social Foundations of Postindustrial Economies*, Oxford University Press: Oxford.

Esping-Andresen, G. and M. Regini (eds) (2000), *Why Deregulate Labour Markets?*, Oxford University Press: Oxford.

Gibbons, R. (1997), "Incentives and careers in organizations", in Kreps, D.M. and K.F. Wallis (eds.): *Advances in economics and econometrics: theory and applications, Seventh World Congress, Vol II*. Cambridge University Press: Cambridge.

Giddens, A. (1998), *The Third Way - The Renewal of Social Democracy*, Polity Press: Cambridge.

Ham, C. (2000), "Series Editor's Introduction", in Smith, P.C. (ed) (2000), *Reforming Markets in Health Care, An Economic Perspective*, Open University Press: Buckingham.

Harris, M. and A. Raviv (1992), "Financial contracting theory", in Laffont, J.-J. (ed.): *Advances in Economic Theory, Sixth World Congress, Vol. II*. Cambridge University Press: Cambridge.

Hart, O. and B. Holmstrom (1987), "The theory of contracts", in Bewley, T.F. (ed.): *Advances in Economic Theory, Fifth World Congress*. Cambridge University Press: Cambridge.

Hillier, B. (1997), *The Economics of Asymmetric Information,* MacMillan Press Ltd.: London.

Laffont, J.-J. (1994), "The New Economics of Regulation Ten Years After", *Econometrica*, 62 (No. 3), 507-537.

Laffont, J.-J. (2000), *Incentives and Political Economy*, Oxford University Press: Oxford.

Laffont, J.-J. and J. Tirole (1993), *A Theory of Incentives in Procurement and Regulation*, The MIT Press: London.

Milgrom, P. and J. Roberts (1992), *Economics, Organization and Management*, Prentice-Hall International: New Jersey.

OECD (1992), The Reform of Health Care – A Comparative Analysis of Seven OECD Countries, Health Policy Studies No. 2, Paris.

OECD (1993a), OECD Health Systems – Facts and Trends 1960-1991, Health Policy Studies No. 3, Vol. I, Paris.

OECD (1993b), OECD Health Systems – The Socio-economic Environment Statistical References, Health Policy Studies No. 3, Vol. II, Paris.

Rasmusen, E. (1994), *Games and Information, Second Edition*, Blackwell: Cambridge.

Shleifer, A. (1985), "A Theory of Yardsstick Competition", *RAND Journal of Economics,* 16, 319-327.

Smith, P.C. (ed) (2000), *Reforming Markets in Health Care, An Economic Perspective*, Open University Press: Buckingham.

Stigler, G. (1971), "The Economic Theory of Regulation", *Bell Journal of Economics*, 2, 3-21.

Tirole, J. (1986), "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations", *Journal of Law, Economics and Organizations*, 2, 181-214.

Tirole, J. (1992), "Collusion and the theory of organizations", in Laffont, J.-J. (ed): *Advances in economic theory, Sixth World Congress, Vol. II.* Cambridge University Press: Cambridge.

Chapter 2

Prospective Funding of General Hospitals in Norway – Incentives for Higher Production?

Prospective Funding of General Hospitals in Norway – Incentives for Higher Production?

by
Egil Kjerstad*

ABSTRACT

In Norway, a new system of Activity Based Financing (ABF) for general hospitals was introduced on a comprehensive basis in July 1997. The main purpose of the reform was to increase activity so that more patients could receive treatment more quickly without reducing the quality of care. In this paper we analyse whether the reform has had any significant effect using two different performance indicators: number of patients treated and production of DRG (Diagnosis Related Group) points. We divide the hospitals into two groups: hospitals owned by counties that have adopted the ABF system, and hospitals owned by counties using other funding systems. The first group then becomes the experiment group, while the second serves as a comparison group. It is argued that fixed-effect models are suitable specifications for this evaluation study, handling selection bias and the influence of unobservable explanatory variables in a consistent manner. We find that the reform has had a significant effect on the number of patients treated and DRG points produced but the results are sensitive as to how the experiment and the comparison group are determined.

Keywords: prospective payment, incentives, fixed-effect models

---

*Foundation for Research in Economics and Business Administration (SNF) and Programme for Health Economics in Bergen (HEB). Postal address: Breiviksv. 40, N-5045 Bergen, Norway. E-mail: egil.kjerstad@snf.no.
Many thanks to Jan Erik Askildsen, Kjell Vaage and Andrew Street for comments on earlier drafts and to Arild Aakvik and Andrew Jones for stimulating discussions of the econometric approach taken in this study. Thanks to participants at the HEB workshop in Bergen, May 2000 and in particular to Wim Groot for valuable comments on the workshop draft. Thanks also to participants at the York Seminars in Health Econometrics. NPR and NIS in Trondheim have provided the data. This research is supported by the Research Council of Norway.

1. INTRODUCTION

For the last ten to fifteen years there has been a steady international move towards changing the way health care is delivered, financed and regulated.[1] Ham (2000) makes an interesting observation commenting on the development. Countries which have traditionally relied on a market in health care are making greater use of regulation and planning. In contrast, those countries which traditionally have relied on regulation and planning are moving towards a more competitive approach. In no country is there complete satisfaction with existing methods of finance and delivery, and there is a search for new policy instruments. Nevertheless, as pointed out by Smith et al (1997), despite the variety of approaches, many features are basically the same: a flow of finances from the population, via a variety of agencies, to the provider of health care. Each transfer of funds can be discussed in terms of a principal-agent relationship: a principal is relying on an agent to perform necessary tasks so that some desired aspects of health care delivery can be secured. Smith et al (1997) conclude that three factors are crucial in order to secure adequate control over health care financing. Firstly, there must be control of patient entry into hospitals. Secondly, there must be a mechanism for remunerating hospitals for treating additional patients. Thirdly, there must be control of physicians by hospital management.

It is the second aspect, transfer of funds to hospitals, which is discussed here. One policy instrument available is the Prospective Payment Systems (PPS) based on Diagnosis Related Groups (DRGs), a financing system aimed first and foremost at general hospitals and first introduced in the USA in 1983. The main aim of any prospective payment system (PPS) is to provide stronger incentives for hospitals to avoid costly technologies that offer little expected benefit to the patients. Bradford and Craycraft (1996), drawing on Pope's (1989) theoretical model, examine the effects of PPS on the behaviour of hospitals with respect to their capital allocations and the efficiency with which they produce in-patient care. They find that PPS hospitals supply

[1] See Smith (ed) (2000) for a comprehensive discussion of the introduction of 'internal markets' in the National Health Service (NHS) in UK. Abbott (ed) (1995) gives insight into fairly recent changes of health care policy and regulations in USA. For a somewhat outdated but still illustrative comparative study of health reforms and health care systems in OECD countries, see OECD (1992) and OECD (1993a, b).

greater levels of capital and produce in-patient care using technically less efficient methods. McClellan (1997) emphasises that most prospective hospital reimbursement systems still have important and under-appreciated features that provide considerable retrospective cost sharing, based on how intensive a patient's treatment turns out to be. McClellan discusses in particular the development of PPS in the USA and finds that to an increasing extent, PPS allows hospitals to pass on the costs of treating a patient more intensively to the purchaser of health services. Adding a degree of retrospective reimbursement makes the system less high-powered compared to a fixed-price system.[2]

Gilman (2000) provides an interesting account of hospitals' response to refinements of the PPS in USA. Hospitals are now faced with multiple reimbursement incentives created by the introduction of procedure-based DRGs in addition to the 'first generation' PPS based on patient diagnosis only. Gilman finds that the incentives now created have conflicting effects on hospital resource use. Firstly, he finds no evidence of a moral hazard effect caused by the introduction of high priced treatment-based DRGs, i.e. changes in level of services per discharge for a given severity of admissions. Secondly, the study finds evidence of quality competition in DRGs. Thirdly, Gilman finds evidence in support of the selection effect documented by Ellis and McGuire (1996), i.e. altering the average severity of admissions independent from any changes in treatment policy.

Newhouse (1996) reviews theoretical analyses of PPS based systems and shows that it is optimal to base reimbursement partially on actual costs of treatment. In other words, fixed-price contracts are not optimal. The optimality of cost sharing, i.e. the extent to which a producer shares the reported cost of production with the purchaser, reflects the tension between incentives to provide adequately intensive treatment and incentives to

---

[2] Assume that a PPS consists of a set of individual patient fixed-price contracts. Let p_{ij} denote the DRG price associated with the treatment of patient i with diagnosis j. Of course, the price may be different from the hospital's own cost in treating the patient, c_{ij}. The hospital's net monetary benefit per patient is given by $\pi = p_{ij} - c_{ij}$. If $p_{ij} > c_{ij}$, the hospital runs a surplus on that patient but incurs a loss if $p_{ij} < c_{ij}$. Thus, the PPS represents a high-powered incentive scheme and may change the hospital's incentives considerably compared to a cost-plus contract. For instance, a hospital that puts no weight on patients' benefit of service would have incentives to lower the level of service to its minimum since it is the residual claimant of the cost savings.

minimise costs in health-care production. Ellis (1998) predicts that provider competition under PPS will create incentives for hospitals to select patients ('cream skimming').

In Norway, a country with a long tradition of regulating the health care sector relying basically on central planning, a DRG based financing system (Activity Based Financing, ABF for short) was introduced on a comprehensive basis in July 1997.[3] The Government had used a variant of ABF as early as 1991 but only on a probationary basis and with the participation of only four hospitals. Here, we do not dwell with the experiences made during the project that lasted from 1991 to 1993 but the funding of the participating hospitals was based on diagnosis related groups (DRG) and fixed DRG prices in combination with block grants.[4] During the probation period the relative shares of DRG based funding and block grants changed. In 1993 forty percent was activity-based and sixty percent was made up by a grant. The present ABF model started out with the same shares but has been changed twice. From January 1999 the shares are 50/50. The shares are 40/60 in the period studies here. The main purpose of the reform was to increase activity levels so that more patients could receive treatment more quickly without hospitals compromising on quality.

The introduction of a 'per case' based funding model represents a novelty in Norway. This reform, alongside patients' 'free' choice of hospitals (certain restrictions are imposed), waiting list guarantees and a steady increase in the use of competitive tendering in public procurement of services like nursing and home care for the elderly, are examples of a trend towards experimenting with market-like governance structures.

Our aim is to analyse whether the ABF reform has had any significant effect on the number of patients treated at hospitals throughout the country. We also analyse whether the reform has had any significant effect on the production of DRG points since the level of DRG points, rather than the number of patients, determines the monetary compensation from the state to counties, the owner of the hospitals. We find that the reform has an effect on hospitals' performance.

---

[3] For an overview of the Norwegian health care system, see van der Noord, Hagen and Iversen (1998).

[4] Magnussen and Solstad (1994) conclude, based on findings from the experiment, that the change of funding system has not had any substantial effect on hospital efficiency.

The paper is organized as follows. In the next section a theoretical model explaining possible responses to PPS at hospital level is presented. A presentation and discussion of the development and implementation of PPS in Norway follows in section 3. The econometric model is presented in section 4. In section 5 follows a presentation of the data used in the analysis. Results are presented in section 6, while concluding remarks are gathered in section 7.

2. PPS AND INCENTIVES AT HOSPITAL LEVEL – A MODEL

A prospective payment system and its possible consequences for the resource allocation at hospitals can be illustrated using a simple deterministic and static model. Imagine that PPS is introduced and replaces a funding system based on cost-plus contracts. A hospital admits and diagnose a patient with certainty. The hospital decides with certainty the patient's treatment cost but neither the state nor the owner of the hospital have this information. The treatment cost depends on the hospital's technology and the patient's age, constitution, way of life (smoking, drinking etc.), factors assumed here to be known to the hospital. The state and the owner can only observe aggregate costs (the total treatment cost of all patients) ex post. It is not possible to disentangle the individual treatment cost based on aggregate ex post costs. The state receives a report from the hospital on the individual patient's intensity of treatment but the state cannot verify information concerning individual patients.

Different types of treatments for a given diagnosis and treatment of different diagnoses can be measured in DRG points. Thus it is possible to compare the intensity of treatment across individuals and diagnosis. Let β_{ij}° represent the reported intensity of treatment measured in DRG points for patient i with diagnosis j. The reported value does not have to be identical to the real intensity of treatment β_{ij}, i.e. $\beta_{ij}^{\circ} = \beta_{ij} + \varepsilon_{ij}$, where ε_{ij} represents 'creep' ($\varepsilon_{ij} > 0$). The size of the 'creep' for a given diagnosis must be 'reasonable', i.e. we assume that the government has enough knowledge to rule away extreme values, and that the hospital knows this. Let $\bar{\beta}_j$ denote the maximum 'reasonable' level of treatment. p denotes the fixed price per DRG point of reported treatment β°_{ij}.

Total treatment cost of patient i is given by $C_i = \beta_{ij}\, c_i(\beta_{ij}) + s_{ij}$ where $c_i(\beta_{ij})$ is patient i's cost per DRG point produced for treatment of diagnosis j. The treatment cost increases as the intensity of treatment β_{ij} increases and at an increasing rate ($C_i' > 0$, $C_i'' > 0$). s_{ij} is the level of service patient i with diagnosis j receives. Service is measured as the monetary equivalent of time spent on nursing while hospitalised (i.e. length of stay), time spent on rehabilitation and provision of 'hotel services' like catering and laundry.

A patient's benefit of service measured in monetary terms is given by the function $B_i(s_{ij})$. The benefit function has $B'_i > 0$, $B''_i < 0$ and $B_i(0) = 0$. The hospital knows $B_i(s_{ij})$ and values patient utility as $\alpha B_i(s_{ij})$ where α ($0 \leq \alpha \leq 1$) is the weight the hospital attaches to the patient's welfare while hospitalised.[5]

Let D_j denote the set of possible diagnoses for which a patient may be treated. The hospital has the following maximisation problem:

$$(1) \quad \max_{\beta_{ij},\, s_{ij}} \sum_{i \in D_j} (\beta^{\circ}_{ij}\, p - \beta_{ij}\, c_i(\beta_{ij}) - s_{ij} + \alpha B_i(s_{ij}))$$

s.t.

$$(1a) \quad \sum_{i \in D_j} (\beta^{\circ}_{ij}\, p - \beta_{ij}\, c_i(\beta_{ij}) - s_{ij}) \geq 0$$

$$(1b) \quad \beta^{\circ}_{ij} \leq \overline{\beta}_j$$

We formulate the Lagrangian

$$(2) \quad L = \sum (\beta^{\circ}_{ij}\, p - \beta_{ij}\, c_i(\beta_{ij}) - s_{ij} + \alpha B_i(s_{ij})) - \lambda_1 (\sum (-\beta^{\circ}_{ij}\, p + \beta_{ij}\, c_i(\beta_{ij}) + s_{ij})) - \lambda_2 (\beta^{\circ}_{ij} - \overline{\beta}_j)$$

[5] It is assumed that all patients have equal benefit of treatment (β). The benefit is normalised to zero.

where λ_1 and λ_2 are Lagrange multipliers (constrained to be nonnegative). The optimal level of treatment β_{ij}^* and service s_{ij}^* is subject to the constraints (1a and 1b), the nonnegative constraints on the multipliers, the complementary slackness conditions (not explicitly given here) and the first-order conditions:

(3) $p = c_i(\beta_{ij}) + \beta_{ij}\, c_i'(\beta_{ij}) + \lambda_2/(1 + \lambda_1)$

(4) $\alpha B'_i(s_{ij}) = 1 + \lambda_1$

From (4) we see that the patient is offered a first-best level of service if (i) the hospital is altruistic ($\alpha = 1$) and (ii) a zero profit constraint is not binding ($\lambda_1 = 0$). The less weight the hospital puts on patient welfare - α decreases - the less service is given to the patient. Likewise, if constraint (1a) is binding, i.e. $\lambda_1 > 0$. In this case, the hospital does not earn a monetary surplus when the service level is first-best.

The optimal treatment intensity, β_{ij}^*(measured in DRG points), of a patient depends on several factors. If the optimal level of service s_{ij}^* is such that it is possible to report β_{ij}^* $< \overline{\beta}_j$ and still earn a positive mark-up on that patient, the elasticity of the individual patient's cost function is decisive:

(3') $(p - c_i(\beta_{ij}))/\, c_i(\beta_{ij}) = (\beta_{ij}\, c_i'(\beta_{ij}))/\, c_i(\beta_{ij})$

We have the following propositions:

<u>Proposition 1</u>. The mark-up, i.e. the difference in percent between the fixed per DRG point remuneration p and the treatment cost per DRG point $c_i(\beta_{ij})$, is higher for a cost sensitive patient compared to a relative cost insensitive patient. The more cost sensitive patient receives a smaller β_{ij}^* compared to a less cost sensitive patient with the same diagnosis and the same benefit of service function ($\alpha B_i(s_{ij}) = \alpha B_k(s_{kj})$).

If two patients have different benefit of service functions ($\alpha B_i(s_{ij}) \neq \alpha B_k(s_{kj})$), but identical cost per DRG point ($c_i(\beta_{ij})) = c_k(\beta_{kj})$), the patients will receive different levels of service but equal level of treatment. Moreover, it follows that:

Proposition 2. The hospital has an incentive to 'creep' on every patient. The size of the 'creep' is determined by the difference between β^*_{ij} *and* $\overline{\beta}_j$.

The size of the 'creep', ε_{ij}, is determined by

$$(5) \qquad \beta^{\circ}_{ij} p - \beta^*_{ij} c_i(\beta^*_{ij}) - s^*_{ij} \geq 0 \quad \Rightarrow \quad (\beta^*_{ij} + \varepsilon_{ij}) p \geq \beta^*_{ij} c_i(\beta^*_{ij}) + s^*_{ij} \quad \Rightarrow$$
$$\varepsilon_{ij} p \geq \beta^*_{ij} c_i(\beta^*_{ij}) + s^*_{ij} - \beta^*_{ij} p$$

Proposition 3. If the necessary 'creep' brings the reported treatment intensity (β°_{ij}) *to a level that is in excess of that which is reasonably expected by the Government* ($\overline{\beta}_j$) *for a given diagnosis, the patient is not 'profitable' for the hospital and will not receive treatment.*

In cases where the patient is a marginal one, i.e. in cases where the hospital does not receive a monetary surplus on the patient given the optimal level of treatment β^*_{ij} and service s^*_{ij}, 'creeping' is in the interest of the marginal patient. Without it the patient will not be treated and like other 'unprofitable' patients he or she will be directed to another hospital[6] or end up on the hospital's waiting list. It follows that it is not evident that a hospital will have incentives to treat as many patients as possible. Remuneration is tied to the patients' weight or intensity of treatment measured in DRG points, not to number of patients per se. We will return to this point in the econometric study.

Based on this discussion, we can identify five different categories of possible effects:

- Average hospital length-of-stay and number of unplanned readmissions
  Hospitals can cut costs and earn higher overhead per patient by reducing the length of stay. In turn, this can increase number of unplanned readmissions if patients discharged prematurely fail to recover and require further hospitals treatment.

[6] If there exists a well defined market for guest patients in the sense that hospitals know each others' cost structures, such a 'trade' can be effective and even welfare enhancing. Whether the 'trade' is welfare enhancing depends on the weight put on travel time and distance between a patient's home and the hospital.

- Nursing and access

  Hospitals can cut costs by reducing intensity of nursing and by refusing[7] to treat patients who are expected to be relatively costly ('cream skimming') or by closing treatment units, i.e. reducing capacity.[8]
- Severity of the case mix

  Hospitals can increase income by 'DRG-creep', i.e. increasing the weight of the average patient.
- Financial conditions of hospitals

  Hospitals may experience higher degrees of financial stress, particularly those hospitals with higher costs than average.
- Effects on cost.

  Hospitals respond to ABF by increasing capacity utilisation thereby bringing average cost down.

Our theoretical results are in line with what others have found. For instance, Hodgkin and McGuire (1994) show that PPS may affect the hospitals' choice of intensity of care, which affects the demand for admissions.[9] Ellis and McGuire (1996) argue that hospitals, in response to a change in reimbursement incentives, may change the intensity of services provided to a given set of patients, change the type or severity of patients or change their market share. They find that each of the effects, moral hazard effect, selection effect and practice-style effect respectively, can influence average resource use in a population.

To the extent that a hospital is able to cover its costs through 'creep' or by selecting particular patients, the number of treated patient may increase in our model. That result is implicit in the model though, assuming that a hospital is free to choose among patients in a waiting list extending capacity. If waiting list patients are cost sensitive

[7] The refusal may not be explicit but rather done by postponing treatment (waiting list).
[8] The last point is not evident but if the hospital's costs for a particular treatment are so high that even a reasonably sized "creep" cannot make the hospital break-even, closing of units may be an option.
[9] For a review of the effects of Medicare's prospective payment system, see Folland, Goodman and Stano (1997).

(relatively high marginal cost of treatment) fewer patients may be treated compared to a situation with traditional cost-plus funding.

Going beyond our simple static model, hospitals will have incentives to utilise new and more cost effective treatment. In our model this means changing the cost functions, i.e. lowering marginal treatment cost across all diagnoses or increase specialisation by focusing on a subset of diagnoses. If the Government under PPS also introduces binding waiting list guarantees[10] for patients, a hospital may end up in debt depending on the composition of the waiting list patients. Furthermore, hospitals with acute wards cannot (at least in principle) refuse to treat patients. If the average acute patient demands higher intensity of treatment and care compared to the average elective patient, hospitals with a higher proportion of acute patients all things equal will have a harder time running a surplus. A high proportion of acute patients may also halt or interfere counterproductively the flow of elective patients through the hospital, causing fewer patients to be treated compared to a situation with less acute treatment.

3. INSTITUTIONAL CONDITIONS AND POLICY CONTEXT

The so-called activity based funding system (ABF) of general hospitals in Norway constitutes a rather complex governance structure and a discussion of the specific institutional conditions is necessary to motivate our econometric models. Formally the ABF system is a contract for hospital services between the central Government and the counties. As such, the system is basically guiding the transfer of funds from central to local government. The hospitals receive their remuneration from the owner, the counties, paying with ABF funds and other non-earmarked funds received from the state or from other sources, such as income taxes. The Government's intention is that counties should implement an ABF model in their remuneration of individual hospitals, something that has been done only to a limited degree. Various funding models are pursued across counties. Up until the end of 1998, several hospitals were still financed

[10] Treatment in less than six months as the case is in Norway for a subset of diagnoses.

by block grants and these grants were not necessarily activity based in terms of DRG points produced.

In broad terms, the ABF model works as follows. The Government agrees with each county on how many DRG points (*TDRG*) that will be delivered the forthcoming year. The activity level may build on last year's achievements adjusted for changes in human resource capacity, application of new technology, etc. The agreed upon number of DRG points are transformed into an ABF budget. Each DRG point is priced at NOK 28.289 (the average treatment cost across diagnoses, 1999). As from 1999 the state covers 50% of the value of the budget. Thus, using 1999 as an example, the budget is given by $TDRG \cdot 0{,}5 \cdot 28.289$. Another way to see this is that the prices set by the state for specific DRGs are reduced by 50% when the counties get paid from Government for the health services provided. The counties will have to cover the rest of the expenses themselves (or on the level of treatment of an individual patient, cover the deficit between the DRG price and a hospital's cost of treatment). Thus, health care has to compete for financial resources with other services financed from the counties' budget like high school education, cultural activities and communication. The counties should therefore be concerned with the cost effectiveness in the health sector. The lower the cost of treatment, the lower is the monetary equivalent of the fraction of the DRG price to be covered by the county in order to avoid deficit in the sector. The Norwegian ABF system can in principle be termed a Prospective Payment System since it builds on predetermined cost levels for the different diagnoses and not on reported costs from the individual hospital.

The ABF funds are transferred to the counties in ten instalments during the budget year. At the end of the year the delivered points are counted and the account between state and counties are settled. If a county has delivered more services (produced more DRG points) than planned, it receives additional funds from the state. However the Government has put an upper limit on the growth of the DRG index from one year to another, thus 'punishing' counties that exceed that limit (or encouraging counties to stay within that limit).[11] If the activity level turns out to be less than expected, the counties will owe money to the central Government.

Some counties have implemented a budget process akin to the process between the state and the counties described above. Others have chosen a more traditional budgetary process in which the ABF funds are not directly visible in the hospital budgets although they may agree on activity level measured in DRG points. A few counties measure activity by other means, such as treatment of patients using categories other than DRGs. The incentives to increase activity and contain costs will also be affected by the internal budget process, i.e. between hospital management and the clinics at the hospitals. Very few, if any, Norwegian hospitals had implemented a comprehensive activity based financing of clinics by the end of 1998. Some counties do punish hospitals if the activity level at the end of the year is lower than what has been laid out in the budget. Punishment is imposed by reducing next year's budget by the value of the under production measured at DRG prices.[12]

The Norwegian ABF system has elements that can be termed retrospective, as discussed by McClellan (1997). For instance, ABF gives block grants based on hospital status (central, regional, local) and status has bearing on the type of health services the hospital can offer. Furthermore, the system gives outlier payments, i.e. supplemental payments for unusually expensive or lengthy admissions. Still, the ABF system bases these block grants on cost estimates determined in advance, not on reported costs from the individual hospital.

Another important question is whether the county is paying hospitals the 50% fraction, as would be necessary to give hospitals a full DRG price. We assumed that this is the case in our deterministic model in section 2. If a county does not pay their share, obviously the hospitals will have to be even more selective in terms of who they treat to avoid a deficit. Likewise, a surplus can be taken away or partly taken away in next

[11] The counties will have to pay the gross marginal cost of the DRG points in excess of the limit. Thus, if the production of these DRG points (patients) involves staff overtime the last points earned will typically be very costly for the hospital.

[12] The activity level is normally measured in DRG points. The NOK value of one DRG point is given by the Government, making the transformation to a NOK budget easy.

year's budget process thereby effectively diluting the incentives to contain costs. Last, but not least, hospitals as an organisation may not even act in terms of what one should expect from a residual claimant of cost saving. On the contrary, hospitals may find it more favourable to run a deficit, assuming that the deficit will be covered by the state and/or the counties. In this way the organisation can avoid costly reorganisation.[13] Another option, and a traditional one, is to avoid deficit, or reduce it, by closing units, i.e. reducing capacity. In sum, the possible effects of a fixed-price ABF system outlined above may be weaker in reality than in our model discussed in section 2, except for the point concerning closing down units. Assuming that the hospital is both cost-conscious and focused on increasing income, closure of units is a last resort in our model.

Other aspects of the ABF system may give countervailing incentives, too. DRG points can be interpreted as number of in-patient stays at a hospital adjusted for the case-mix of patients. The DRG points are calculated as follows: $\sum_j (Weight_j \cdot Stays_j)$ where j denotes (DRG) type of diagnosis. The weights vary across diagnosis and may be changed by the government on a yearly basis.[14]

Number of stays is counted at clinic level and with an important feature: Imagine that a patient is in need of treatment at different clinics, say both surgical and medical treatment. The treatment with the highest DRG weight is counted, not the composite stay. Regardless of the number of clinics at which the patient is treated, the stay is counted as one and the DRG points earned are given by the highest DRG weight.

Generally, DRG points may increase for a variety of reasons:

1. An increase in number of stays across diagnosis, which is the basic pronounced aim of the reform.
2. An increase in some DRGs higher than the reduction in other DRGs.
3. Changes in the case-mix, i.e. a change towards DRGs with higher weights, without an increase in the number of treated patients.

[13] Costs in terms of the administration's disutility of effort, costs in terms of disutility of effort of the medical management when hardening the budget restriction at clinic level etc.

[14] In 1999 several changes were made. Our analysis will be based on data for the period 1995-1998. The details for 1999 are therefore not dealt with here.

4. Changes in the case-mix and a reduction in number of patients.

Cases 3 and 4 can on the other hand be a result of

i. More careful coding of diagnostic information compared to earlier (note that the DRG system was in use before the ABF was introduced in 1997).
ii. Categorising patients in a more severe DRG than necessary (creeping).
iii. Choosing procedures earning more points (creeping).

Thus, an increase in DRG points produced is not necessarily the same as an increase in the number of treated patients, as we will return to in the following sections.

The government has realised that it is difficult to separate the different causes behind an increase in DRG points and has put a limit on the 'creep', regardless of reason, from one year to another.[15] 'Creep' is measured using the DRG index. The index is given by $\sum_j DRG/\#stays$. The index gives the average severity or DRG weight of patients treated that year. In effect, the government does not pay for activities that bring the index above an increment of 1% from one year to the next. As a consequence, some counties punish the individual hospitals by reducing next year's budget if they are too productive in the sense that the index increases by more than 1%.

All things equal, the index constraint imposed by the government creates incentives to change the case-mix towards lower weight DRGs. A hospital can increase the number of DRG points, and thereby income, by increasing the number of stays with relatively low DRG weights. The government has realised this and the 'creep' index is calculated by subtracting in the numerator and the denominator selected commonly used DRGs with small weights.

Finally, the ABF system may not give incentives to increase the number of patient treated, as discussed above, but may still create incentives to produce more DRG points

[15] This is probably partly due to fear of 'creeping' but also an aim to keep a lid on the aggregate spending in the sector.

since monetary remuneration is tied to DRG points. Still, even though ABF hospitals have to agree to a funding system that is activity based, the funding system does not provide direct incentives for the individual doctor, nurse or any other staff member (except perhaps for the Director of the hospital) to change their effort level.

4. ECONOMETRIC METHOD

In this study we analyse the effect of introducing ABF on the number of treated patients and the number of DRG points produced. Both performance indicators are derived from the model in section 2 and the discussion in the preceding section. An additional important motivation for using number of patients treated is the fact that the Government highlighted increasing throughput as the main goal of the reform. Use of DRG points as a performance indicator is motivated through the discussion in section 3.

Although the ABF system was introduced on a comprehensive basis for allocating state funds to counties in 1997, not all counties have introduced or mimicked the ABF system locally for paying hospitals. Based on interviews with administrative staff at county and hospital level, we divide the hospitals into two groups: hospitals owned by counties that have mimicked the ABF system, and hospitals owned by counties using other funding systems. The former type, which are in majority, then becomes the experiment group, while the latter type serves as a comparison group. In our main analysis we draw the distinction between the two groups of counties and hospitals depending on whether the county and the hospitals stick to the pre-reform funding principle of block grants. Thus, counties that in the period we study have mainly applied block grants not conditioned on realised DRG point at hospital level encompass the comparison group.

We discuss the sensitivity of the results by raising the threshold of entering the experiment group. In the second set of models the comparison group consists of counties using block grants and block grants based on DRG points produced, while the experiment group consists of counties with block grants and 'ratchet effect', and ABF funding of hospitals. Hospitals in the experiment group, although not funded in exactly the same manner due to variations across counties, are basically faced with 'contracts' (budgets) stating the number of DRG points to be produced and they receive remuneration on DRG based prices decided by the state.[16]

Using a comparison group is the obvious way to find the effect of the ABF system on the performance indicators focused on in this study. Still, a possible problem is that counties self-select into the respective groups. In that case the experiment group and the comparison group are no longer randomly drawn. Behaviour after the reform probably reflects both the selection process and the effects of the ABF.

In Figure 1 we observe the hospitals in one period before (*t-1*) and one period after (*t+1*) the reform took place. In evaluating the reform, what is the effect of the change? Figure 1 illustrates one of the challenges when evaluating social reforms (see for instance Heckman and Hotz (1989) and Moffitt (1991) for overviews of the broader issues involved).

Suppose that *t* represents the year in which ABF was introduced. Focusing only on the trends with endpoints (A, B) it may be tempting to view the difference A-B as the 'program effect' caused by ABF. This is not an appropriate way to evaluate the experiment. Using also points (C, D), reflecting the productivity differences which exist independently of the experiment, the correct way of measuring the effect is (A–B) – (C–D). What this formulation says is that the difference A–B overstates the effect of ABF. The challenge is the pre-experiment differences in productivity, which has to be accounted for to avoid over (or under) estimating the effect of the reform. As illustrated in Figure 1, the effect of the funding reform is A – A′.

[16] See Table A1 in the appendix for an overview of counties and funding systems and the division of counties into two groups. Also see the appendix for summary statistics of the two groups of hospitals (Tables A2 – A6).

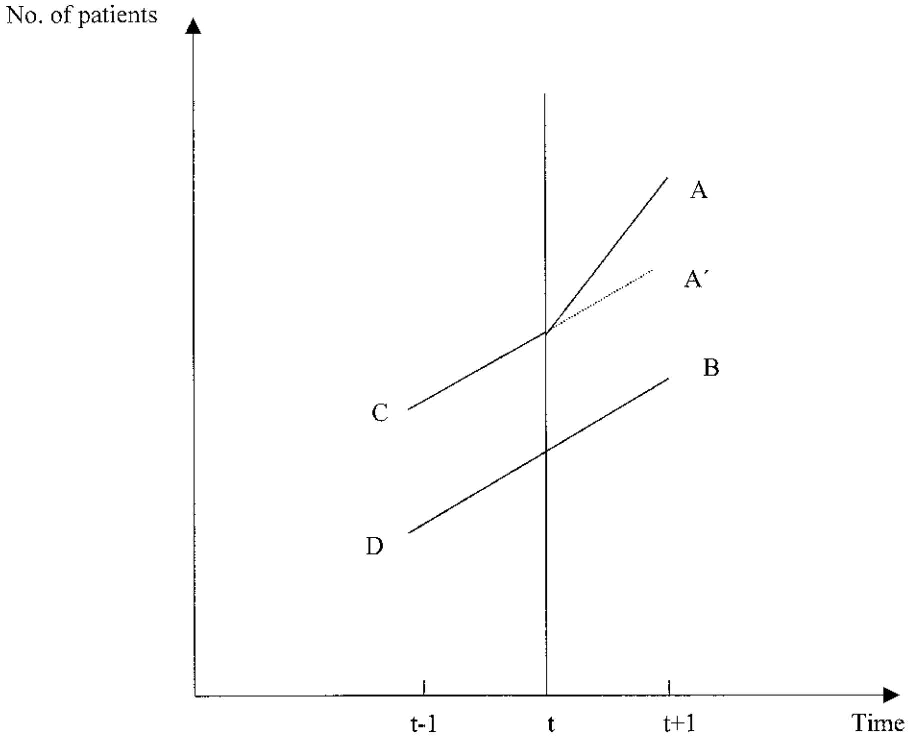

Figure 1. The evaluation problem

We have collected data for 59 hospitals over the time period 1995-1998 allowing us to perform an evaluation in line with Figure 1. We estimate a difference-in-difference model using two time periods only, i.e. a before – after study.

The basic econometric model is as follows:

(1) $$Y_{it} = X_{it}\beta + Z_i\delta + D_{it}\gamma + \alpha_i + \eta_{it}$$

where Y_{it} is number of treated (and unweighted) patients (or the number of DRG points produced). Both dependent variables vary across time t and hospital i. X_{it} is a matrix of explanatory variables that varies across time and hospitals while Z_i is a matrix of explanatory variables that varies across hospitals but for each hospital is constant over time. D_{it} is a dummy variable depending on whether the hospital is categorised as ABF hospital or not in period 2. (In period 1 all hospitals have value 0). α_i represents the time-invariant and unobserved hospital specific effects. η_{it} is the error term for hospital i in period t. β and γ are coefficients to be estimated. However, as will shortly be explained, δ is not.[17]

Changing the time indexing compared to Firgure 1, let $t = 1$ denote the period before the introduction of ABF and $t = 2$ represent the time period after the introduction of ABF. Applying a first difference transformation the model becomes:

(2) $$Y_{i2} - Y_{i1} = (X_{i2} - X_{i1})\beta + (Z_i - Z_i)\delta + (D_{i2} - D_{i1})\gamma + (\alpha_i - \alpha_i) + (\eta_{i2} - \eta_{i1})$$

In the model, α_i accounts for any hospital specific effect that is not included in the regression.[18] This transformation implies that all time-invariant observable and unobservable features (for the researcher) of the hospitals 'cancel out' and the model is reduced to:

(3) $$Y_{i2} - Y_{i1} = (X_{i2} - X_{i1})\beta + D_{i2}\gamma + (\eta_{i2} - \eta_{i1})$$

As discussed earlier, the number of DRG points produced at hospitals for a given time period may be 'contaminated'. For instance, the ABF system may induce hospitals to 'creep' although the underlying true patient-mix has not changed. In other words, the hospitals may change the registrations of patients to earn more DRG points but this change does not represent any true and profound change in the workload of the hospitals. We assume that the patient mix is constant over time and analyse the changes in the number of (unweighted) patients treated at the individual hospital. Note that we

[17] The model is referred to as difference-in-difference model because it identifies differences between experiment and comparison groups and because it is based on differenced variables.
[18] There might be interaction between the ABF dummy and other explanatory variables, i.e. the effect on Y may not be simply additive but multiplicative as well. We have not added interaction terms here but the regression models we run have such terms.

do not assume away any influence between patient mix and the observable explanatory variables X_{it}. The fixed-effect model is robust against cases where Cov (X_{it}, α_i) $\neq 0$, i.e. we do not need to assume independence. On the other hand, using a difference-in-difference model means that we cannot single out the effect of δ alone on the dependant variable. So although we loose some information (i.e time-invariant observable features enter as unidentifiable components in the fixed effect), we gain in that unobserved heterogeneity is dealt with. Note that an important component of this heterogeneity is assumed to be the way in which the hospitals adapt to the DRG based system.

Two possible selection variables are the management style and strategies developed at county and hospital level, features we assume to be time-invariant and unobservable. Management more eager to apply modern management tools, like management by objectives, profit centres etc. may be less reluctant to implement ABF. The fixed-effect model enables us to control for this type of selection by eliminating all unobservable and time-invariant variables that can explain why hospitals are divided into different groups. Using the difference-in-difference model, the dependent and the explanatory variables are stated as changes in the variable values between time periods t=1 and t=2. Thus, the values across groups are comparable along a difference scale. In equation 3, the experiment group or participation group is given the value D_{i2} = 1 and the comparison group D_{i2} = 0. The estimated γ coefficient will tell us whether the reform has had any significant effect on the number of treated patients and on number of DRG points produced.

5. DATA

The data is collected from two main sources: The NPR/NIS registers covering patient and hospital specific data and through a project initiated qualitative study of the extent of ABF implementation locally.

We apply several explanatory variables when estimating the effect of ABF on the two performance indicators number of patients treated and number of DRG points produced at the individual hospital. The explanatory variables are linked to the results of our theoretical model discussed in section 2. Recall the patient individual cost function $C_i = \beta_{ij} c_i(\beta_{ij}) + s_{ij}$ and the benefit of service function $B_i(s_{ij})$.

The change in number of guest patients[19], which are paid for by fixed prices based on a centrally decided tariff outside the ABF system, is used as an indicator of hospitals' motivation to increase the number of DRG patients treated and thereby DRG based revenues. Assuming that the hospitals know their cost structure, an increase in the number of guest patients could lead to lower levels of production if the groups as a whole earn more 'overhead' per guest patient compared to DRG patients or vice versa. This goes for both performance indicators.

Reduction in quality is one possible response to a prospective payment system, as argued in section 2. In the empirical analysis, we assume that length of stay and number of readmissions can be used as quality indicators (s_{ij}), keeping in mind that we analyse aggregate effects, not patient specific effects. Reduction in average length of stay, measured in days, is one way to reduce quality. Reduced length of stay can be a result of technological improvements, though. The coefficient is expected to be negative for both performance indicators. A positive change in the number of readmissions is an indication of reduction in quality. In the case of number of in-patients treated, one could argue that hospitals have incentives to increase production by discharging patients earlier or lowering the level of nursing intensity compared to non-ABF hospitals. Both actions can lead to a higher probability of complications which then 'materialise' in a higher number of readmissions.

However, with respect to DRG points, given that it is possible to "send patients home on Friday afternoon and readmit them on Monday morning" and in that way earn more points, an increase in readmissions is a deliberate action not necessarily at odds with quality. If this is the case, a positive change in readmissions should result in a positive

[19] A guest patient is a patient that is treated at a hospital in a county which he or she is not resident. Counties may make bilateral agreements on how many patients that will be exported/imported during a period of time, often on a yearly basis.

change in DRG points produced.[20] Thus, a positive change in readmissions driven by increasing DRG points may give a positive effect on number of treated patients, too.

When interpreting our results, we will discuss to which degree average length of stay and number of readmissions are useful as indicators of quality.

The expected sign of number of acute patients treated is hard to determine. However, change in the number of acute patients may adversely influence both the number of treated patients and DRG points produced since it is reasonable to assume that acute wards are costly and may interrupt the treatment and admissions of elective or planned cases.

A change in capacity utilisation should also influence positively both number of patients treated and DRG points produced.

We use change in average number of available beds as measure of change in capacity. Average number of available beds is calculated using total number of available beds divided by the number of operative days during a year (excluded closures in connection with public holidays, vacation etc). Higher capacity implies higher number of treated patients and likewise with production of DRG points. Assuming that the case-mix is constant across groups of hospitals, an increase in capacity should result in increased production.

A change in hospitals' net expenses, the difference between gross expenses and gross income, is here interpreted as a financial indicator of intensity of treatment. Our interpretation is that the greater the 'deficit' the greater the intensity of treatment all other things equal. In other words, running a large deficit means that more resources are spent per patient compared to a situation where the deficit is smaller.

[20] Carey and Burgess jr (1999) explore the relationship between cost and quality of hospital care. High likelihood of existence of measurement error in quality in the cross sectional data leads to application of novel instrumental variable techniques. They find that mortality and readmission indices are adjusted inadequately for illness severity.

Furthermore, in section 2 it was also argued that hospitals could respond to a Prospective Payment System (PPS) by differentiating the patient individual intensity of treatment (β_{ij}) and nursing (s_{ij}) depending on the patient's type. In the empirical analysis we are not able to capture such effects directly. We suggest using change in number of doctors per bed (number of beds measures the hospital's capacity) as indicator of intensity of treatment and change in nurses per bed as indicator of intensity of nursing An increase in number of doctors per bed should give a positive effect on number of patients and DRG points produced. The same pattern should be expected concerning changes in number of nurses per bed. More human resources should lead to a higher number of treated patients and possibly also a higher number of DRG points produced.

Last, we use a dummy for ABF hospitals as discussed earlier and interaction terms between the ABF dummy and the explanatory variables.

It can be argued that the dependent variables and the explanatory variables are jointly determined. We need to solve the inconsistency problem caused by endogenous variables. A standard approach is to lag the explanatory variables.[21] Then it can be argued that the explanatory variables (X_{it}) in the model are independent of all subsequent structural disturbances. Thus, the variables are predetermined and can be treated, at least asymptotically, as if they were exogenous in the sense that consistent estimates can be obtained when they appear as regressors.[22]

Using model specifications where both the dependent and explanatory variables are measured as differences in levels, the Cook-Weisberg (C-W)) test[23] shows that heteroscedasticity is present in the specified regression models. The log-log transformations turn out to mitigate this problem, i.e. the models pass the C-W test after transformation, and performs well when testing for omitted variables (RESET test). Furthermore, the size of the hospitals in the sample ranges from approximately 1000

[21] We have also employed current values of the explanatory variables. These model specifications perform poorer than the lagged models regarding tests for omitted variables (RESET test) and heteroscedasticity.
[22] See discussion in Greene (2000), chap. 16.
[23] See STATA reference manual release 6 Volume 3.

treated patients per year to approximately 48 000 patients per year (see Table A2 in the appendix). The log-log transformations allow us to compare growth rates rather than changes in levels. Finally, using log-log specifications on both sets of models allow all results to be discussed in terms of elasticities.

Our research strategy is to apply the full set of explanatory or control variables and their respective interaction terms. Although the ABF dummy captures whether there is a significant difference between the two groups of hospitals (different intercepts), the interaction terms capture whether the reform influences the performance indicators through the other variables (different slopes).

6. RESULTS

In Model 1 and Model 2 in Table 1 number of patients and number of DRG points are the dependent variables, respectively. The experiment group consists of hospitals that are funded on a DRG based system. The comparison group consists of hospitals using traditional block grant funding.

The constant terms pick up effect of time on the dependent variable. In Model 1, the constant term is positive and significant. There is an increase in number of patients treated over time, i.e. 1998 compared to 1996.

Number of guest patients (ΔGUESTS) has an insignificant common effect on number of patients treated. Common in the sense that the estimated coefficients of the control variables are the parts that are common for both groups, while the total effect for the ABF group is this part pluss the part stemming from the interaction variables.[24] Also change in average length of stay (ΔDAYS) has insignificant effects on number of

[24] Note that we comment on the explanatory variables in the ordering of Table 1. The coefficients of the control variables indicate the common effect on the dependent variable. The interaction terms have to be considered too, when commenting on the effect at ABF-hospitals. For simplicity, we can illustrate the point by writing the estimated equation as $Y = \beta_0 + \beta_1 Var1 + \beta_2 ABFVar1$. The (partial) derivative of the equation is $\partial Y/\partial Var1 = \beta_1 + \beta_2 ABF$. β_1 is the coefficient for the comparison group (ABF=0), while the sum of the coefficients gives the effect at ABF hospitals. We test whether the compound effect is significant by testing $\beta_1 + \beta_2 = 0$ or not.

treated patients. Interpreting average length of stay as a proxy variable of quality, we find that at hospitals belonging to the comparison group, there appears to be no negative relationship between change in production and change in quality.

A positive change in number of readmissions (ΔREADMISSIONS) appears to have a positive effect on number of treated patient. Interpreted as change in quality, this means that a reduction in quality (positive change in readmissions) has a positive effect on production.

Table 1. OLS regressions
Model 1: Number of patient treated
Model 2: Number of DRG points produced

	Model 1 Patients		Model 2 Points	
	Coeff.	Robust Std. Err.	Coeff.	Robust Std. Err.
CONSTANT	0,099 ***	0,018	0,029	0,034
ΔGUESTS	-0,009	0,027	0,025	0,042
ΔDAYS	-0,226	0,215	0,171	0,227
ΔREADMISSION	0,175 ***	0,050	0,123 *	0,074
ΔEMERGENCY	0,348 ***	0,098	0,385 ***	0,105
ΔUTILISATION	-0,005	0,220	-0,226	0,246
ΔCAPACITY	0,326	0,246	0,138	0,231
ΔNETEXPEN	-0,674 ***	0,121	-0,533 ***	0,173
ΔNURSES	-0,179 *	0,096	0,029	0,124
ΔDOCTORS	-0,076 *	0,045	-0,183 ***	0,044
ABF	-0,083 ***	0,027	-0,068 *	0,040
ABFΔGUESTS	0,113 **	0,052		
ABFΔDAYS				
ABFΔREADM	-0,130 **	0,065		
ABFΔEMERG	-0,448 **	0,181	-0,303 *	0,177
ABFΔUTILIS				
ABFΔCAPAC			-0,496 **	0,242
ABFΔNETEX	0,679 ***	0,156	0,731 ***	0,248
ABFΔNURSES	0,406 ***	0,122		
ABFΔDOCTORS			0,137 *	0,079
R^2	0,51		0,18	
CW[1]	0,67 χ^2=0,18		0,13 χ^2=2,25	
RESET (Ramsey test)	0,28 F(3,49)= 1,31		0,92 F(3,41)= 0,16	

[1] Cook-Weisberg test for heteroscedasticity using fitted values for dependent variable.

*** Significant at 1% level. ** Significant at 5% level. * Significant at 10% level.

Furthermore, at non-ABF hospitals, a one percent change in the ratio between emergency cases to total number of patients (ΔEMERGENCY) treated leads to an inelastic but positive response in terms of number of treated patients.

Neither change in capacity utilization (ΔUTILISATION), nor change in capacity (ΔCAPACITY) have significant effects on activity. Both results are surprising, and contradict our expectations drawn from the model presented in section 2.

Interestingly, a positive change in net expenses (ΔNETEXPEN) leads to a reduction in number of treated patients at non-ABF hospitals. We will comment on this result below.

A positive change in nurses (ΔNURSES) has a negative effect on number of treated patients at non-ABF hospitals. In terms of what economic theory predicts, we should expect to find a positive absolute effect (not necessarily a positive marginal effect). Likewise with a positive change in number of doctors per bed (ΔDOCTORS), but we find that it leads to a (small) negative change across all hospitals.[25]

The ABF dummy is significant and negative. Note, though that the effect of the reform has to be evaluated taking the interaction terms into account as well.[26] We will return to this shortly but let us first comment on the individual explanatory variable's effect at ABF hospitals.

[25] We should keep in mind that outpatient activity is not part of the analysis due to missing data and under-reporting for several of the hospitals in the period 1995-98. Furthermore, in the data available there is no way to decide where different types of man-years have been employed: in-patient care or outpatient care. In 'macro' there has been an increase in outpatient treatment over the years (SAMDATA (1999)). Thus, the overall activity level at hospitals are higher than accounted for in the analysis and may explain the negative effect on in-patient care, as response to an increase in number of doctors per bed.

[26] For illustrative purpose, we write the equation as $Y = \beta_0 + \beta_1 ABF + \beta_2 ABFvar1 + \beta_3 ABFvar2$. The effect of ABF is found using the compound effect of $\beta_1 + \beta_2 var1 + \beta_3 var2$. We test whether the coefficients $\beta_1 + \beta_2 + \beta_3 = 0$ and evaluate the effect of ABF by using average values of the explanatory variables for the experiment group of hospitals.

While a change in number of guest patients has a negative effect on production at non-ABF hospitals, the opposite is the case at ABF hospitals (ABFΔGUESTS). A positive change in guest patients leads to a positive change in production.[27]

The effect of a change in average length of stay (ABFΔDAYS) is non-significant also at ABF hospitals.[28] A change in quality (positive or negative) has no significant effect on production.

Change in readmissions (ABFΔREADM) leads to significant effect at both groups of hospitals. The effect is different between the two groups of hospitals, but in size not in sign. At ABF hospitals the effect of an increase in readmissions, is less positive compared to the situation at non-ABF hospitals. Interpreted as a change in quality, this means that a reduction in quality has a smaller effect at ABF hospitals compared to non-ABF hospitals. Based on the discussion of the model presented in section 2, a quality change could be expected to have a larger effect at ABF hospitals. However, we should keep in mind the discussion in section 5 regarding average length of stay and readmission as quality indicators.

At ABF hospitals, more emergency cases relative to the overall number of patients (ABFΔEMERG) treated leads to a significant reduction in number of treated patients while the opposite is true at non-ABF hospitals.

Note also that higher deficit (ABFΔNETEXP) leads to hardly any change in the number of treated patients at ABF hospitals (but still a significant effect) while we saw that non-ABF hospitals reduced the total number of patients treated. This difference can be attributed to the incentives built into the DRG based funding systems. Whether it is a

[27] With reference to footnote 24, test shows that the compound effect is significant at ten- percent level (F-value of 2.34). Similar tests are performed for the other compound effects (explanatory variable and interaction term), too. All test results are significant on one- percent level of significance unless otherwise stated.

[28] The approach taken is that we start out with a full model specification (all explanatory variables and their interaction terms). Insignificant interaction terms are dropped one by one until only significant interaction terms are present in the model specification, using also the RESET test (Ramsey test) and the Cook-Weisberg test for heteroscedasticity as guidance to which combination of interaction terms to use in the regressions.

wise response to a growing deficit to decrease activity, as we find non-ABF hospitals do, or increase production/keeping status quo, as ABF hospitals do, depends on the cost of treating (additional) patients compared to the income the treatment generates. Assuming that hospitals know their marginal treatment costs across diagnosis, this result is an indication of higher efficiency at ABF hospitals compared to non-ABF hospitals.

Another difference between the two groups of hospitals is connected to the effect of changing the number of nurses per bed (ABFΔNURSES). ABF hospitals appear to significantly increase production when intensity of nursing increases, while non-ABF hospitals decrease production.

What then about the total effect of the reform on number of treated patients? Following the procedure illustrated in footnote 26, we first test whether the coefficients are significantly different from zero and find them to be so (F-value is $F(9,43) = 9,39$). Evaluating the effect of the reform using the average values of the explanatory variables gives an elasticity of 0.025. In other words, ABF hospitals treat more patients compared to non-ABF hospitals, i.e. the reform has a significant positive but inelastic effect on production.

Turning to Model 2 in Table 1, note that the time effect is not significant. Next, neither changes in number of guest patients (ΔGUESTS) nor changes in average length of stay (ΔDAYS) have significant impact on DRG points produced (the same pattern as in Model 1).

As expected, a positive change in number of readmissions (ΔREADMISSION) leads to more DRG points produced and henceforth, more DRG based income. A positive change in the emergency-elective patient mix (ΔEMERGENCY) leads to more DRG points produced, as it does for number of treated patients.

Change in capacity utilisation (ΔUTILISATION) has non-significant effect on DRG points produced. As a matter of fact the effect is non-significant on both performance indicators, which is unexpected. The same is true for change in capacity (ΔCAPACITY). These results are unexpected and counterintuitive. We have no good explanations for these results.

A positive change in net expenses (ΔNETEXPEN), i.e. a greater deficit, gives fewer DRG points produced at non-ABF hospitals. At non-ABF hospitals, financed through block grants, a greater deficit could well be associated with lower production both in terms of patients and in terms of points. As pointed out in section 2, whether this is a sound response, depends on whether the hospitals have lower or higher marginal costs than the DRG prices offered to them. We return to this point when discussing the effect at ABF hospitals.

A positive change in number of doctors per bed (ΔDOCTORS), leads to a reduction in points at non-ABF hospitals, as it does with number of patients treated. As pointed out above, this result is hard to explain. One possible explanation is that to the extent doctors (increasingly) get engaged in administrative work, more doctors per bed does not necessarily lead to more clinical activity at hospitals.[29]

The ABF dummy is significant but again, before we conclude regarding the overall effect on DRG points produced, let us discusses the effect of the explanatory variables at ABF hospitals.

As with number of patient treated, a change in the emergency patients to total number of patients ratio (ABFΔEMERG) has significant effect at ABF hospitals, too. The effect is significant and positive.

The effect of a positive change in capacity (ABFΔCAPAC) appears to be negative at ABF hospitals, a result that is hard to explain. However, the F-test shows that the effect is insignificant.

[29] We cannot verify such a change in the use of doctors or for that matter nurses with the data available, though. In aggregate, there has been an increase by 28% in number of man years (doctors) in the period 1994-1998, while number of beds has decreased by 2% in the same period (SAMDATA (1999)).

More interestingly, the effect of a positive change in net expenditures (ABFΔNETEXP) is a significant increase in number of DRG points at ABF hospitals, while the effect is a reduction at non-ABF hospitals. These results are similar to the ones found in Model 1. One possible explanation is that ABF hospitals are more efficient compared to non-ABF hospitals or assume that their marginal costs are lower than the DRG price.

A change in number of doctors per bed (ABFΔDOCTORS) also has different impact among the two groups of hospitals. The effect is significant and negative but less so at ABF hospitals.

Finally, performing the test and the calculation as described above regarding the effect of the reform on number of treated patients, we find that the overall effect of the reform is significant and positive. ABF hospitals produce more DRG points compared to non-ABF hospitals.

We have performed the same analysis using a different way to divide the hospitals into the comparison group and the experiment group (see Table A1 in the Appendix), basically making it 'more difficult' to be termed an ABF hospitals. The main results are that fewer variables are significant. In particular, the constant terms are no longer significant, neither are the ABF dummies. On both performance indicators, the overall effect of the reform turns out to be insignificant, too.

7. CONCLUDING REMARKS

The main goal of this study has been to evaluate the effect of the ABF reform on the number of treated patients and the number of DRG points produced. Based on the theoretical model developed in the paper, one would expect to find effects both on the number of patients treated and possibly on the number of DRG points produced. Model specifications using interaction terms gave evidence that there are significant differences between the two groups of hospitals in terms of the impact on number of patients treated and DRG points produced. Since the main purpose of the reform was to increase the activity level so that more patients could receive treatment, from a policy perspective the reform may be termed a success given our empirical results. On the other hand, the results are sensitive as to how the experiment group and the comparison group are determined.

In practice, the Norwegian ABF funding system is more complex and inhibits more features than our simple model could capture. We have argued that one of the main shortcomings is the fact that ABF is a 'contract' between the state and the owner of the hospitals, the counties, and not the hospitals themselves, possibly 'diluting' the incentives for higher efficiency. Thus, this feature may be part of the reason for finding only weak effects of the ABF reform on the performance indicators focused here. However, qualifications can be made to the study itself, too.

First, the sample is small. Second, we have divided the hospitals into two groups based on qualitative data collected by interviewing personnel working with the ABF system at county level. The information they gave about the implementation of the system at hospital level may only partly be correct.

Another fact that can bias our results is that we have not been able to incorporate the effect of changes in the in-patient/out-patient mix due to lack of data. Last, also due to lack of available data, we have not been able to incorporate variables that could better measure quality changes than those employed. Rogers et al (1990) find, based on experiences made in the USA, that mortality following hospitalisation has been unaffected by the introduction of PPS but that PPS has increased the likelihood that a patient will be discharged in an unstable condition. We have used average length of stay and readmission as quality variables but application of both variables can be criticized for being only partly useful for the purpose. Changes in treatment, technology and/or the way patient are rehabilitated after surgery can in fact lead to a shorter period of hospitalization. Thus, shorter average length of stay can be a sign of quality improvements for some patients. Still, for other groups of patients a shorter stay is of less quality compared to a longer stay. Using change in average length of stay (across all patients) is of course also a 'rude' measure since it may hide the fact that some patients have been hospitalized for longer periods of time, implying an increase in quality. Readmissions are here lumped together, making no distinction between planned and unplanned readmissions. The latter category would be of most relevance as a quality indicator. It should be of interest to follow up this study taking quality changes more explicitly into consideration, a perspective that should be of considerable political interest, too.

APPENDIX

Table A1. Counties and funding systems. ABF dummy. Period 1997-1998

County	Block grant	Block grant based on DRG points	Block grant based on DRG points + 'ratchet effect'[1]	ABF funding of hospitals	ABF dummy[2]
Finnmark			X		1
Troms	X				0
Nordland			X		1
Nord-Trøndelag				X	1
Sør-Trøndelag			X		1
Møre og Romsdal		X			(0)1
Sogn og Fjordane	X				0
Hordaland			X		1
Rogaland		X			(0)1
Vest-Agder			X		1
Aust-Agder	X				0
Telemark			X		1
Vestfold		X			(0)1
Buskerud	X				0
Oslo				X	1
Akershus			X		1
Østfold				X	1
Oppland				X	1
Hedmark			X		1

[1] Some counties reward and/or punish hospitals in next year's budget depending on whether the hospital has delivered more or less DRG points compared to the level that the parties have agreed upon.
[2] Counties with two entries are borderline cases. In the first set of models, counties using block grants become the comparison group and the rest of the counties become the experiment group. In the second set of models, counties using block grants and block grants based on DRG points become the comparison group. We have run the models with both specifications but report in detail only the results from the first set of models.

Table A2. Summary statistics. All hospitals. Levels

Variable	Obs	Mean	Std.Dev	Min	Max
PATIENTS98	59,00	10998,59	10580,95	1045,00	48053,00
PATIENTS96	59,00	10512,46	10017,97	1122,00	45114,00
POINTS98	59,00	11610,75	11831,67	1233,10	50455,65
POINTS96	59,00	11694,38	11980,48	1301,52	51429,96
GUESTS97	59,00	1321,81	3968,41	37,00	28734,00
GUESTS95	59,00	1196,95	3824,66	40,00	28134,00
DAYS97	59,00	5,93	0,93	4,40	9,47
DAYS95	59,00	6,17	1,07	4,92	11,46
READMISSIONS97	59,00	1373,81	1426,96	65,00	7204,00
READMISSIONS95	59,00	1253,20	1338,17	81,00	7009,00
EMERGENCY97	59,00	7210,29	6823,55	398,00	32261,00
EMERGENCY95	59,00	6753,56	6489,52	491,00	31333,00
UTILISATION97	59,00	82,36	8,96	57,77	105,97
UTILISATION95	59,00	83,20	9,48	55,00	105,00
CAPACITY97	59,00	213,33	220,23	27,25	970,33
CAPACITY95	59,00	210,90	211,88	30,00	922,00
NETEXPEN97	59,00	315931,80	345478,70	28767,27	1398671,00
NETEXPEN95	59,00	266512,20	278092,60	26129,88	1102412,00
DOCTORS97	59,00	0,39	0,12	0,18	0,81
DOCTORS95	59,00	0,36	0,09	0,23	0,68
NURSES97	59,00	1,85	0,31	1,32	2,72
NURSES95	59,00	1,76	0,27	1,20	2,28

Explanation of variables

The variable PATIENTS is number of in-patients treated during a one-year period. POINTS denotes the number of DRG points produced during the same time period. GUESTS: number of guest patients treated. DAYS: average length of stay measured in days. READMISSIONS: number of readmissions during a year. EMERGENCY: Number of acute patients treated. UTILISATION: utilisation of beds measured in percent of total 'day and night' beds available during a year. Capacity utilisation is measured by taking total number of patients' days during a year divided with average number of beds available times the factor 100/365(366). CAPACITY: average number of beds available during a year. DOCTORS: Number of man-years per bed. NURSES: Number of man-years per bed. NETEXPEN: the difference between gross costs and gross income, measured in NOK, fixed prices (1998).

Table A3. **Summary statistics. The Large ABF group/Experiment group**
Dependent variables 1998-1996 differences
Explanatory variables 1997-1995 differences

Variable	Obs	Mean	Std. Dev.	Min	Max
PATIENTS	48,00	551,35	859,27	-369,00	4507,00
DRGPOINTS	48,00	-48,42	638,01	-1636,48	2716,04
GUESTS_1	48,00	133,69	307,69	-63,00	1510,00
DAYS_1	48,00	-0,21	0,42	-1,99	0,69
READM_1	48,00	140,83	205,87	-352,00	899,00
EMERG_1	48,00	505,38	1004,93	-1799,00	5257,00
CAPACITY_1	48,00	3,70	17,81	-15,97	81,33
UTILISAT_1	48,00	-0,18	5,02	-13,61	9,97
NETEXPEN_1	48,00	53648,80	74671,87	-3709,33	300097,90
DOCTORS_1	48,00	0,04	0,06	-0,11	0,25
NURSES_1	48,00	0,09	0,17	-0,27	0,53

Table A4. **Summary statistics. The Small Comparison group**
Dependent variables 1998-1996 differences
Explanatory variables 1997-1995 differences

Variable	Obs	Mean	Std. Dev.	Min	Max
PATIENTS	11,00	201,55	559,01	-975,00	1297,00
DRGPOINTS	11,00	-237,33	523,43	-1611,65	226,70
GUESTS_1	11,00	86,36	114,07	-73,00	272,00
DAYS_1	11,00	-0,38	0,32	-0,91	0,01
READM_1	11,00	32,36	169,66	-371,00	337,00
EMERG_1	11,00	244,45	708,11	-304,00	2230,00
CAPACITY_1	11,00	-3,10	16,74	-45,79	16,57
UTILISAT_1	11,00	-3,72	4,93	-16,31	4,44
NETEXPEN_1	11,00	30964,61	38403,67	1110,09	108755,30
DOCTORS_1	11,00	0,01	0,04	-0,05	0,09
NURSES_1	11,00	0,10	0,16	-0,15	0,35

Table A5. Summary statistics. The Small Experiment group Dependent variables 1998-1996 differences Explanatory variables 1997-1995 differences

Variable	Obs	Mean	Std. Dev.	Min	Max
PATIENTS	38,00	524,47	944,58	-369,00	4507,00
DRGPOINTS	38,00	-93,37	698,55	-1636,48	2716,04
GUESTS_1	38,00	158,05	341,73	-63,00	1510,00
DAYS_1	38,00	-0,24	0,45	-1,99	0,69
READM_1	38,00	154,58	208,54	-172,00	899,00
EMERG_1	38,00	493,39	1105,72	-1799,00	5257,00
CAPACITY_1	38,00	4,41	19,51	-13,62	81,33
UTILISAT_1	38,00	-0,29	5,44	-13,61	9,97
NETEXPEN_1	38,00	59029,31	82062,66	-3709,33	300097,90
DOCTORS_1	38,00	0,05	0,06	-0,04	0,25
NURSES_1	38,00	0,10	0,17	-0,27	0,53

Table A6. Summary statistics. The Large Comparison group Dependent variables 1998-1996 differences Explanatory variables 1997-1995 differences

Variable	Obs	Mean	Std. Dev.	Min	Max
PATIENTS	21,00	416,76	536,31	-975,00	1409,00
DRGPOINTS	21,00	-66,03	454,21	-1611,65	473,96
GUESTS_1	21,00	64,81	89,84	-73,00	272,00
DAYS_1	21,00	-0,26	0,30	-0,91	0,25
READM_1	21,00	59,14	180,69	-371,00	359,00
EMERG_1	21,00	390,38	620,77	-304,00	2230,00
CAPACITY_1	21,00	-1,16	13,46	-45,79	16,57
UTILISAT_1	21,00	-1,83	4,56	-16,31	5,81
NETEXPEN_1	21,00	32030,45	33460,89	1110,09	108755,30
DOCTORS_1	21,00	0,01	0,05	-0,11	0,12
NURSES_1	21,00	0,08	0,15	-0,19	0,35

REFERENCES

Abbott, T.A. III (ed) (1995), *Health Care Policy and Regulation*, Kluwer Academic Publishers: Boston.

Bradford, W.D. and C. Craycraft (1996), "Prospective Payments and Hospital Efficiency", *Review of Industrial Organization*, Vol. 11, 791-809.

Carey, K. and J.F. Burgess jr (1999), "On Measuring the Hospital Cost/Quality Trade-off", *Health Economics*, 8, 509-520.

Ellis, R.P. and T.G. McGuire (1996), "Hospital Response to Prospective Payment: Moral Hazard, Selection and Practice-style Effects", *Journal of Health Economics*, Vol. 15, 257-277.

Ellis, R.P. (1998), "Creaming, Skimping and Dumping: Provider Competition on the Intensive and Extensive Margins", *Journal of Health Economics*, Vol. 17, 537-555.

Folland, S., A.C. Goodman and M. Stano (1997), *The Economics of Health and Health Care*, Second Edition, Prentice-Hall International: London.

Gilman, B.H. (2000), "Hospitals Response to DRG Refinements: The Impact of Multiple Reimbursement Incentives on Inpatient Length of Stay", *Health Economics*, Vol. 9, No. 4, 227-294.

Greene, W.H. (2000), *Econometric analysis*, Fourth Edition, Prentice-Hall International: London.

Ham, C. (2000), "Series Editor's Introduction", in Smith, P.C. (ed) (2000), *Reforming Markets in Health Care, An Economic Perspective*, Open University Press: Buckingham.

Heckman, J.J. and V.J. Hotz (1989), "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training", *Journal of the American Statistical Association*, December, 862-880.

Hodgkin, D. and T.G. McGuire (1994), "Payment Levels and Hospital Response to Prospective Payment", *Journal of Health Economics*, Vol. 13, 1-29.

Magnussen, J. and K. Solstad (1994), "Case-based Hospital Financing: the Case of Norway", *Health Policy*, Vol. 28, 23-36.

McClellan, M. (1997), "Hospital Reimbursement Systems: An Empirical Analysis", *Journal of Economics and Management Strategy*, Vol. 3, No. 1, 91-128.

Moffitt, R. (1991), "Program Evaluation with Nonexperimental Data", *Evaluation Review*, June, 291-314.

Newhouse, J.P. (1996), "Reimbursing Health Plans and Health Providers: Efficiency in Production versus Selection", *Journal of Economic Literature*, Vol. 34, September, 1236-1263.

OECD (1992), The Reform of Health Care – A Comparative Analysis of Seven OECD Countries, Health Policy Studies No. 2, Paris.

OECD (1993a), OECD Health Systems – Facts and Trends 1960-1991, Health Policy Studies No. 3, Vol. I, Paris.

OECD (1993b), OECD Health Systems – The Socio-economic Environment Statistical References, Health Policy Studies No. 3, Vol. II, Paris.

Pope, G.C. (1989), "Hospital Nonprice Competition and Medicare Reimbursement Policy", *Journal of Health Economics*, Vol. 8, 147-172.

Rogers, W.H., D. Draper, K.L. Kahn, E.B. Keller, L.V. Rubinstein, J. Kosecoff and R.H. Brook (1990), "Quality of Care Before and After Implementation of the DRG-based Prospective Payment System – a Summary of Effects", *JAMA*, Vol. 264, No.15, 1989-1994.

SAMDATA (1999), Sykehussektoren 1998 – fra rammefinansiering til ISF, NIS-rapport 3/99, Trondheim (in Norwegian).

Smith, P.C., A. Stepan, V. Valdmanis and P. Verheyen (1997), "Principal/Agent Problems in Health Care Systems: an International Perspective", *Health Policy*, Vol. 41, 37-60.

Smith, P.C. (ed) (2000), *Reforming Markets in Health Care, An Economic Perspective*, Open University Press: Buckingham.

van der Noord, P., T.P. Hagen and T. Iversen (1998), The Norwegian Health Care System, Economics Department Working Papers No. 198, OECD, Paris.

Chapter 3

Skill Formation among Vocational Rehabilitation Clients – Public Policy vs Private Incentives

Skill Formation among Vocational Rehabilitation Clients – Public Policy vs Private Incentives

by

Arild Aakvik[*] and Egil Kjerstad[**]

ABSTRACT

In this paper we analyse individual vocational rehabilitation clients' decisions to enter active training or not. Although the Government pays the direct costs of training, the composition of the total costs of training may be decisive for individual choices. Based on labour market theory, we relate background characteristics of the clients to monetary opportunity costs and non-monetary costs of training, arguing that training choices are a consequence of differences in costs of training. We use a ten percent sample of participants in educational programs, work related training and non-participants who entered the Norwegian vocational rehabilitation sector in the period from 1989 to 1993, a total of 6653 persons. We find that the background characteristics of persons investing in educational training differ along several dimensions compared both to persons attending work related training and to clients not participating in training at all.

Keywords: public policy, private incentives, costs of training, educational training, work related training

---

[*] Programme for Health Economics in Bergen (HEB).

[**] Foundation for Research in Economics and Business Administration (SNF) and Programme for Health Economics in Bergen (HEB).

Many thanks to Jan Erik Askildsen, Kjell Vaage and Andrew Jones for comments on earlier drafts. Thanks also to seminar participants at SNF. The Norwegian Social Data Service (NSD) in Bergen provided the data. All the data are gathered from sources at The Directorate of Labor, The Social Insurance Organization and Statistics Norway. Both researchers thank the Ministry of Labor and the Research Council of Norway for financial support. None of the mentioned institutions are in any circumstances responsible for the analysis or for the conclusions drawn from it.

1. INTRODUCTION

Conditions like ill health and/or mismatch of individuals' skills and the demands in the labour market can for some lead to a permanent withdrawal from the labour market. For instance, people with ill health - making them unable to perform their traditional craft or to perform according to employers' expectations - may end up as long term unemployed and finally, with a disability pension. However, unemployed may still have a chance to find employment by acquiring new skills in accordance with their health condition and the demands in the labour market. Governmental interventions in the areas of post-schooling employment schemes and training programs may be means of achieving this. Such governmental interventions are permanent fixtures of most OECD countries and the interventions normally have two basic goals: to reduce unemployment and poverty by increasing skills of certain groups of the population. Both objectives require that the schemes or programs increase the probability of employment and/or increase the earnings of the participants above what they would otherwise achieve.

An almost generic Governmental view is that the main purpose of offering educational and work related training programs to partly disabled or hard to employ workers is to enhance the participants' human capital and productive skills, increase individual employment prospects and, in turn, reduce transition to disability pension. However, from a client's point of view participation or non-participation in a training program may be the result of a more complex decision making process influenced by intrinsic, and, for the Government (and researcher), unobservable characteristics of the individual client.[1] Although there are observable costs, such as fees, connected to training programs, other cost components, such as disutility of effort, are not observable (directly at least). Thus, public policy in the area of manpower training confronts private incentives partly driven by factors not observable by policy makers, program administrators or researchers.

[1] The caseworker's opinion of a client's need of training may have an impact on the choice made by a client. Furthermore, a client may be rationed in the sense that the most preferred training program is not available. We have neither information allowing us to measure caseworkers' impact on the choices made by clients nor the degree of rationing.

There is a huge literature devoted to estimating the effect of training programs (e.g. Ashenfelter (1978), Bassi (1984) and Heckman, Hotz and Dabos (1987)) and to analyzing the particular econometric issues involved when estimating such effects (e.g. Heckman and Robb (1985), Heckman and Hotz (1989), Moffitt (1991) and Heckman and Smith (1996)). The topic of sample selection and the deleterious effects on the properties of conventional estimators such as least squares has been at the core of much recent work. The basic problem is that selection bias arises when a non-random selection process determines participation in training programs. Researchers do not observe all explanatory variables influencing program participation and outcome.

Selection bias is of course not only an econometric issue. The ambition of Governmental programs is to 'hit the right people', an ambition usually stated quite explicitly. The selection bias may be a 'real life problem' if the Government allocates scarce resources mainly to program participants who without a training program would do as well in the labour market as with such training. Thus, when evaluating the impact of training programs on earnings or on the probability of getting employed one should take into account the fact that the impact might be correlated with variables determining participation in the first place.

The importance of unobservable characteristics of a client becomes particularly evident when noting that skill formation can be acquired in a variety of situations with different levels of training costs for an individual. Generally, participants in training programs incur three types of costs and the composition of cost components can affect the training decisions taken by the individuals. Firstly, participants usually have to pay a tuition fee to participate in a training program. Sometimes the employer or a third party, like local or central Government, pays this direct cost. Secondly, it is often assumed that participation in a training program gives rise to non-monetary costs (NMC) or disutility of training born by the individual. Thirdly, the clients incur monetary opportunity costs (MOC) in terms of lost income while attending training.[2]

[2] See for instance Elliott (1991) for a discussion of private costs of human capital investment.

It is the complex decision making process at individual level we find interesting, an issue that most often is treated implicitly in the evaluation literature. By shifting focus from estimating mean effects of training to modelling and testing an individual's decision to participate, we take one step back in the decision making process compared to much of the evaluation literature. Basically, our main aim is to study the selection process itself. We analyse whether clients participating in active training separate themselves from non-participants and whether clients participating in general and specific training programs separate themselves from each other in terms of background characteristics.

Our approach is to use the background characteristics as indicators of an individual's cost of training and thereby his or her incentives to participate. Our main hypothesis is that differences in training costs are decisive in forming the individual's incentive to participate in training or not. Heterogeneity in innate abilities of an individual and/or the particular welfare benefits the client is eligible to receive, are among the important variables that influence costs of training. Specifically, based on human capital and signaling theory[3], we relate differences in previous education and age to differences in non-monetary costs (NMC) of training. Differences in previous income, spouse's income and status as recipient of VR benefits give raise to differences in monetary opportunity costs (MOC) of training.

The data consist of a random 10 percent sample of persons who entered the Norwegian vocational rehabilitation sector in the period from 1989 to 1993, a total of 6653

[3] The main difference between human capital and signaling interpretations of training is that signaling models allow for attributes that are not observed by the firm to be correlated with training. According to Weiss (1995) sorting models (signaling and screening models) of education (training) can best be viewed as extensions of human capital models. Sorting models extend human capital theory models by allowing for some productivity differences that firms do not observe to be correlated with the costs or benefits of schooling. In fact, Weiss argues that sorting models subsume all the features of human capital models. However, while human capital theory is concerned with determining the return to schooling, sorting models, while allowing for learning, focus on the ways in which schooling serves as either a signal or filter for productivity differences that firms cannot reward directly.

persons.[4] We find that the background characteristics of persons investing in educational training differ along several dimensions compared both to persons in work related training and to VR clients not participating in any training. Firstly, participants in educational training programs have background characteristics that indicate comparatively low NMC (disutility of effort) of participating in training programs.

Secondly, they also have background characteristics that indicate comparatively low MOC. Persons attending work related training have background characteristics indicating both higher NMC and higher MOC compared to participants in educational training. Non-participants have background characteristics indicating relatively high NMC of training compared to both of the other groups of clients.

The paper is organised as follows. In section 2 we present some important features of the Norwegian VR sector. In section 3 we model an individual's decision-making process, make a closer distinction between different cost components of training and discuss the incentives to train. The data used in the analysis are presented in section 4 along with descriptive statistics. Our main findings are presented in section 5 while conclusions and policy implications are drawn in section 6.

2. INSTITUTIONAL FEATURES OF THE VR SECTOR

Public concern about the level of investment in human capital among the unemployed and those outside the labour market is quite evident in OECD (1998). The report gives a description of the importance of OECD Governments attributing to the strategic role of human capital investments. Based on an international adult literacy survey, the report

[4] VR clients participating in educational training are usually integrated into ordinary classes in the public school system, or attend classroom training aimed specifically at unemployed people. The purpose of these programs is to enhance an individual's human capital in a way that is compatible with demands from a relatively large number of different types of jobs, i.e. general training. Work related training usually includes one or more of the following: employment in the public sector, wage subsidies, physical rehabilitation, sheltered work, and vocational training in specific occupational skills sometimes leading up to a vocational certificate. Work related training programs are often targeted at enhancing individual skills in a more narrow range of jobs, i.e. specific training.

concludes that job-related training by employed people constitutes a high proportion of all adult education and training activity. On average, people who are not employed are less likely to participate in job-related training. Those outside the labour market are more likely than employed people to participate in education and training unrelated to work. Also, participation in continuing training is strongly related to educational attainment. Those with less initial human capital appear to lack incentives or opportunities to acquire more later in life, creating the risk of exclusion. This is true for all countries in the study.[5] It is notable that the differences in participation rates between countries are as great as differences between well and poorly educated groups within countries. See also Chapman (1993) for a comprehensive discussion of different aspects of training: theory, empirical evidence and policy issues.

In Norway, concern about the gap of education between partly disabled and non-disabled workers has resulted in comprehensive use of educational programs of a general character. Education in the ordinary school system is the most widely used VR program aimed at partly disabled workers in Norway. Also, the VR sector offers several different work related training programs to enhance specific skills.[6] [7]

The Norwegian vocational rehabilitation sector has expanded rapidly since the national social insurance act was passed in 1966. The number of participants in training programs has stabilised during the 1990s. There are around 35,000 persons in different VR programs each day, which is around 1.5 percent of the labour force in Norway. The expansion has neither been guided by a firm knowledge of the overall economic impact of the training programs, nor by knowledge of who selects the different training

---

[5] Norway is not among the countries covered by the survey. Our study will be a contribution to analyzing whether the selection mechanism found elsewhere in Europe is replicated for Norway. See also OECD (1996) for a review of problems associated with assessing and certifying occupational skills in vocational education and training.

[6] There is a gap of general and specific skills between partly disabled and both unemployed and employed non-disabled workers in Norway. Less than 70 percent of the vocational rehabilitation (VR) clients in Norway have high school or more, while the same number is 75 percent for ordinary unemployed. In the labor force around 85 percent have high school or more. Furthermore, Hansen (1996) finds that VR clients have less work experience and are less likely to hold a vocational certificate compared to non/disabled workers.

[7] Note that job search assistance is not considered training. Such assistance is open for everyone at local labor market agencies at no direct cost to the user.

programs. Given the amount of public funds channelled into these areas of public policy and the ultimate goals of the spending, this may seem at odds with the present focus on accountability in the public sector.

The VR sector offers income maintenance payments and training programs for individuals with reduced productivity in the labor market due to medical conditions. The service varies in substance and duration across clients, reflecting a diverse clientele and broad orientation of vocational rehabilitation. All schooling and labour market training is free – no fees are paid by the participants - and extra expenses, such as commuting, are covered by the Social Security Office. The local Social Security Office and medical doctors usually assist in the application process. The local labour market authorities evaluate whether training may help increase the applicant's employment prospective, or help people keep their jobs. The decision to accept a person into a training program is mainly taken by caseworkers with the labour market authorities and local managers of vocational rehabilitation centres. The vocational rehabilitation administrators and caseworkers base participation decisions on subjective judgement regarding employment prospects, and upon available training slots in areas where the applicant has shown interest. In RTV (The National Social Insurance Organization) (1985) it was emphasised that an "evaluation of the clients' total situation in each case should be considered when a participation decision is made. The main inclusion criteria are health, age, personal characteristics, social conditions, education, and labor force experience."

The candidates for program participation may also influence the participation decision by information supplied to the program administrator about personal preferences, motivation, ability, appearance, etc. Thus, self-selection by individuals may be a contributing factor towards the training program a client is offered.

There may be a substantial time lag between the time of application, the creation of a rehabilitation plan, and acceptance into a program by caseworkers. During this period individuals may choose to drop out due to lack of motivation or acceptance of a disability pension. It may also be the case that some individuals receive and accept a job offer prior to the start of a training program and thus drop out. We do not directly observe who is offered a training slot but for some reason does not attend training.

Many of those who apply for training are eligible for rehabilitation benefit payment. The decision to accept rehabilitation benefits is made by the local Social Security Office, usually after a recommendation by a medical doctor.[8] However, program participants may also rely on other benefits, such as unemployment benefit, sickness benefit or social assistance. Also, a smaller group is accepted for disability pensions prior to going into a training program, but these persons are not 100 percent disabled, and are thus at least able to take part time jobs. Most of the VR clients who applied for training have been on sickness benefit before entering the vocational rehabilitation benefit scheme. The rehabilitation benefit is usually two-thirds of the gross income from the previous year subject to maximum and minimum benefit restrictions. Health status is the legal eligibility criterion for VR benefit but labour market prospects and the local Social Security Office, or the medical doctor, may also implicitly take social integration into account.

There is no maximum number of weeks a client may receive vocational rehabilitation benefit but normally periods do not exceed 3-4 years. Active program participation is only a part of the VR system. Clients may receive VR benefit for several years without participating in an active program.

Benefits usually cease upon return to work. Finally, the clients we are studying are all individuals who have a potential for returning to the labour market. They have not received medical diagnoses that conclude with 100 percent disability.

[8] Some institutional changes have been made in VR responsibilities in Norway. As from 1994, the labor market authorities decide on both rehabilitation benefit payments and program participation. These changes do not apply for our data.

3. THE ECONOMETRIC MODEL

We assume that VR clients are bounded rational, i.e. they cannot possibly foresee all the things that might matter for the decision to participate or not. Given that they have not perfect foresight, clients have to make decisions to participate in training or not based on less than perfect information. While the effects on income from training choices are influenced by a range of exogenous factors (e.g. the future conditions in the labour market, how well others do in the courses etc.), costs of training are to be born nearer in time and are basically intrinsic to the individual. Thus, it is more difficult for an individual client to sort choices according to expected training effects on income than to sort according to cost levels. Clients are myopic.

Ex ante, i.e. before a client has made the choice, a client's utility of a particular training choice can be represented by the utility function. $U_{ji} = I_{ji} - C_{ji}$. I_{ji} is expected income, i.e. income after having participated in a training program or, as the case may be, from not participating in a training program at all. C_{ji} is a vector representing training costs for individual *i* choosing alternative *j*. Training costs are zero if non-participation. Expected income is based on a probability distribution of finding work and knowledge of the distribution of the wage level(s) in jobs foreseen as being available to an individual ex ante. We assume that clients do not act on the differences in income but make their choices based on a ranking of costs associated with the different options available. Important for formulating the econometric model to be tested here is that we cannot observe a client's expected income level but we do observe indicators of the costs associated with training. We return to this shortly.

The difference in utility between state *j* and state *k* then becomes

$$U_{ji} - U_{ki} = C_{ki} - C_{ji}$$

Let U_{ji} be individual i's utility from choosing alternative $y_i = 1$ and U_{0i} be individual i's utility from choosing alternative $y = 0$. We have that[9]

(1) $P_i = \Pr(y_i = 1) = \Pr(U_{1i} > U_{0i})$

An individual's utility from the alternatives available to him/her has both a deterministic and a stochastic component. Using the case of two alternatives (participation, non-participation) as an example,

$$U_{1i} = U^{\circ}_{1i} + \varepsilon_{1i}$$
$$U_{0} = U^{\circ}_{0i} + \varepsilon_{0i}$$

where U°_{1i} and U°_{0i} are deterministic components and ε_{1i} and ε_{0i} are stochastic for participation and non-participation respectively.

Further specification of U°_{ji} is necessary. Let us initially divide the covariates into two groups. Let W_{ji} include variables that characterise the choices in question and let Z_i include variables that characterise the individual. We then have

$$U^{\circ}_{1i} = W'_{1i}\delta + Z'_i\gamma_0$$
$$U^{\circ}_{0i} = W'_{0i}\delta + Z'_i\gamma_0$$

We can now write the ith individual's probability of choosing alternative $y_i = 1$ as a function of the net benefit of choosing participation versus non/participation:

(2) $\Pr(y_i = 1) = \Pr(U_{1i} > U_{0i}) = \Pr(\varepsilon_{0i} - \varepsilon_{1i} < U^{\circ}_{1i} - U^{\circ}_{0i})$

[9] We model the decision making process as a utility maximizing process and base our model on McFadden (1973, 1976, 1978) and Domencich and McFadden (1975). McFadden's provision of a discrete utility theory has come to serve as a theoretical basis for discrete-choice models, as we employ here.

Let $U^{\circ}_{1i} - U^{\circ}_{0i} = (W'_{1i} - W'_{0i})\delta + Z'_i(\gamma_1 - \gamma_0) = X'_i\beta$

With different notation $X'_i = [(W_{1i} - W_{0i})', Z'_i]$ and $\beta = \begin{bmatrix} \delta \\ \gamma_1 - \gamma_0 \end{bmatrix}$. Thus

(3) $\quad \Pr(y_i = 1) = \Pr(U_{1i} > U_{0i}) = \Pr(\varepsilon_{0i} - \varepsilon_{1i} < X'_i\beta) = \Pr(\varepsilon^{\circ}_i < X'_i\beta)$

or

(4) $\quad \Pr(y_i = 1) = F(X'_i\beta)$

where $F(X'_i\beta)$ is the cumulative distribution of ε°_i evaluated at $X'_i\beta$. McFadden (1973) has shown that if each ε_{ki} (k = 0,1) is independent and identical type I extreme value distributed, then the distribution of the difference $\varepsilon_{0i} - \varepsilon_{1i} = \varepsilon^{\circ}_i$ will be logistic, and the choice probabilities can be expressed by the multinomial logit.

According to the multinominal logit model, individual *i* will choose alternative *j* among *J* alternatives with probability (normalising by setting $\gamma_0 = 0$)

(5) $$P_{ji} = \Pr(y_i = j) = \frac{e^{Z'_i\gamma_j}}{1 + \sum_{k=1}^{J} e^{Z'_i\gamma_k}}, j = 1,\ldots,J$$

We estimate and report the marginal effects specifically and they are given by the expression (for variable Z_{ik}):

(6) $$\frac{\partial P_{ji}}{\partial Z_{ik}} = P_{ji}(\gamma_j - \sum_{k=1}^{J} P_{ji}\gamma_j)$$

Turning to the more specific issues of this study, note that a VR client faces three choices: non-participation, educational training programs and work related training programs. The aim is to develop the reduced form estimation model linking, as briefly discussed above, background characteristics of the individual and training costs. We then go on linking the cost components to the probabilities of entering the three different states of 'non-participation', 'educational training' or 'work related training'.

In our reduced form maximum likelihood estimation, it is the background characteristics of the individuals that are used as explanatory variables, not the cost components. The cost components are not observable for the researcher, and measurements would under any circumstances be difficult to perform. Still, we assume that it is possible to make ordinal rankings within each cost component. We do this by stating the sign of partial derivatives of the cost functions, i.e. the direction of change in costs to changes in background characteristics. These effects are then used to state the expected sign of the probabilities of entering the different states as response to changes in the background characteristics, i.e. the expected sign of the marginal effects given by equation (6).

The Government pays the direct costs of training[10] but a client participating in training incurs other types of costs, too. In particular, there are two other basic cost components that may matter for the choices that the individuals are making: the non-monetary costs of training (C_N) and the monetary opportunity costs of training (C_M). Human capital investment theory and signalling theory inspired the division between non-monetary and monetary costs. The division allows a closer study of the driving forces behind the choices made by clients since we allow the composition and 'size' of the cost components to vary among clients.

Let total VR costs be represented by the cost function $C = f(C_N, C_M)$. The non-monetary costs of training (NMC) can be thought of as disutility or psychological costs. These are costs experienced by all individuals participating in training but the cost varies depending on how difficult one finds learning to be. Here NMC is represented by the function

$$C_N = g[\underset{-}{\text{EDU}}, \underset{+}{\text{AGE}}]$$

[10] In the Norwegian VR system clients can make their choices under a regime of no direct financial constraints i.e. the tuition fee is paid for by the Government. Direct training costs are therefore less relevant here.

Both education and age are commonly used as variables explaining differences in non-monetary costs of training and the signs of the partial derivatives are in line with assumptions made in signalling theory. The lower the former education level is and the higher is the age, the higher is the level of anguish and anxiety associated with participation.

According to job-market signalling theory[11], training is correlated with differences among workers that were present before training choices were made. An important unobservable ability correlated with training is the individual's capability of adapting to new requirements at the work place. Firms may make inferences about productivity differences from training choices, and the clients respond to this inference process by choosing different levels of training or different courses. One of the main hypotheses in signalling theory is that persons with low non-monetary costs of training or low disutility of effort are inclined to signal high productive ability by choosing or self-selecting to participate in training. The argument rests on a negative correlation between productive ability and size of the non-monetary costs of training, i.e. the higher the productive ability the lower the disutility of effort and vice versa.

Thus, here we assume that a client who has a relatively high educational level before entering VR is assumed to have a comparatively low disutility of education due to the fact that he or she has already completed educational programs in the past. In contrast, the older one gets the more costly it is to participate in terms of disutility.

On the other hand, in the standard human capital models training is treated as an investment decision[12] where the costs of training also consist of the monetary opportunity costs. Monetary opportunity costs take the form of foregone market opportunities. Time devoted to investment in human capital precludes the possibility of

[11] Spence's (1973,1974) analysis rests on the assumption that persons of low-ability find signaling through training more costly than do high-ability persons. More precisely, the marginal non-monetary cost (or disutility) of education and work specific training is higher for low-ability than for high-ability persons for every level of training (Gibbons (1992)).

[12] See Becker (1962,1964). Acemoglu and Pischke (1998) and Barron, Berger and Black (1999) are two recent contributions in the 'Becker tradition'. Also see Elliott (1991) for a text book presentation of human capital theory.

devoting that time to other market or non-market activities. Here, the monetary opportunity costs (MOC) of both types of training - work related or educational - is given by

$$C_M = h[\underset{+}{\text{INCOME}}, \underset{-}{\text{SPOUSEINC}}, \underset{-}{\text{VRBEN}}]$$

We assume that the individuals still perceive former income as a benchmark of their worth in the labour market. Thus, foregone market opportunities or monetary opportunity costs increases with higher former income level. On the other hand, we assume that the opportunity costs decrease as spouse income increases, i.e. the household subsidises the training costs for the individual. Finally, in Norway clients may receive VR benefit (dummy variable in the regression) while undergoing training. We assume that the opportunity costs decrease if the client is a recipient of VR benefits.[13]

The 'priors' on the likely estimates of participation in either of the programs are given by the following partial derivatives (non-participation is given by opposite signs)

(7) $\partial \text{Prob}(y_i=1)/\partial C_M(.) < 0$ and $\partial \text{Prob}(y_i=1)/\partial C_N < 0$

or in terms of marginal effects (equation (6)):

(8) $$\frac{\partial P_{1i}}{\partial EDU_i} > 0, \frac{\partial P_{1i}}{\partial AGE_i} < 0$$

(9) $$\frac{\partial P_{1i}}{\partial INCOME_i} < 0, \frac{\partial P_{1i}}{\partial SPOUSEINC_i} > 0 \text{ and } \frac{\partial P_{1i}}{\partial VRBEN_i} > 0$$

[13] It is possible to receive different types of financial support while participating in training but VR benefits (based on former income) represents normally the highest level of support available. For some work related training programs, in particular training at so called Labor Market Enterprises, it is possible to receive wages instead of VR benefits but that happens only later in the training process, i.e. at least six months after first enrolling in active training.

To sum up, our research strategy is not to estimate directly the different cost components' impact on the probability of participating in training or not[14] since direct measurement of cost is difficult to achieve. Rather we use indicators and let them enter a reduced form equation. In the reduced form estimation we allow the effects of changes in the background variables to influence the probability of participation directly, i.e. we estimate the marginal effects of changes in the variables on the probability of entering the different states.[15]

If participation or non-participation is completely random, there is no reason to expect results as predicted by the model. We hypothesise that the process contains systematic components and expect to find the following results:

1. Other things equal, younger and more educated clients are more likely to participate in active training programs, since they have comparatively low NMC. Vice versa for older and less educated clients.
2. Other things equal, clients with comparatively low previous income, high spouse income and being recipients of VR benefits are more likely to participate in active training, since they have comparatively low MOC. Vice versa for clients with comparatively high previous income, low spouse income and non-recipients of VR benefits.
3. Clients participating in active training differ systematically in terms of background characteristics and the composition of cost components, compared to non-participants.
4. Other things equal, clients with comparatively low NMC are more likely to participate in educational training programs than in work related training.

Point 3 is our main hypothesis and builds on points 1 and 2. Point 4 demands more discussion since it does not follow directly from (8) and (9).

[14] In the case of participation: the probabilities of choosing general or specific training.

[15] See Table 1 in the appendix for an overview of the total set of variables used in the empirical analysis.

Why should participants in educational training differ from participants in work related training? Spence's (1973, 1974) work on job-market signalling builds on the assumption that less productive or less able individuals have higher disutilty of educational training compared to more productive individuals. Thus, other things equal, using education as a signal of productive ability is more costly for the less productive. This implies that those clients with relatively high disutility of effort, other things equal (in particular equal MOC), are discouraged from participating in educational training programs. Obviously, discouraged clients have two other options. One of the options is to participate in work related training. Work related training is a more direct way of communicating information about productive ability compared to educational training. The individuals reveal information about their abilities since they are typically trained within a firm where their productivity is observed as opposed to indirect revelation (or signalling) through participation in educational programs. Loh (1994) analyses employment probation as a screening mechanism[16] and finds that probation induces self-selection. Those who accept jobs with probationary employment tend to be more efficient workers and less likely to quit than those who take jobs without probation. Building on Loh, work related training could function as a screening mechanism to discourage the least qualified VR clients from participating in work related training. And since signalling through educational training is also more costly for these clients, they will be more likely not to participate in active training at all.

Still, we cannot a priori rule out that clients with the comparatively lowest NMC, i.e. the clients with lowest disutility of effort and according to Spence the more able ones, may pool with other types of clients in work related training. After all, they have the least to fear being directly observed at the work place. Pooling of types is also possible in the educational training programs given that the less productive clients may have incentives to 'hide' as more productive clients, a point also discussed by Spence.

[16] Loh (1994) terms probation a sorting mechanism. We prefer to use sorting as a common term for both screening and signaling. Screening is the case where the uninformed party or the principal, designs and offers a contract to the agent before the agent takes any action. Signalling is the case where the agent chooses his/her action before the principal offers a contact (see e.g. Hillier (1997)).

Lastly, as Lang (1994) points out, the distinguishing characteristic of a sorting model is that knowing an individual's education provides employers with information about that individual's productivity which would be unknown otherwise. In human-capital models, education is not informative, because employers observe productivity directly. Thus, the models are distinguished by the role of education in conveying the individual's private information about his productivity. However, in both models units of human capital will be generated in the same way from inputs of innate ability and schooling/training.

All in all, we expect to find that the choices made follow a 'hierarchical structure':

- Individuals with comparatively low NMC will be more likely to choose educational training than work related training, other things equal.
- Individuals who are not discouraged from participating in active training but with comparatively high NMC will choose work related training, other things equal.
- Individuals who feel discouraged from participating in work related training due to possible revelation of low productive abilities and with comparatively high NMC of training will not participate in active training at all, other things equal.

Lastly, in the regressions we also use information on other background variables such as medical diagnosis, disability status, whether the client receives social benefits and whether the client was employed the year before entering the VR system. Note though that in the regressions we are only able to capture disability status and social benefit status as dummy variables.

4. DATA

Our data consist of people who were directed to the local labour market authorities for participation in a training program during the period 1989 to 1993. We have relatively detailed information on socioeconomic background characteristics, labour market participation, and health status for the persons in our sample. We observe gender, age, number of children below 18 years of age, education measured in years and type of education, own income before entering the VR system, spouse's income, work experience in years, medical diagnosis, and several social security and labour market states, such as training status, vocational rehabilitation benefit and social security benefits. All of these variables change over time, except gender.

Table 2 of the appendix show that mean age is lowest for participants in educational training. The gender variable shows that of the 1065 participants participating in educational training most are women (approximately 52 percent). The opposite is true both for work related training and for non-participation (approximately 61 percent out of 3927 cases and 54 percent out of 1661 cases are men). Approximately 70 percent of the participants in educational training have high school or longer education, while the numbers are down to approximately 60 percent for participants in work related training and 53 percent for non-participants. The share of clients with college or university education is also higher for the educational group.

While only approximately 18 percent of the participants in educational training had received social benefits, the share is 31 percent for work related training and 24 percent for non-participants. The share of clients with disability status is 31 percent in the non-participation group compared to 17.5 percent in the work related and 8 percent in the educational group.

Former income level is on average higher for the educational group and the same is the case for the share of the clients receiving VR benefits. Number of years of working experience is highest for non-participants (approximately 12 years), while work related training and educational training average 10 and 8 years, respectively. The average level of spouse's income is highest for non-participants.

5. RESULTS

Several individual characteristics appear to have an impact on the choice of whether to participate in active rehabilitation or not, and in case of participation, in which type of program to participate. The results from the multinomial logit model, reported as marginal effects, are given in Table 3 of the appendix. We start out the discussion of the result focusing on the central variables AGE, EDU, INCOME, SPOUSEINC and VRBEN.

Table 3 shows that the clients' age appears to increase the probability of not participating in active training. For those clients that do participate, the older a client is, the more likely it is that he or she will participate in work related training and the less likely the client chooses educational training. The signs of the effects align well with the discussion in section 3. The non-monetary cost of training is assumed to increase with age, reducing the (partial) probability of participation, as the data confirm. Explaining the distinction between those that participate in educational training and work related training involves a more subtle argument. For clients participating in work related training, it must be that the non-monetary costs (or disutility of effort) of educational training is higher for every time or effort 'unit'[17] compared to participation in work related training. As age increases, the non-monetary costs of educational training increase relatively more compared to work related training. Thus, the probability of choosing educational training decreases with age.

Former education, the second variable influencing the non-monetary costs of training, also 'behaves' nicely but with fewer significant results compared to the age variable. Clients with only secondary school are more likely *not* to train compared to clients with high school education. Table 3 in the appendix show that compared to those clients with only secondary school (or less), high school education reduces the probability of non-

[17] Think of the non-monetary costs of training in terms of 'disutility units', for instance disutility per hour, day or month.

participation. For clients with college education there is no significant difference compared to clients with only secondary school, though.

The results are in line with what we predicted in section 3. Higher education reduces the non-monetary costs of training and increases the probability of participation. We also find that college education reduces the probability of choosing work related training, while increasing the probability of choosing educational training programs. High school education on the other hand is not significant when it comes to explaining the choice of work related training but it is significant when explaining educational training. These results support the assumption made above that there are differences in disutility of effort between the two types of training programs. For clients that are relatively highly educated, the (marginal) non-monetary costs or disutility of educational training is lower for every time or effort 'unit' compared to work related training and vice versa for clients with secondary education or less.

Next, we turn to the variables explaining the monetary opportunity costs of training, INCOME, SPOUSEINC and VRBEN.

In section 3, we argued that being a recipient of VR benefits decreases the monetary opportunity costs of training and thereby increases the probability of taking part in active training. That prediction is confirmed in Table 3. Receiving VR benefits reduces the probability of non-participation. It is interesting to note that VR benefits also decrease the probability of taking part in work related training while increasing the probability of educational training. Adding that the size of spouse income significantly influences the probability of educational training, while having a non-significant effect on the probabilities of non-participation and work related training, gives strength to the arguments that (i) non-participants are different from participants and (ii) that participants in educational training differ from participants in work related training. The differences are most naturally interpreted as differences in NMC and MOC, as predicted in section 3.

Former level of income does not seem to have any significant effect on non-participation but the higher the former income is, the less likely it is that a client chooses work related training, while it increases the probability of participation in educational training. These results are also in line with the predictions in section 3.

Other explanatory variables also entered the regressions, and Table 3 shows that being male decreases the probability of non-participation. For those that participate, being male increases the probability of work related training while decreases the probability of education training. Turning the results around: women are more likely to choose educational training or non-participation and less likely to participate in work related training.

Being married significantly increases the probability of not participating in any program. The same is true for clients having status as (partly) disabled. As the number of years of work experience increases, the more likely it is that clients opt for non-participation. Being employed the year before registering as VR client has the opposite effect in the sense that it increases the probability of entering work related training. Recipients of social benefits have lower probability of taking part in educational training, while having children under 18 years of age increases the probability of taking part in educational training.

The dummy variables representing broad groups of diagnosis also have significant impact on some of choices made by VR clients (compared to the base category Diagnosis 0).

Finally, Table 3 shows that VR enrolment in 1993 significantly reduces the probability of non-participation while increasing the probability of participation.

We have tested[18] both whether the vector of all the coefficients in each state are significantly different between states, and whether single coefficients are different

[18] The test results are not reported in tables. Results are available from the authors upon request.

across states. We find that the vector of coefficients is significantly different overall, i.e. the explanatory variables generally have different effects on the probability of entering the different states. Furthermore, most of the single coefficient comparisons show the same. The important non-significant test results are the AGE coefficient; the dummy variables SOCBEN and VRBEN, which have non-significant differences when comparing non-participation and work related training. Comparing the EXPER variable, we find that there is no significant difference between work related and educational training. Likewise, SPOUSEINC is significantly different only between work related and educational training, while JOB_1 is significantly different between non-participation and work related training only.

To sum up, we find that our results support the view that participants in educational training differ from the participants in the other two groups. Interpreting the choice made by non-participants is not self-evident but we find, as hypothesised, that non-participants single themselves out compared to the active clients.

6. CONCLUDING REMARKS

We have estimated probabilities of entering training using a relatively large set of background characteristics of the individuals as explanatory variables. In the theoretical discussion sub-sets of these variables are linked to the two main cost components facing VR clients: non-monetary costs of training and the monetary opportunity costs of training. Participants in educational programs have both lower non-monetary costs of training and lower monetary costs of training compared to participant in work related training. Non-participants as a group are 'poor' in terms of decisive background characteristics. We have argued that being 'poor' increases the costs of training and lessens incentives to invest in human capital.

Our results partly fit an investment-signaling dichotomy.[19] Participants in educational training have background characteristics that indicate comparatively low disutility of training as we define it. We also find that the monetary opportunity costs of training is comparatively low for this group of clients making signalling an even more attractive option. A narrow interpretation of signalling theory is that we should expect to find that clients with relatively high educational levels dominate among participants in educational training and that active clients separate themselves into the two different programs depending on factors such as former educational level and age. Another interpretation is that while participation in educational training can be interpreted as a signalling decision, participation in work related training is the same as screening.

Participants in work related training do differ from the other groups of clients in terms of relevant background characteristics and although signalling theory does not rule our pooling of types, clients taking active part in training seem to take separate actions. Thus, participating in work related training could be interpreted as an investment decision rather than a signalling decision since the client reveals private information through his/her conduct at work. This may be a debatable conclusion. A different interpretation is that clients in practice have to choose between a signalling mechanism (participation in educational training) or a screening mechanism (participation in work related training) dismissing the pure investment argument altogether. Even non-participation could be termed a signal, perhaps the strongest negative signal of ability seen from prospective employers' point of view.

From a governmental perspective the main aim of active training is that clients should invest in human capital and productive skills so that, in turn, transition to disability pension is reduced. As we have tried to show, an individual client may find it more rational to either stay out of training altogether or seek to separate themselves from other clients by choosing different types of training. From a policy perspective the underlying reasons that may guide an individual's decision to choose educational

[19] See for instance Cohn, Kiker and Mendes De Oliveira (1987), Hungerford and Solon (1987), Belman and Heywood (1997) and Kroch and Sjoblom (1999) for empirical tests of education as human capital or signal. Based on these studies, signalling theory seems to lack decisive empirical support.

training versus work related could be of only minor interest as long as some kind of investment is made. On the other hand, using work related training and educational training as a sorting mechanism can be useful if clients opting for a disability pension stay out of active training anyhow and leave the VR resources available to clients with aspirations of entering the labour market again. We find results pointing in this direction since disability status significantly increases the probability of non-participation in active training.

Appendix

Table 1. Variables used in multinomial logit regression

Variable name	Definitions
EDU	Educational training, dummy variable.
WRT	Work related training, dummy variable.
NPRG	Non-participation, dummy variable.
AGE	Age, in years.
MALE	Dummy variable (1=male, 0=female).
HIGHSCH	Clients with at least high school education prior to VR, Dummy variable.
COLLEGE	Clients with at least college/university education prior to VR, Dummy variable.
CHILD18	Clients with children less then 18 years of age, Dummy variable.
MARRIED	Dummy variable (1=married, 0=not married)
SOCBEN	Dummy variable indicating if the person is receiving social benefit having the value one if the person is or has received benefits in the year prior to entering the vocational rehabilitation sector, and zero otherwise.
DISABIL	Disability pension is a dummy variable having the value one if the person is disabled (received disability pension) in the year prior to entering the vocational rehabilitation sector, and zero otherwise. The disability is less then 100 percent.
VRBEN	Dummy variable indicating if the person is receiving VR benefits.
INCOME	Income and spouse's income are measured in 100,000 Kroner (NOK). Income is measured for the year before the application.
EXPER	Number of years of working experience prior to VR.
SPOUSEINC	See INCOME
JOB_1	Dummy variable having the value one if the individual is employed in the previous period, and zero otherwise. A person is employed if she has an employment spell of at least 90 days in the particular year.
DIAGN0 – DIAGN5	Medical Diagnosis (0-5) are dummy variables. Medical diagnosis 0 is used as the base category in the regressions. More information on these dummy variables is available upon request from the authors.
YEAR89 – YEAR93	Year of entering VR (89-93) are dummy variables. Year89 is used as the base category in the regressions.

Table 2. Descriptive statistics

Variable	Mean	Std.Dev.	Minimum	Maximum	Cases
EDUCATIONAL TRAINING					
AGE	31.7596	9.0649	16.0000	64.0000	1065
MALE	.4836	.5000	.0000	1.0000	1065
HIGHSCH	.7042	.4566	.0000	1.0000	1065
COLLEGE	.1052	.3069	.0000	1.0000	1065
CHILD18	.3718	.4835	.0000	1.0000	1065
MARRIED	.3587	.4798	.0000	1.0000	1065
SOCBEN	.1831	.3869	.0000	1.0000	1065
DISABIL	.0779	.2682	.0000	1.0000	1065
VRBEN	.5897	.4921	.0000	1.0000	1065
INCOME	.9372	.7790	.0000	4.4690	1065
EXPER	8.9483	6.8366	.0000	26.0000	1065
SPOUSEINC	.6678	1.0788	.0000	10.9000	1065
JOB_1	.5418	.4985	.0000	1.0000	1065
DIAGN0	.2056	.4044	.0000	1.0000	1065
DIAGN1	.3793	.4855	.0000	1.0000	1065
DIAGN2	.1587	.3656	.0000	1.0000	1065
DIAGN4	.0620	.2412	.0000	1.0000	1065
DIAGN5	.0516	.2214	.0000	1.0000	1065
DIAGN6	.1427	.3500	.0000	1.0000	1065
YEAR89	.2272	.4192	.0000	1.0000	1065
YEAR90	.1972	.3981	.0000	1.0000	1065
YEAR91	.2188	.4136	.0000	1.0000	1065
YEAR92	.1634	.3699	.0000	1.0000	1065
YEAR93	.1934	.3952	.0000	1.0000	1065
WORK RELATED TRAINING					
AGE	35.4026	11.4586	15.0000	69.0000	3927
MALE	.6114	.4875	.0000	1.0000	3927
HIGHSCH	.5944	.4911	.0000	1.0000	3927
COLLEGE	.0530	.2240	.0000	1.0000	3927
CHILD18	.2832	.4506	.0000	1.0000	3927
MARRIED	.3491	.4768	.0000	1.0000	3927
SOCBEN	.3107	.4628	.0000	1.0000	3927
DISABIL	.1757	.3806	.0000	1.0000	3927
VRBEN	.4270	.4947	.0000	1.0000	3927
INCOME	.7765	.7482	.0000	4.3810	3927
EXPER	10.3305	8.2056	.0000	26.0000	3927
SPOUSEINC	.5477	.9099	.0000	10.5000	3927
JOB_1	.4856	.4999	.0000	1.0000	3927
DIAGN0	.2223	.4159	.0000	1.0000	3927
DIAGN1	.3020	.4592	.0000	1.0000	3927
DIAGN2	.3079	.4617	.0000	1.0000	3927
DIAGN4	.0532	.2245	.0000	1.0000	3927
DIAGN5	.0354	.1848	.0000	1.0000	3927
DIAGN6	.0792	.2701	.0000	1.0000	3927
YEAR89	.2401	.4272	.0000	1.0000	3927
YEAR90	.2175	.4126	.0000	1.0000	3927
YEAR91	.1956	.3967	.0000	1.0000	3927
YEAR92	.1693	.3751	.0000	1.0000	3927
YEAR93	.1775	.3821	.0000	1.0000	3927

NON-PARTICIPANTS					
AGE	38.5966	11.4593	16.0000	67.0000	1661
MALE	.5370	.4988	.0000	1.0000	1661
HIGHSCH	.5352	.4989	.0000	1.0000	1661
COLLEGE	.0686	.2529	.0000	1.0000	1661
CHILD18	.3372	.4729	.0000	1.0000	1661
MARRIED	.4654	.4990	.0000	1.0000	1661
SOCBEN	.2444	.4299	.0000	1.0000	1661
DISABIL	.3113	.4632	.0000	1.0000	1661
VRBEN	.5099	.5001	.0000	1.0000	1661
INCOME	.8565	.7485	.0000	4.1110	1661
EXPER	12.0692	7.8085	.0000	26.0000	1661
SPOUSEINC	.7296	1.0274	.0000	10.2000	1661
JOB_1	.4967	.5001	.0000	1.0000	1661
DIAGN0	.1559	.3629	.0000	1.0000	1661
DIAGN1	.4257	.4946	.0000	1.0000	1661
DIAGN2	.2185	.4134	.0000	1.0000	1661
DIAGN4	.0518	.2216	.0000	1.0000	1661
DIAGN5	.0476	.2129	.0000	1.0000	1661
DIAGN6	.1005	.3008	.0000	1.0000	1661
YEAR89	.2487	.4324	.0000	1.0000	1661
YEAR90	.2571	.4372	.0000	1.0000	1661
YEAR91	.2179	.4130	.0000	1.0000	1661
YEAR92	.1987	.3991	.0000	1.0000	1661
YEAR93	.0777	.2677	.0000	1.0000	1661

Table 3. Marginal effects. Multinomial Logit Model

Variables	Non-Participation	Work related	Educational
Constant	-.2391*	.2267*	.0124
	(.0314)	(.0351)	(.0236)
AGE	.0020**	.0038*	-.0059*
	(.0008)	(.0009)	(.0007)
HIGHSCH	-.0435*	-.0190	.0625*
	(.0117)	(.0133)	(.0095)
COLLEGE	-.0070	-.1150*	.1220*
	(.0234)	(.0262)	(.0155)
INCOME	.0065	-.0483*	.0417*
	(.0098)	(.0109)	(.0068)
SPOUSEINC	.0017	-.0111	.0094**
	(.0072)	(.0080)	(.0049)
VRBEN	-.0264**	-.0595*	.0859*
	(.0118)	(.0129)	(.0082)
MALE	-.0414*	.0855*	-.0441*
	(.0133)	(.0146)	(.0092)
MARRIED	.0359**	-.0194	-.0165
	(.0156)	(.0176)	(.0118)
DISABIL	.1423*	-.0394**	-.1029*
	(.0138)	(.0170)	(.0135)
EXPER	.0030**	-.0022	-.0008
	(.0012)	(.0014)	(.0010)
JOB_1	-.0275**	.0375*	.0100
	(.0128)	(.0142)	(.0091)
SOCBEN	.0198	.0227	-.0425*
	(.0142)	(.0156)	(.0105)
CHILD18	-.0126	-.0110	.0236**
	(.0135)	(.0152)	(.0100)
DIAGN1	.0919*	-.0893*	-.0026
	(.0164)	(.0177)	(.0111)
DIAGN2	.0179	.0405**	-.0584*
	(.0176)	(.0189)	(.0126)
DIAGN4	.0549**	-.0375	-.0175
	(.0272)	(.0293)	(.0181)
DIAGN5	.0739**	-.1217*	.0478**
	(.0292)	(.0323)	(.0198)
DIAGN6	.0869*	-.1182*	.0312**
	(.0220)	(.0239)	(.0140)
YEAR90	.0325**	-.0221	-.0104
	(.0155)	(.0176)	(.0118)
YEAR91	.0108	-.0112	.0004
	(.0161)	(.0181)	(.0117)
YEAR92	.0128	-.0030	-.0098
	(.0169)	(.0191)	(.0127)
YEAR93	-.1731*	.1290*	.0440**
	(.0209)	(.0215)	(.0125)
N= 6653			
χ^2 = 1135.04			
DF = 44			

* Significant at 1%. ** Significant at 5%.

REFERENCES

Acemoglu, D. and J.-S. Pischke (1998), "Why do firms train? Theory and evidence", *Quarterly Journal of Economics*, 79-119.

Ashenfelter, O. (1978), "Estimating the effect of training programs on earnings", *The Review of Economics and Statistics*, Vol. 6, No. 1, 47-57.

Barron, J.M., M.C. Berger and D.A. Black (1999), "Do Workers Pay for On-the-Job Training?", *Journal of Human Resources*, Vol. 34, No. 2, Spring, 235-252.

Bassi, L.J. (1984), "Estimating the effect of training programs with non-random selection", *Review of Economics and Statistics*, February, 36-43.

Becker, G.S. (1962), "Investment in human capital: a theoretical analysis", *Journal of Political Economy*, Vol. 70, Supplement, 9-49.

Becker, G.S. (1964), *Human Capital*, NBER: New York.

Belman, D. and J.S. Heywood (1997), "Sheepskin effects by cohort: implications of job matching in signaling model", *Oxford Economic Papers*, Vol. 49, 623-637.

Cohn, E, B.F. Kiker and M. Mendes De Oliveira (1987), "Further evidence on the screening hypothesis", *Economics Letters*, Vol. 25, 289-294.

Chapman, P. G. (1993), *The Economics of Training*, LSE Handbooks in Economics, HarvesterWheatsheaf: London.

Domencich, T. and D. McFadden (1975), *Urban Travel Demand*, North-Holland: Amsterdam.

Elliott, R. F. (1991), *Labor Economics*, McGraw-Hill: London.

Gibbons, R. (1992), *Game theory for applied economists*, Princeton University Press: New Jersey.

Hansen, H.-T. (1996), Hvem er på attføring og hva blir gjort? SNF-rapport 32/96 (in Norwegian).

Heckman, J.J. and R. Robb jr (1985), "Alternative methods for evaluating the impact of interventions", in Heckman J.J. and B. Singer (eds): *Longitudinal analysis of labor market data*. Cambridge University Press: Cambridge.

Heckman, J.J., V.J. Hotz and M. Dabos (1987), "Do we need experimental data to evaluate the impact of manpower training on earnings?", *Evaluation Review*, Vol. 11, No. 4, August, 395-427.

Heckman, J.J. and V.J. Hotz (1989), "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training", *Journal of the American Statistical Association*, Vol. 84, No. 408, December, Applications & Case Studies, 862-874.

Heckman, J.J. and J.A. Smith (1996), "Experimental and Nonexperimental Evaluation", in Schmidt, G., O'Reilly, J. and K. Schomann (eds): *International handbook of labour market policy and evaluation*. Elgar: Cheltenham.

Hillier, B. (1997), *The Economics of Asymmetric Information*, MacMillan Press Ltd.: London.

Hungerford, T. and G. Solon (1987), "Sheepskin effects in the returns to education", *The Review of Economics and Statistics*, Vol. 69, 175-177.

Kroch, E.A. and K. Sjoblom (1999), "Schooling as Human Capital or a Signal – Some Evidence", *The Journal of Human Resources*, 156-180.

Lang, K. (1994), "Does the Human-Capital/Educational- Sorting Debate Matter for Development Policy?", *American Economic Review*, Vol. 84, No. 1, 353-358.

Loh, E.S. (1994), "Employment Probation as a Sorting Mechanism", *Industrial and Labor Relations Review*, Vol. 47, No. 3, 471-486.

McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behaviour", in P. Zarembka (ed.), *Frontiers in Econometrics*. Academic: New York.

McFadden, D. (1976), "Qunatal Choice Analysis: A Survey", *Ann. Economics and Social-Measurement*, Vol. 5, 363-390.

McFadden, D. (1978), "Modelling the Choice of Residential Location", in Karlquist et al. (eds.), *Spatial Interaction Theory and Planning Models*. North-Holland: Amsterdam.

Moffitt, R. (1991), "Program evaluation with nonexperimental data", *Evaluation Review*, Vol. 15, No. 3, June, 291-314.

OECD (1996), *Assessing and certifying occupational skills and competences in vocational education and training*. Paris: OECD.

OECD (1998), *Human Capital Investments: An International Comparison*, Paris: OECD.

RTV (1985), quotation from report made by The National Social Insurance Organization, Oslo.

Spence, A.M. (1973), "Job Market Signaling", *Quarterly Journal of Economics*, 87, 355-74.

Spence, A.M. (1974), *Market Signaling: Informational Transfer in Hiring and Related Processes*, Harvard University Press: Cambridge.

Weiss, A. (1995), "Human Capital vs. Signaling Explanations of Wages", *Journal of Economic Perspectives*, 9, Fall, 133-154.

Chapter 4

Procurement Auctions with Entry of Bidders

International Journal of Industrial Organization
18 (2000) 1243–1257
International Journal of Industrial Organization
www.elsevier.com/locate/econbase

Procurement auctions with entry of bidders

Egil Kjerstad[a], Steinar Vagstad[b,*]

[a]*Foundation for Research in Economics and Business Administration, Breiviken 2, N-5035 Bergen-Sandviken, Norway*

[b]*Department of Economics, University of Bergen, Fosswinckelsgate 6, N-5007 Bergen, Norway*

Received 30 June 1998; accepted 30 November 1998

Abstract

In procurement auctions with a fixed number of bidders there is a tradeoff between cost efficiency and rent extraction. An optimal mechanism, therefore, entails distortions of effort (Laffont and Tirole, 1987). If potential suppliers must sink an entry investment before they can participate in the auction, then decreasing the firms' rent may imply reduced entry. We show that if potential bidders are uninformed before entry, commitment to a plain, nondistortive auction is optimal. In contrast, if potential bidders learn all their private information before entry, the optimal mechanism entails the same distortions as in Laffont and Tirole's static model.

Keywords: Incentive contracts; Auctions; Entry; Procurement

JEL classification: D44; D82; L51

1. Introduction

In many industrial countries, government procurement of goods and services reach 20% or more of GDP (World Bank, 1990). The sheer size of public procurement calls for concern about how it is organized. This paper considers

*Corresponding author.
E-mail address: steinar.vagstad@econ.uib.no (S. Vagstad).

0167-7187/00/$ – see front matter
PII: S0167-7187(98)00055-1

procurement of fixed quantities of well-defined goods or services. In such situations, most procurers seem to agree on using some variant of what can be called a 'plain' auction: Invite potential suppliers to submit bids, and choose the one with the lowest bid. This practice corresponds to what theory prescribed before 1980 (see Vickrey, 1961; Demsetz, 1968), but does not meet the standards set by more recent theory.

What does the recent theory prescribe, then? By the mid-1980s what became known as agency theory was applied to procurement problems and the closely related problem of monopoly regulation. In particular, Laffont and Tirole (1987); Riordan and Sappington (1987); McAfee and McMillan (1987a) have independently shown that a procurer can improve upon a plain auction among a fixed number of contenders for a procurement contract: by auctioning carefully designed incentive contracts instead of fixed-price contracts, the induced cost inefficiencies are more than compensated for by the extraction of information rent.[1] This is still state of the art.

True, this literature has contributed important insights. However, it is founded on the assumption that the number of bidders is fixed. The purpose of the present paper is to explore the consequences of having the number of suppliers depend on how procurement is organized. As there are entry costs for most firms in most industries, the number of firms within an industry should usually reflect the profitability of the industry, and the government is so large a buyer that its contracting practice may affect the profitability of entire industries significantly (e.g., defense procurement and the road construction industry). However, note that as there are often costs associated with bid preparation, also 'small' procurers may experience adverse effects of 'contractual smartness' on the number of bidders.

Our analysis is cast in the framework of a government that wants one out of several firms to carry out an indivisible project, e.g., the delivery of some amount of a well-defined good or service. We combine two related fields of literature on procurement and auctions. One is the already mentioned 'auctioning incentive contracts' literature, focusing on the desirability of rent extracting mechanisms in a setting with a fixed number of potential suppliers (see also Dasgupta and Spulber, 1990). The other field is the recently developed theory of auctions with entry (Samuelson, 1985; McAfee and McMillan, 1987b; Levin and Smith, 1994).[2] Following the literature on auctions with entry we assume that each potential supplier must sink a relation-specific entry investment before he can participate in

[1] In their seminal paper on regulation of a firm with variable output, Baron and Myerson (1982) showed that the government can do better than use a fixed-price schedule. By reducing the output incentives of high-cost types of the firm the government makes it less tempting for a low-cost type to pretend to have high costs, in turn reducing the information rent of a low-cost firm.

[2] The latter theory does not consider incentive contracts, but focuses on optimal fixed-price mechanisms and, in particular, on whether the auctioneer should try to improve a plain auction by using reservation prices or entry fees.

the procurement auction.[3] After entry our setup is exactly as in Laffont and Tirole (1987): Before submitting bids, each firm knows its cost parameter (an adverse selection parameter) and, if chosen to produce, the firm decides how much effort (moral hazard variable) to put forth. Both the cost parameter and the effort are private information for the firm, while their compound effect—the *cost*—is observed by the principal.

Equilibrium entry requires each potential firm to enter if and only if the expected profit is large enough to cover the entry cost. The problem with rent-extracting mechanisms in a model with entry is that reducing the firms' expected rent also reduces their incentives to enter, possibly resulting in higher expected project costs if the number of bidding firms is reduced. This raises the question of the nature of the optimal mechanism. In particular, is the optimal mechanism just a modified Laffont and Tirole (1987) mechanism or is it radically different?

The paper is organized as follows. In Section 2 we present a Laffont and Tirole (1987) model with entry of firms, assuming that the firms do not learn their cost parameters before making their entry decisions (as in Levin and Smith, 1994; McAfee and McMillan, 1987b). In contrast, in Section 3 we study the opposite case, in which the firms learn their cost parameters before entry (as in Samuelson, 1985). Some concluding remarks are gathered in Section 4.

2. A model of uninformed entry

Consider a government that wants an indivisible project to be carried out by one of several agents. Ex ante there are N potential agents. Before an agent can serve the government, however, he must sink an irreversible and completely relation-specific entry investment of size $I > 0$. We assume that N is so large that it is not desirable to have all agents invest. Let n denote the number of agents that actually enter.

Agent (i.e. firm) i's cost at the production stage is given by $C^i = \beta^i - e^i$, where β^i is an intrinsic cost parameter and e^i is agent i's cost-reducing effort. The cost parameters β^i are independently drawn from a common distribution F on an interval $[\underline{\beta}, \bar{\beta}]$. Let f denote the density function.

The production cost C^i is observable and verifiable ex post, while β^i and e^i are

[3] The entry investment may take different forms: expenditures incurred by establishing the organization necessary to be taken seriously as a contender for the project; bid preparation costs (i.e., expenditures in calculating cost figures for the project); R&D expenditures (Tan, 1992; Piccione and Tan, 1996 study different fixed-price mechanisms in situations with variable entry investments, with the variable component interpreted as R&D investments); capacity (due to the size of the government's purchase); or opportunity costs (the supplier may have to decline other customers' demands while preparing delivery).

private information for agent i. In this section it is assumed that the agent must enter before learning his cost parameter (to be relaxed in Section 3 below). The payoff for any agent i that has made the investment is given by

$$u^i = \begin{cases} t^i - I - \tau & \text{if not selected} \\ t^i - I - \tau - \psi(e^i) & \text{if selected} \end{cases} \tag{1}$$

where t^i is the net transfer (in addition to cost reimbursement) to the agent at the production stage, e^i is his effort, τ is a positive or negative entry fee charged by the principal, and ψ is the agent's disutility of effort. It is further assumed that $\psi(0)=\psi'(0)=0$ and that $\psi'(e)>0$ and $\psi''(e)>0$ for $e>0$. If an agent has not made the investment, his utility is normalized to zero.

The principal seeks to maximize welfare w. If n firms have made the investment and firm i is selected to produce, w is given by

$$\begin{aligned} w &= S - (1+\lambda)\left[C^i + \sum_j t^j - n\tau\right] + \sum_j u^j \\ &= S - (1+\lambda)\left[\beta^i - e^i + \psi(e^i) + nI\right] - \lambda \sum_j u^j \end{aligned} \tag{2}$$

where S is the value of having the project carried out and λ is the shadow cost of public funds. The second line is obtained by replacing $\Sigma_j t^j$ by $\psi(e^i) + nI + n\tau + \Sigma_j u^j$ (from Eq. (1)) and C^i by $\beta^i - e^i$ and rearranging. Welfare can be interpreted as the sum of consumers' surplus (the project value minus its net costs) and producers' surplus.[4] The entry cost I, the disutility of effort function ψ, the distribution function F and the parameters S and λ are assumed to be common knowledge.

In what follows we will characterize the optimal procurement mechanism.[5] A mechanism is a set of rules specifying: (i) which agent to select; (ii) cost requirements for the selected agent; (iii) how much the different agents will be paid; and (iv) the amount to be paid upon entry (i.e., the entry fee). Each of these rules can in principle depend on the number of entrants and on the realization of their cost parameters. However, with risk neutrality there is no point in making the entry fee depend on cost parameters that are unknown at the time of entry.

[4] To simplify the exposition, we assume that if a firm has entered there is a price at which production is mutually beneficial for the firm and the principal. A sufficient condition for this is that $S \geq (1+\lambda)\bar{\beta}$ (for the mechanisms that will be considered here).

[5] Attention is restricted to static mechanisms, excluding the possibility of having the bidders enter one at a time and stop further entry if a 'satisfactory' firm has entered, thereby saving entry costs. Burguet and Sákovics (1996) show that an optimal sequential mechanism (with rounds of auctions) is better than the optimal static mechanism, provided that the costs of delay from prolonging the entry process are negligible. See also McAfee and McMillan (1988), who extend the Revelation Principle to sequential mechanisms.

Moreover, we will restrict attention to symmetric equilibria, making also the number of entrants stochastic. By the same logic it is, therefore, no reason to make the entry fee depend on the number of entrants, either. Finally, we restrict attention to direct mechanisms, that is, mechanisms in which the agents are asked to report their cost parameters. The optimal mechanism is then found by maximizing welfare over the set of mechanisms that induce truth-telling. Using the one-to-one correspondences between costs and effort levels and between transfers and utilities once the vector of cost parameters is known, a general direct mechanism can now be written $M=\{e^i_n(\beta), x^i_n(\beta), u^i_n(\beta), \tau\}$, where $e^i_n(\beta)$ is the effort of a selected firm i when n firms have entered and drawn cost parameters $\beta=(\beta^1,\ldots,\beta^n)$; $x^i_n(\beta)$ is the probability by which firm i is chosen in the same situation; $u^i_n(\beta)$ is firm i's utility; and τ is the entry fee.

As already indicated, we restrict attention to symmetric mixed-strategy equilibria, in which each of the N agents enters with the same probability $q\in(0,1)$ and stays out with probability $1-q$.[6] This implies that the number of firms that actually enter, n, follows a binomial distribution with mean qN and variance $(1-q)qN$. In a mixed-strategy equilibrium, each agent's expected returns from entry equals zero, no matter which mechanism the principal uses. (If not, randomized entry cannot be an equilibrium strategy.) Therefore,

$$\sum_{n=1}^{N}\binom{N}{n}q^n(1-q)^{N-n}\sum_{j=1}^{n}E_\beta[u^j_n(\beta)]=0 \tag{3}$$

The value of q that satisfies Eq. (3) characterizes equilibrium in mixed strategies. Taking Eq. (3) into account, expected welfare (i.e., the expected value of the expression in Eq. (2), expectation taken over n as well as β) can be written as follows:

$$E_{\beta,n}[w]=\sum_{n=1}^{N}\binom{N}{n}q^n(1-q)^{N-n} \times\left\{\underset{\beta}{E}\left[\sum_{i=1}^{n}x^i_n(\beta)[S-(1+\lambda)[\beta^i-e^i+\psi(e^i)+nI]]\right]\right\} \tag{4}$$

Different mechanisms may affect expected welfare in two ways. First, a mechanism may affect the equilibrium entry probability q by affecting the terms $\Sigma_j E_\beta[u^j_n(\beta)]$ in Eq. (3). Second, for a given entry probability a mechanism may affect welfare through the term $E_\beta\{\Sigma^n_{i=1}x^i_n(\beta)[S-(1+\lambda)[\beta^i-e^i+\psi(e^i)]]\}$ of Eq. (4).

To find the optimal mechanism, we start by noting that whatever effort,

[6] The analysis in this section builds on Levin and Smith (1994), and we refer the reader to Levin and Smith for the fine details of the argument. Pure-strategy equilibria, in which n agents enter with certainty and the remaining $N-n$ agents stay out, are studied in Kjerstad and Vagstad (1996).

selection and transfer rules (that is, $\{e_n^i(\cdot), x_n^i(\cdot), u_n^i(\cdot)\}$) the principal chooses, any desired entry probability can be achieved by setting an appropriate level of the entry fee, τ. Moreover, as τ does not enter expression Eq. (4) and affect welfare only by altering the equilibrium entry probability, the entry probability is a free decision variable. Therefore, a necessary condition for a mechanism to be optimal is that it maximizes $E_\beta\{\Sigma_{i=1}^n x_n^i(\beta)[S-(1+\lambda)[\beta^i - e^i + \psi(e^i)]]\}$ for each realized number of firms. For subsequent reference, we will call any such mechanism an *ex post efficient* mechanism. Ex post efficiency implies that the agent with the lowest cost parameter is selected, and that he exerts the cost minimizing effort, e^*, defined by $\psi'(e^*)=1$.[7]

Drawing attention to what we in the introduction called *plain auctions* is now convenient. Formally, by a plain auction we mean (in our procurement context) any rule by which: (i) firms bid for fixed-price contracts; and (ii) bidder i is selected to produce and is paid only if his bid, B_i, is the lowest and not exceeding the principal's maximum willingness-to-pay (i.e., iff $\min\{B_k\} \leq S/(1+\lambda)$).[8] Examples of plain auctions are the English, Dutch, first-price sealed bid and second-price sealed bid (Vickrey) auctions. Plain auctions have the appealing property that they entail ex post efficiency: by involving fixed-price contracts only they make the producing agent residual claimant of the savings from cost-reducing effort, hence $e^i = e^*$. Moreover, among all bidders the one with lowest cost $\beta^i - e^* + \psi(e^*)$, that is, lowest β^i, is selected. This proves the following result:

Proposition 1. *With mixed-strategy entry, any ex post inefficient mechanism is inferior to an appropriately set entry fee followed by a plain auction.*

The important assumption driving this result is that from an ex ante point of view bidders have no informational advantage. Therefore, rent extraction is not an issue, and attention can safely be restricted to mechanisms that are ex post efficient.[9]

We are now almost done with the case of mixed-strategy entry. Levin and Smith (1994) have in their analysis of different fixed-price schedules shown that with

[7] Costs are minimized when the marginal disutility of effort equals the marginal reduction in costs resulting from extra effort, $\psi'(e) = -\partial C/\partial e = 1$.

[8] If only one firm has submitted a bid below $(1+\lambda)S$ in a second price auction, then the price is set equal to $(1+\lambda)S$. The number $(1+\lambda)S$ may, therefore, be interpreted as the government's own bid.

[9] For similar reasons, Levin and Smith (1994), Proposition 1, have found that entry fees are better means than reservation prices: both extract rents ex ante, but reservation prices entail ex post inefficiencies in that sometimes the project is not carried out even if the value of the project exceeds the costs.

mixed-strategy entry, adding entry fees to a plain auction reduces welfare (their Proposition 6).[10] Combined with Proposition 1 this implies the following result:

Proposition 2. *With mixed-strategy entry, committing to a plain auction without entry fees is optimal.*

Note, however, that this does not imply that a plain auction implements the first-best allocation, only the best allocation possible given that firms conform to a symmetric mixed-strategy entry equilibrium. There is still a welfare loss due to lack of coordination, as the number of firms that actually enter is a random number that may or may not equal the first-best number.[11]

While mixed-strategy entry is a plausible assumption if there are no coordination devices, pure-strategy entry is plausible if entry decisions can be coordinated, e.g., if prospective bidders make their entry decisions sequentially and the decisions are observable or can be communicated to other prospective bidders before they decide. McAfee and McMillan (1987b) have shown that with pure-strategy entry, a plain auction yields first-best entry. Unlike the mixed-strategy case, however, now the information rent is not totally dissipated, but dissipated down to an integer approximation only. Intuitively, if the principal and the agents are symmetrically informed before entry, the principal can implement the first-best outcome by committing to a mechanism consisting of a plain auction and an entry fee designed to extract all expected rent. If, in contrast, the principal is imperfectly informed about some features of the problem, e.g. the entry cost, at the time she must make the commitment, then she cannot use the entry fee to extract information rent without risking reduced entry. However, also now it remains clear that if entry fees can be used there is no scope for ex post inefficient mechanisms. If one wishes to deviate from a plain auction, the sensible way to do it is to

[10] Intuitively, a plain auction makes the entrant residual claimant of the benefits from his investment, and he has therefore the socially right incentives to enter (see the discussion on p. 593 in Levin and Smith).

[11] We define the *first-best outcome* as the outcome when the principal has the same information as do the agents, she controls the agents' actions (investments and efforts) and maximizes welfare subject to the constraint that any agent's expected utility must be non-negative. Clearly, the first-best allocation requires *ex post efficiency* and no rents. If n firms have entered, expected welfare can then be written

$$W^*(n) = S - (1+\lambda)\left[\int_{\underline{\beta}}^{\bar{\beta}} \tilde{\beta}' \mathrm{d}\left[1-\left[1-F(\tilde{\beta}')\right]^n\right] - e^* + \psi(e^*) + nI\right]$$

where the integral represents the expected value of the lowest cost parameter in a sample of size n. As there are diminishing marginal benefits and constant marginal costs associated with entry, $W^*(n)$ has a unique maximizer n^*.

regulate entry directly, by taxing entry. After entry there is no reason to deviate from a plain auction.

3. Informed entry

In the preceding we have assumed that the agents do not learn their private information until after entry. In this section we explore the consequences of having the firms learn their cost parameters *before* they make their entry decision, building on the work by Samuelson (1985). As in the preceding section, we assume that the principal commits to a mechanism before the firms enter, and that the firms observe the number of entrants before they submit bids.

In a symmetric equilibrium, each of the N agents enters if his costs are sufficiently low, that is, if β^i is smaller than or equal to a common cutoff parameter denoted $\hat{\beta}$. This implies that the number of firms that actually enter, n, still follows a binomial distribution, now with mean $F(\hat{\beta})N$ and variance $[1 - F(\hat{\beta})]F(\hat{\beta})N$. To simplify notation, let $U_n^i(\beta^i) = E_{\beta^{-i}}[u_n^i(\beta)]$ denote the expected utility of individual i having cost parameter β^i when exactly n firms have entered (expectation taken over all other firms' cost parameters $\beta^{-i} = (\beta^1, \dots, \beta^{i-1}, \beta^{i+1}, \dots, \beta^n)$).

For $\hat{\beta}$ to be an equilibrium cutoff cost parameter, the expected utility of a firm with this cost parameter must equal zero. That is,

$$\sum_{n=1}^{N} \binom{N-1}{n-1} F(\hat{\beta})^{n-1} [1 - F(\hat{\beta})]^{N-n} U_n^i(\hat{\beta}) = 0 . \tag{5}$$

Note that the equilibrium condition Eq. (5) does not fix the utility of an entrant with parameter $\hat{\beta}$ for any particular number of entrants, only the expected utility, expectation taken over all possible number of other entrants. (Among all possibilities, it can be shown that setting $U_n^i(\hat{\beta}) = 0$ is as good as anything else.)

Let

$$P_n(\hat{\beta}) \equiv \binom{N}{n} [F(\hat{\beta})]^n [1 - F(\hat{\beta})]^{N-n}$$

denote the binomial probability that exactly n firms enter. The principal's full problem can then be written as follows:

$$\max_{x_n^i(\cdot), e^i(\cdot), U_n^i(\cdot), \hat{\beta}} \Bigg\{ \sum_{n=1}^{N} P_n(\hat{\beta}) E_\beta \times \left[\sum_{i=1}^{n} x_n^i(\beta) [S - (1+\lambda)[\beta^i - e^i(\beta) + \psi(e^i(\beta))]] - \lambda \sum_{i=1}^{N} U_n^i(\beta^i) \right] - (1+\lambda) N F(\hat{\beta}) I \Bigg\} \tag{6}$$

subject to

$$U_n^i(\beta^i) \geqslant -I,\ n=1,\ldots,N,\ \beta^i \in [\bar{\beta}, \hat{\beta}], \tag{7}$$

$$\frac{\mathrm{d}U_n^i(\beta^i)}{\mathrm{d}\beta^i} = -E_{\beta^{-i}}[x_n^i(\beta)\psi'(e^i)],\ n=1,\ldots,N,\ \beta^i \in [\underline{\beta}, \hat{\beta}],\ \text{and} \tag{8}$$

$$\sum_{n=1}^{N} \binom{N-1}{n-1} F(\hat{\beta})^{n-1}[1-F(\hat{\beta})]^{N-n} U_n^i(\hat{\beta}) = 0\,. \tag{9}$$

The objective function Eq. (6) is simply the expected value of w as given in Eq. (2), expectation taken over possible values of n and possible values of β, given n. The constraints arise for the following reasons. First, since an agent who has invested can always secure a utility of $-I$ by not accepting a contract, this number is a lower bound on individual utilities, hence Eq. (7). Second, after entry, marginal incentives operate exactly as in Laffont and Tirole (1987) fixed-n model, hence Eq. (8). The final constraint Eq. (9) is the equilibrium cutoff condition discussed above.

We will solve this problem in two steps. First, suppose the principal chooses a cutoff $\hat{\beta}$ and implements this choice by setting a vector $(U_1^i(\hat{\beta}),\ldots,U_N^i(\hat{\beta})) = (\hat{U}_1,\ldots,\hat{U}_N)$ such that $\hat{U}_n \geqslant -I$ for all n and such that Eq. (9) holds. (The question of which $\hat{\beta}$ to choose and which vector to support it will be studied below.) When $(U_1^i(\hat{\beta}),\ldots,(U_N^i(\hat{\beta}))$ is given, so is $\hat{\beta}$, and the remaining problem is for each n to solve the following:

$$\max_{x_n^i(\cdot),e^i(\cdot),U_n^i(\cdot)} \left\{ E_\beta \left[\sum_{i=1}^{n} x_n^i(\beta)[S-(1+\lambda)[\beta^i - e^i(\beta) + \psi(e^i(\beta))]] - \lambda \sum_{i=1}^{N} U_n^i(\beta^i) \right] \right\} \tag{10}$$

$$\text{s.t. } U_n^i(\hat{\beta}) \geqslant \hat{U}_n,\ \text{and} \tag{11}$$

$$\frac{\mathrm{d}U_n^i(\beta^i)}{\mathrm{d}\beta^i} = -E_{\beta^{-i}}[x_n^i(\beta)\psi'(e^i)],\ \beta^i \in [\underline{\beta}, \hat{\beta}]\,. \tag{12}$$

Except the possibility that the reservation utility level $\hat{U}_n$ may be different from zero, each of these n problems is identical to the problem solved by Laffont and Tirole (1987), with cost parameters β^i drawn from $[\underline{\beta}, \hat{\beta}]$ according to a cumulative distribution function $\tilde{F}(\beta^i) \equiv F(\beta^i)/F(\hat{\beta})$. Their solution concerning effort and selection must, therefore, also be the same, while the expected utility of the firms (for any given number n of entrants) may differ from the corresponding utility in Laffont and Tirole (1987) to the extent $\hat{U}_n$ differ from zero. Therefore, the most efficient firm is chosen to produce, its effort incentives are given by

$$\psi'(e^i)=1-\frac{\lambda}{1+\lambda}\frac{\tilde{F}(\beta^i)]}{\tilde{f}(\beta^i)}\psi''(e^i)=1-\frac{\lambda}{1+\lambda}\frac{F(\beta^i)}{f(\beta^i)}\psi''(e^i)$$

(note that the hazard rate is unaffected by truncation of the distribution F from above), while the expected rent of a firm with cost parameter $\beta^i\in[\underline{\beta},\hat{\beta}]$ is given by

$$U_n^i(\beta^i)=\hat{U}_n+\int_{\beta^i}^{\hat{\beta}}\left[1-\frac{F(\tilde{\beta})}{F(\hat{\beta})}\right]^{n-1}\psi'(e(\tilde{\beta}))\mathrm{d}\tilde{\beta}.$$

What remains is to find the optimal level of entry, $\hat{\beta}$. Following Samuelson (1985) we now exploit the fact that we need not condition on n (we conditioned on n in the analysis above in order to exploit the similarities with the Laffont and Tirole (1987) analysis): for any given $\hat{\beta}$, n is a function of the vector of cost parameters $(\beta^1,\ldots,\beta^N)$. Moreover, since the optimal mechanism is only concerned with the most efficient firm, we need only consider that particular firm too. Let $G(\beta^i)=1-[1-F(\beta^i)]^N$ denote the cumulative distribution function of the first-order stochastic in the entire sample of size N, while $g(\beta^i)=N[1-F(\beta^i)]^{N-1}f(\beta^i)$ denotes the corresponding density. Expected utility of a firm with cost parameter β^i (unconditioned on n) can be written

$$\begin{aligned}U^i(\beta^i)&=\int_{\beta^i}^{\hat{\beta}}[1-F(\tilde{\beta})]^{N-1}\psi'(e(\tilde{\beta}))\mathrm{d}\tilde{\beta}\\&\quad+\sum_{n=1}^{N}\binom{N-1}{n-1}F(\hat{\beta})^{n-1}[1-F(\hat{\beta})]^{N-n}\hat{U}_n\,.\end{aligned}$$

From the equilibrium entry condition Eq. (5) we know that the last term equals zero. This proves that it does not matter how each entry-dependent reservation utility, $\hat{U}_n$, is set; only their weighted average matters.

Now the sum of expected utilities is found by taking expectation also over β^i and multiplying by N. The result is

$$\begin{aligned}E_\beta\left[\sum_{i=1}^{N}U^i(\beta^i)\right]&=N\int_{\underline{\beta}}^{\hat{\beta}}\int_{\beta^i}^{\hat{\beta}}[1-F(\tilde{\beta})]^{N-1}\psi'(e(\tilde{\beta}))\mathrm{d}\tilde{\beta}f(b^i)\mathrm{d}\beta^i\\&=N\int_{\underline{\beta}}^{\hat{\beta}}F(\tilde{\beta})[1-F(\tilde{\beta})]^{N-1}\psi'(e(\tilde{\beta}))\mathrm{d}\tilde{\beta}\,.\end{aligned}\tag{13}$$

Consequently, expected welfare can be written

$$\begin{aligned}E_\beta[w]&=\int_{\underline{\beta}}^{\hat{\beta}}[S-(1+\lambda)[\tilde{\beta}-e(\tilde{\beta})+\psi(e(\tilde{\beta}))]]\,\mathrm{d}G(\tilde{\beta})-\lambda NE_\beta\left[\sum_{i=1}^{N}U^i(\beta^i)\right]\\&\quad-(1+\lambda)NF(\hat{\beta})I\end{aligned}\tag{14}$$

Differentiating Eq. (14) with respect to $\hat{\beta}$ and rearranging using Eq. (13) yields the following optimality condition:

$$\frac{\mathrm{d}}{\mathrm{d}\hat{\beta}} E_{\beta}[w] = [S - (1+\lambda)[\hat{\beta} - e(\hat{\beta}) + \psi(e(\hat{\beta}))]]g(\hat{\beta}) - \lambda NF(\hat{\beta})[1 - F(\hat{\beta})]^{N-1}\psi'(e(\hat{\beta})) - (1+\lambda)Nf(\hat{\beta})I = 0 \qquad (15)$$

We see that increasing the cutoff entails one benefit and two cost components. First, the increase may trigger entry by the first firm, implying a gross (of the entry cost) social surplus of $S - (1+\lambda)[\hat{\beta} - e(\hat{\beta}) + \psi(e(\hat{\beta}))]$. This happens with probability $g(\hat{\beta})$. Second, as the cutoff increases, more rent has to be given up to the firms, as measured by the term $NF(\hat{\beta})[1 - F(\hat{\beta})]^{N-1}\psi'(e(\hat{\beta}))$.[12] Third, the expected cost of entry is $NF(\hat{\beta})I$. Increasing the cutoff marginally therefore leads to an increase in aggregate entry cost of $Nf(\hat{\beta})I$, and this component has weight $(1+\lambda)$. Using that $g(\hat{\beta}) = N[1 - F(\hat{\beta})]^{N-1}f(\hat{\beta})$, this expression can be further simplified:

$$f(\hat{\beta})[S - (1+\lambda)[\hat{\beta} - e(\hat{\beta}) + \psi(e(\hat{\beta}))] = \lambda F(\hat{\beta})\psi'(e(\hat{\beta})) + \frac{(1+\lambda)If(\hat{\beta})}{[1 - F(\hat{\beta})]^{N-1}} \qquad (16)$$

Except the entry cost term, Eq. (16) is identical to the condition for optimal shutdown of high-cost firms found in the Laffont and Tirole (1993) (p. 318) model without entry. The condition also bears some resemblance with the condition for procurement cost minimization in Samuelson (1985) fixed-price model. With our notation, Samuelson's condition, his Eq. (8) reads

$$\beta_0 - \hat{\beta} = \frac{F(\hat{\beta})}{f(\hat{\beta})} + \frac{I}{[1 - F(\hat{\beta})]^{N-1}} \qquad (17)$$

where β_0 is the principal's own cost parameter. (The relation between the principal's own production costs and her project valuation is given by $S = (1+\lambda)[\beta_0 - e^* + \psi(e^*)]$.) Rewriting Eq. (16) yields

$$\beta_0 - \hat{\beta} - [e^* - \psi(e^*) - e(\hat{\beta}) + \psi(e(\hat{\beta}))] = \frac{\lambda}{1+\lambda}\frac{F(\hat{\beta})}{f(\hat{\beta})}\psi'(e(\hat{\beta})) + \frac{I}{[1 - F(\hat{\beta})]^{N-1}} \qquad (18)$$

[12] This expression has a particularly simple interpretation if we restrict attention to second price (or Vickrey) mechanisms, in which the winner is paid according to the runner-up's cost parameter (see Laffont and Tirole, 1987, for details). Increasing the cutoff will then only affect the rent if $n = 1$. The rent increases by $\psi'(e(\hat{\beta}))$ per unit increase of the cutoff, it occurs with probability $P_1(\hat{\beta}) = NF(\hat{\beta})[1 - F(\hat{\beta})]^{N-1}$ and has weight λ in the principal's welfare function.

Cost minimization can be interpreted as the result of having a very high shadow cost of public funds, implying that $\lambda/(1+\lambda)\approx 1$. The remaining differences between Eq. (17) and Eq. (18) are due to distorted effort. Given $\hat{\beta}$, the left-hand side of Eq. (18) is smaller than the left-hand side of Eq. (17), the intuition being that the benefit of additional entry is smaller when the entrant's effort is not optimal. Similarly, the first term on the right-hand side of Eq. (18) is smaller than the first term on the right-hand side of Eq. (17), because distorted effort implies that less rent is given up to inframarginal firms as the cutoff increases.

It is not transparent from Eq. (17) and Eq. (18) how the optimal $\hat{\beta}$ changes when we introduce distortive mechanisms. However, implicit differentiation of Eq. (18), treating $\hat{e}=e(\hat{\beta})$ as an independent variable, shows that $d\hat{\beta}/d\hat{e}<0$ for $\hat{e}\in(e(\hat{\beta}),e^*)$, implying that the optimal cutoff is higher under distortive mechanisms than under fixed-price contracts. This suggests that the two rent-extracting devices (setting the cutoff level lower than one would have done if information were symmetric, and distorted effort) are substitutes.[13]

The following Proposition sums up the discussion of informed entry:

Proposition 4. *With informed entry, the optimal mechanism is characterized as follows:*

(i) among the entrants, the firm with the lowest cost parameter is selected;

(ii) the selected firm's effort satisfies $\psi'(e^i)=1-(\lambda/(1+\lambda))[F(\beta^i)/f(\beta^i)]\psi''(e^i)$;

(iii) any cutoff parameter $\hat{\beta}\in[\underline{\beta},\overline{\beta}]$ *can be implemented by setting the vector of entry-dependent reservation utilities* $(U_1(\hat{\beta}),\ldots,U_N(\hat{\beta}))$ *such that* $U_n^i(\hat{\beta})\geq -I$, $n=1,\ldots,N$, *and* $\sum_{n=1}^{N}\binom{N-1}{n-1}F(\hat{\beta})^{n-1}[1-F(\hat{\beta})]^{N-n}U_n^i(\hat{\beta})=0$, and

(iv) the optimal cutoff satisfies $f(\hat{\beta})[S-(1+\lambda)[\hat{\beta}-e(\hat{\beta})+\psi(e(\hat{\beta}))]=\lambda F(\hat{\beta})\psi'(e(\hat{\beta}))+\dfrac{(1+\lambda)If(\hat{\beta})}{[1-F(\hat{\beta})]^{N-1}}$.

In light of the preceding section the distortion result may be surprising. Although rent extraction is costly in terms of reduced entry, we find that effort should be distorted in the same way as in a model without entry. However, entry concerns are misleading here. Entry is an argument for leaving some rent to

[13] The first-best or symmetric information cutoff level can be found by setting $e(\hat{\beta})=e^*$ and cancelling the rent term of condition Eq. (15). The result reads $\beta_0-\hat{\beta}=I/[1-F(\hat{\beta})]^{N-1}$, which is identical to the condition for maximization of social surplus found in Samuelson (1985) (his Eq. (4)). (These should of course be identical, because symmetric information reduces our model to a fixed-price model.)

marginal firms (i.e., firms with $\beta^i = \hat{\beta}$), but it is no argument against extracting rent from inframarginal ones.

We conclude this section with a remark on implementation. While it is true that any given cutoff parameter can be implemented in many ways, some ways are more intuitive than others. If we restrict attention to mechanisms by which only the winner is paid, we have that $U_n^i(\hat{\beta}) = -I$ for $n \geqslant 2$ and $U_1^i(\hat{\beta}) = (I/[1 - F(\hat{\beta})]^{N-1}) - I$. This suggests the following rather simple scheme: the principal commits to pay any agent an entry subsidy of $(I/[1 - F(\hat{\beta})]^{N-1})$ whenever he is the sole bidder (in which case he will be the winner), and nothing if he is not. After entry, the principal implements Laffont and Tirole (1987) fixed-n mechanism. This latter part does not require any commitment.

4. Concluding remarks

In the introduction we asked whether allowing for entry in Laffont and Tirole (1987) procurement model would imply that the optimal mechanism changes radically or only modestly. We have seen that the answer depends on the information structure. First, if the potential suppliers have no private information at the time they make their entry decisions, distortive mechanisms do more harm than good, while plain auction mechanisms perform surprisingly well: A plain auction without entry fees is always optimal under mixed-strategy entry. Also with pure-strategy entry a plain auction is part of the optimal mechanism, but now regulating entry may be desirable.[14]

In contrast, if the potential suppliers know their costs at the time they make their entry decisions, distortive mechanisms are back in business: to extract information rent from those firms that have learnt that their costs will be low, the effort of firms with higher costs is optimally distorted, in exactly the way proposed by Laffont and Tirole (1987). However, the optimal mechanism is so constructed that it leaves positive gross rents to bidders of the worst possible type, implying that although the effort incentives part of Laffont and Tirole's mechanism carries over to the model with entry, the principal must commit to this particular deviation from Laffont and Tirole's mechanism.[15]

[14] This should be good news for procurers, as what we recommend is that they should continue to do what they have anyway done for decades now. Plain auctions have other advantages not studied here. For instance, plain auctions requires little information. To set up a plain auction the principal need not acquire cost information ex post, and, consequently, the firms need not waste resources on cost padding either.

[15] In fact, such commitment is crucial when prospective bidders are informed before entry: When agents have private information prior to entry, then entry is a signal of low costs, and this signaling can be shown to deter investment completely unless the principal before entry commits to a mechanism that protects the agent's rent (see Erbenová and Vagstad, 1999).

In both cases, implementation of the optimal mechanism requires commitment to a mechanism before the entry decisions are made. Such commitment is often difficult, because what is the optimal mechanism changes in the course of the game (the problem is sometimes referred to as 'time inconsistency', other times as the 'holdup' problem). Whether the solution to this problem is explicit or implicit contracts ('repetition' or 'reputation' concerns), it is crucial that the mechanism is easy to describe in advance, and that ex post violations are easy to detect. In the two cases studied in this paper neither of these should be a problem. With uninformed entry we have recommended the use of a plain auction. Plain auctions are easy to describe, and checking whether a particular auction is plain is also easy. With informed entry the optimal mechanism is far more complicated, but here we can exploit the implementation using an entry subsidy to sole bidders. The entry subsidy is just a number and by that easy to describe, and it should not be difficult to check whether there is only one bidder, either. No commitment to the remaining elements of the optimal mechanism is needed.

Our most general policy conclusion is that procurers should think twice before trying to improve upon a plain auction, making sure that the benefits of being 'smarter' exceed the costs, paying particular attention to the effect on entry.

Acknowledgements

Helpful comments and suggestions from Dag Morten Dalen, Nils-Henrik Mørch von der Fehr, Paul Klemperer, Kjell Erik Lommerud, Espen Moen, Tore Nilssen, Trond E. Olsen and two anonymous referees are gratefully acknowledged. Thanks also to seminar audiences at the University of Bergen, University of Bonn, Humboldt University, and the following conferences: 1994 EARIE (Chania), 1995 ESWC (Tokyo), Public Policy and Private Incentives (Bergen 1996), EEA 1996 and ESEM 1996 (both Istanbul). Financial support from the Norwegian Research Council and NFR-Ruhrgas is acknowledged. The usual disclaimer applies.

References

Baron, D., Myerson, R., 1982. Regulating a monopolist with unknown costs. Econometrica 50, 911–930.

Burguet, R., Sákovics, J., 1996. Reserve prices without commitment. Games and Economic Behavior 15, 149–164.

Dasgupta, S., Spulber, D.F., 1990. Managing procurement auctions. Information Economics and Policy 4, 5–29.

Demsetz, H., 1968. Why regulate utilities. Journal of Law and Economics 11, 55–65.

Erbenová, M., Vagstad, S., 1999. Investors facing opportunistic government: Is it really good to 'know the market' before investing? Scandinavian Journal of Economics, forthcoming.

Kjerstad, E., Vagstad, S., 1996. Procurement auctions with entry of bidders, Mimeo, Norwegian Research Center in Organization and Management, Bergen.

Laffont, J.-J., Tirole, J., 1987. Auctioning incentive contracts. Journal of Political Economy 95, 921–937.

Laffont, J.-J., Tirole, J., 1993. A theory of incentives in procurement and regulation, The MIT Press.

Levin, D., Smith, J.L., 1994. Equilibrium in auctions with entry. American Economic Review 84, 585–599.

McAfee, R.P., McMillan, J., 1987a. Competition for agency contracts, RAND Journal of Economics 18, 296–307.

McAfee, R.P., McMillan, J., 1987b. Auctions with entry, Economics Letters 23, 343–347.

McAfee, R.P., McMillan, J., 1988. Search mechanisms. Journal of Economic Theory 44, 99–123.

Piccione, M., Tan, G., 1996. Cost-reducing investment, optimal procurement and implementation by auctions, International Economic Review 37, 663–685.

Riordan, M.H., Sappington, D.E.M., 1987. Awarding monopoly franchises. American Economic Review 77, 375–387.

Samuelson, W.F., 1985. Competitive bidding with entry costs. Economics Letters 17, 53–57.

Tan, G., 1992. Entry and R&D in procurement contracting. Journal of Economic Theory 58, 41–60.

Vickrey, W., 1961. Counterspeculation, auctions, and competitive sealed tenders. Journal of Finance 16, 8–37.

World Bank, 1990. World Development Report, Oxford University Press.

Chapter 5

Auctions vs Negotiations – A Study of Price Differentials

Auctions vs Negotiations – A Study of Price Differentials

by
Egil Kjerstad*

ABSTRACT

Recent contributions in auction and bargaining theory suggest that a procurer should place more faith in the power of competition among alternative suppliers than in his or her own negotiating skill. Based on data from approximately 230 contracts between procurers and suppliers of medical and surgical articles, we test whether auctions and bargaining result in significantly different prices. We also test whether the market structure, i.e. the number of potential suppliers, depends on the particular trading procedure chosen and whether number of suppliers matters for prices. The main result is that auctions do not appear to result in significantly lower prices compared to negotiations.

Keywords: auction, bargaining, price differentials, endogenous market structure

* Foundation for Research in Economics and Business Administration (SNF) and Programme for Health Economics in Bergen (HEB). Thanks to Sturla Alvheim for helping me collect the data and thanks to Jan Erik Askildsen, Kjell Vaage, Steinar Vagstad and Christian Andersen for comments on earlier drafts. Thanks also to participants at the York Seminars in Health Econometrics and to participants at the Ninth European Workshop on Econometrics and Health Economics, Tinbergen Institute, Amsterdam 13. – 16. September 2000. Many thanks to Erik Schokkaert, Andrew Jones and to two anonymous referees for valuable comments on previous versions of the paper. This research is supported by the Research Council of Norway.

1. INTRODUCTION

Consumption of articles ranging from bandages via hip joints to pacemakers, amounts on average to around 14% of total gross expenditures at hospitals.[1] Although the dominant cost component in hospital care is labour costs, the effects of various ways of organising procurement should be of considerable interest for policymakers and administrators.

Variations in expenditures across hospitals can be attributed to quantitative effects reflecting different medical and surgical procedures or different degrees of specialisation. Quality differences may also influence the relative shares of total gross expenditures between hospitals. Last but not least, variations can be attributed to price differentials.

The prices to be paid for articles used in medical and surgical treatment are rarely posted.[2] Rather the demand for articles is publicly tendered and prices are determined after implementation of some form of auction mechanism or after a bargaining process between a procurer and one or more potential suppliers. The particular way products are traded may have impact on the degree of competition for contracts, measured as the number of suppliers participating in the auction or the bargaining process. Thus, prices per unit not only depend on the quantity and the quality of the goods, but may also depend on the trading procedure chosen.

Generally, the arrangement of procurement of goods and services varies. For instance, central government procurement in EU/EEA countries is often internationally tendered in compliance with EU directives, while local governments may use national or even local distributors or producers. We also observe that for the same type of goods and

[1] Norwegian data (NIS (1997)).

[2] See Arnold and Lippman (1998) for a theoretical discussion of the choice between posting a price and bargaining for the seller of an asset who is imperfectly informed about both buyer valuations and buyer bargaining abilities. See Bester (1993) for a discussion of the competitive market setting.

services central governments use auction mechanisms while local governments use negotiations.[3]

Differences in trading procedures can often be explained by differences in value of the tender. Central governments are obliged to use the international market place and auction mechanisms because orders are generally above the EU thresholds.[4] Local governments with purchases below the relevant EU thresholds can use the national or local market place and choose from a larger set of trading procedures. The essence of the European Union procurement regime is the insistence that major contracts be awarded according to specific procedures designed to ensure *openness*: active advertising and preference for open tendering; *equal treatment*: neutral specifications and objective award criteria; and *transparency*: a clearly defined set of rules, applied in a predictable manner and subject to public inspection.

The EU scheme does not explicitly state economy and efficiency in public procurement as an overall objective. Still, it is clear that the underlying premises are that the enforcement of openness, equal treatment and transparency leads to economic and efficient results compared to other trading procedures not in compliance with the underlying premises.

Bargaining represents an important alternative class of trading procedures to auctions or competitive bidding. Given that procurers with orders below EU thresholds can choose from a wider set of procedures, should they opt for negotiations or auctions? And for that matter, should procurers with orders exceeding the thresholds rather split the purchases into smaller orders substituting negotiations for auctions?

Based on the theoretical literature on auctions and bargaining, auctions have the upper hand. Milgrom (1989) argues that bargaining is a costly way to determine prices in a

---

[3] In Britain compulsory competitive tendering (CCT) is institutionalised by law to be the way local governments must organise their purchases of a wide variety of goods and services. As part of the ongoing restructuring of the public sector many other countries have introduced the use of auctions or other competitive mechanisms to expose (former) public monopolies and private "walk over" providers to competition.

[4] Generally ECU 200 000 for goods and services and ECU 500 000 for major capital work projects.

society where time is especially valuable. Milgrom concludes that bargaining is a trading institution that is best avoided when there is enough competition for auctions to be used. McAfee and McMillan (1996) argue that competition is a good substitute for bargaining skills and conclude that a procurer having a choice between negotiating with a sole supplier and organizing a bidding contest among several potential suppliers should do the latter. The bottom line is that to get a good price, a buyer should place more faith in the power of competition among alternative suppliers than in his or her own negotiating skill. Bulow and Klemperer (1996) conclude the same on the basis of far more general settings. Competitive bidding by suppliers will yield lower average prices than negotiating with a smaller number of suppliers. Thus, the value of negotiation skills is small relative to the value of additional competition given by auction.

Milgrom (1989), McAfee and McMillan (1996) and Bulow and Klemperer (1996) build on the assumption that the number of bidders is given. They do not explicitly deal with how the choice of trading procedure may effect the number of possible suppliers for a given contract and thereby also the expected price to be paid. Rather than assuming that the number of bidders is exogenous and constant, French and McCormick (1984), Samuelson (1985), McAfee and McMillan (1987a), Tan (1992), Engelbrecht-Wiggans (1993), Levin and Smith (1994) and Kjerstad and Vagstad (2000) discuss the importance of auction design for the number of bidders and ultimately the expected price to be paid by the procurer. According to this literature price and market design should not be analysed as separate issues but rather treated in a way that handles the (possible) linkage between trading procedures, number of bidders and prices.

We take as a starting point that the (expected) prices depend on how competitive a particular trade turns out to be and that the initial choice of trading procedure determines the number of suppliers actually participating in the first place, i.e. the market structure is endogenous. Using the price to be paid by the procurer as the crucial outcome variable, we test the hypothesis that auctions result in significantly lower prices compared to negotiated prices, applying data from markets for medical and surgical articles. Of course, this study does not represent a final judgement of the pros and cons of auction and bargaining. Hopefully it represents a step in the direction of testing empirically the recommendations that can be drawn from the theoretical literature.

Surprisingly little empirical work is done comparing the relative merits of auctions versus negotiations. Lack of comparable data is the most likely reason for this deficit. On empirical grounds at least, one should be cautious a priori stating that a procurer opting for negotiations will always have a thinner market than a colleague opting for auctions, and that bargaining outcomes (e.g. prices) are always dominated by the outcomes from some form of bidding mechanism.

Our analysis builds on information obtained from purchasers of medical and surgical articles in Norway. We have collected data on approximately 230 contracts made between procurers and suppliers during the period 1997 to 1999. Most of the articles are used on a daily basis and in relatively large quantities during the contract period. Our data provide information about price (per unit prices), the size of the purchase, the duration of the contracts, the date of the agreement and whether the purchase was organised centrally by the hospital owner or decentralised at hospital level. We also know whether the purchase was tendered in the EU or negotiated, and we know how many suppliers that took part in the auctions and how many suppliers that were part of the bargaining process. First-price sealed-bid auctions are the basic auction mechanism applied for the products studied here. The bargaining contracts on the other hand basically wind up a two-step process in which samples of products are tried, quality learned and prices are then negotiated.

We find evidence suggesting that compared to negotiations, auctions lead to significantly thicker markets in terms of suppliers but not to significantly lower prices.

This paper is organised as follows. In the next section price determination and market design are briefly discussed. In section 3 a short presentation of the survey follows, and the data on which the paper builds are presented. The econometric models are presented in section 4. Section 5 summarises the main results. Concluding remarks are gathered in section 6.

2. PRICE DETERMINATION AND MARKET DESIGN

Two broad categories of trading procedures are studied here.[5] Auctions are markets with explicit trading rules that specify precisely how market clearing determines prices. Auctions are primarily a response to asymmetric information problems. A public procurer may not know the suppliers' reservation prices and uses an auction to extract information. Auctions may be useful for agency reasons too. Dishonest dealings or favouritism can be more costly when a contract is put up for auction. Auctions may be designed in a multitude of ways but different auction mechanisms can, under certain circumstances, produce the same expected revenue for the seller or the same expected price to be paid by a procurer (Vickrey (1961)). Auctions are competitive by nature but the number of participants in the bidding process may be endogenous as discussed below.

Bargaining is an activity that can be associated with many trading situations. Some examples are bargaining between employer and employees over wages, between a consumer and a retailer over prices or between wholesalers and large consumers or consumer groups, as is the case in this paper. Furthermore, bargaining is commonly modelled as a bilateral monopoly, i.e. all negotiations take place between two agents. This is convenient for modelling purposes but not necessarily a realistic assumption in many cases. In the literature bargaining is often modelled with complete information, i.e. seller and buyer have the same information about all relevant aspects of the trade. A common result in such models is that the more patient part in the bargaining game reaps the largest benefits from trade and that trade takes place instantaneously. There is also a literature where it is assumed that the seller may have incomplete information about the buyer's willingness to pay or that the buyer has incomplete information about the

[5] See for instance Rasmusen (1994) for a short discussion of auction mechanism and bargaining problems. There exists a huge literature on both topics. Thorough surveys are Milgrom (1987), McAfee and McMillan (1987b), Kennan and Wilson (1993), and Klemperer (1999).

seller's reservation price (or the asymmetry may be two-sided). Fudenberg and Tirole (1991) and Osborne and Rubinstein (1990) give a presentation and discussion of several such models. The course of the negotiation is basically influenced by two considerations: the agents' impatience and the agents' opportunities for trade with other partners.

Ståhl (1972) and Rubinstein (1982) (the Rubinstein-Ståhl bargaining model) found that sequential bargaining under complete information yields a unique, Pareto-efficient outcome in which the parties reach an agreement without delay. Bargaining power is determined by the players' discount rates, i.e. players who are more patient do better. Osborne and Rubinstein (1990) discuss the relationship between bargaining theory and the operation of markets. They cover several model specifications. Among them are complete information models allowing for sequential bargaining with more than one party (the analysis is limited to the three party case). Using as benchmark the bilateral bargaining game of alternating offers in which the procurer makes the first offer, Osborne and Rubinstein show that the price to be paid may be lower when allowing for more suppliers. Still, the extent to which the procurer can benefit from the existence of an additional supplier depends on how credible the procurer's threat is to abandon a supplier, which again depends on the supplier's (known) reservation prices.

Inefficiencies tend to be introduced when modelling sequential bargaining games with incomplete information. When bargaining is inefficient, the choice of economic institutions, i.e. the rules of the game, can influence the efficiency of the outcome. Another feature is that incomplete information bargaining models tend to result in many equilibria (see for instance Fudenberg and Tirole (1991) for a discussion), i.e. different prices can be achieved.

One-sided-offer bargaining games give strong results, though. In these games the seller's production or opportunity cost is known, while the buyer's valuation of the good is private information. Furthermore, the seller makes the offers and the buyer responds by yes or no in each period, i.e. no price offer is given in return, only acceptance or rejection.

To illustrate the outcome of a game of one-sided asymmetric information, we adopt Cresta's (1991) exposition of Fudenberg and Tirole (1983). The game may not be a very realistic description of the bargaining process between suppliers and consumers of surgical and medical articles, but it shows that it is difficult a priori to argue that bargaining always results in relatively low or high prices, even in cases where the number of equilibria is limited.

Model:

A buyer and a seller bargain over the price of a non-divisible item. Both are risk-neutral. The value of the good (in money units) is s for the seller and b for the buyer. The seller does not know with certainty the buyer's valuation. Assume that the buyer can be one out of two types $\underline{b}$ and $\bar{b}$, with $\underline{b} < \bar{b}$. We furthermore assume that s < $\underline{b} < \bar{b}$. The seller has a prior probability distribution over $\{\underline{b}, \bar{b}\}$ and we assume this to be {1/2, 1/2}. Bargaining takes place in two stages. At the beginning of the first stage, the seller announces a price, which the buyer either accepts or rejects. If the buyer refuses the offer, the seller renews his offer in the second period. The bargaining process ends with the second period.

If an agreement is reached at price p_1 at the end of period 1, the players' payoffs are

Seller: $p_1 - s$

Buyer: $b - p_1$

If an agreement is reached with price p_2 at the end of period 2, the payoffs are:

Seller: $\delta^S(p_2 - s)$

Buyer: $\delta^B(b - p_2)$

where δ^S and δ^B are discount factors. Otherwise, the payoffs are $(\delta^S - 1)\,s$ for the seller and 0 for the buyer.

Restricting the discussion to the two cases where parameter values give (i) $\underline{b} > (\bar{b} + s)/2$ and (ii) $\underline{b} < (\bar{b} + s)/2$. Case (i) means that the buyer's lowest valuation ($\underline{b}$) is higher than the seller's second-period expected profit, estimated at the beginning of the second period. Case (ii) represents the opposite situation. Assume that the product is valueless after period 2.

Cresta (1991) shows for case (i) that two perfect Bayesian equilibria can be realized.

(1) The seller plays the minimum price in each period; the buyer regardless of his type accepts that price in the first period.

(2) The seller proposes an intermediate price b^i in the first period and the minimum price $\underline{b}$ in the second.[6]

For case (ii), depending on the parameter values, the following hold in equilibrium:

(1) The seller plays the minimum price and the buyer buys the item in the first period regardless of his type.

(2) The seller announces an intermediate price in the first period, which is accepted only by a type $\bar{b}$ buyer, and the minimum price $\underline{b}$ at the second period, which is always accepted.

(3) The seller proposes the maximum price in each period and only a type $\bar{b}$ buyer purchases the item; there is some probability that he will buy, depending on discount rates, in the first period, and always buys in the second period if he refused in the first period.

The results show how the information of the players can increase in a bargaining process when the players act rationally. The increase in information is reflected in the Bayesian revision of the prior probabilities. In contrast to the case of complete information, bargaining does not necessarily stop at the first stage of the game. The solution is therefore not necessarily Pareto optimal (first best) since the discount factors are strictly less than 1 and positive. On the other hand, the results show that it is not necessarily self-evident that bargaining always will result in (relatively) high prices.

---

[6] The intermediate price b^i is the highest first-period price accepted by a type $\bar{b}$ buyer knowing that the second-priod price offer will be $p_2 = \underline{b}$.

Imagine that the game described above captures the bargaining game between a matched pair of a supplier and a buyer. Imagine that we have several such matched pairs bargaining at the same time and that no information is transmitted between pairs of players.[7] Price variations can then be explained by different production costs (heterogeneity among suppliers), different valuations of the good (heterogeneity among buyers) and different discount factors across players.

Continuing the discussion of the impact of incomplete information on trade outcomes, let us next turn to auctions. To be specific, let us assume that the auction is of the first-price sealed bid type with N risk neutral bidders with values of the item for sale independently drawn from a uniform density from 0 to some amount v*. Denote player i's value by v_i. Let us study player 1. It can be shown that[8] the optimal bid for player 1 is E(v) = $\frac{(N-1)v_1}{N}$. Player 1 should bid a fraction $\frac{(N-1)}{N}$ of his value minus ε. As the number of competitors increases, the optimal bid for player 1 moves closer to his value. Based on this simple model, the recommendation is clear-cut. The more bidders that participate in the auction, the higher is the expected revenue from the auction (or the lower is the price to be paid).

Note that the bidder with the highest value will be the winner regardless of number of bidders, i.e. the outcome is Pareto optimal. In the case of a buyer faced with N sellers, the same result will prevail. The fixed-price auction induces first-best cost reducing effort.

Laffont and Tirole (1987), Riordan and Sappington (1987) and McAfee and McMillan (1987a) analyse optimal auction design in cases where there are asymmetric information between the principal and the agent and there are costs associated with raising public funds to finance procurement. For instance, Laffont and Tirole (1987) show that

[7] This is not an unrealistic description. The impression is that few procurers know much about what other hospitals or groups of hospitals achieve in terms of prices or other terms in the contracts. The main reason for this is probably that they lack incentives for doing so. They are paid on a fixed salary basis and the owners of the hospitals have a long tradition of non-interference in procurement and logistic decisions made locally.

[8] The result is not a general rule. See Rasmusen (1994) for a discussion.

Governments can do better than securing first-best effort. Instead of using fixed-price contracts, they propose the auctioning of incentive contracts. The cost inefficiencies induced by incentive contracts are more than compensated for by the extraction of information rent. Thus, incentive contracts can be termed rent extracting mechanisms or rent extracting trading procedures.

In the work of Laffont and Tirole (1987), Riordan and Sappington (1987) and McAfee and McMillan (1987a) it is assumed that the number of bidders is fixed or given exogenously. In the literature on auctions with entry, on the other hand, it is assumed that each potential supplier must sink a relation specific entry investment before he can participate in the procurement auction. The entry investment may take different forms. Expenses incurred establishing an organisation that is necessary in order to be taken seriously as a contender for the project; bid preparation costs such as costs associated with calculating cost figures for the project; capacity and opportunity costs in the sense that the supplier may have to decline other customer's demands while preparing delivery, are some examples. Equilibrium entry requires that each potential firm enters if and only if the expected profit is large enough to cover entry cost, i.e. the number of bidders is endogenous.

Kjerstad and Vagstad (2000), using a model with an endogenous number of bidders, ask whether auctioning of incentive contracts instead of using fixed-price contracts is necessarily a wise recommendation to give a public procurer. They suggest that rent-extracting mechanisms reduce the firm's expected rent and thereby reduce their incentives to enter, possibly resulting in higher expected costs if the number of firms is reduced. Their study illustrates the point that the choice of trading procedure or market design may have significant effects on the market structure and expected prices.[9]

[9] Kjerstad and Vagstad (2000) conclude that procurers should think twice before trying to improve upon a plain auction, making sure that the benefits of being "smarter" exceed the costs, paying particular attention to the effects on entry. A plain auction is recognized by i) firms bid for fixed-price contracts and ii) bidder i is selected if his bid is the lowest and not exceeding the procurer's maximum willingness to pay. Examples of plain auctions are the English, Dutch, first-price sealed bid and second-price sealed bid (Vickrey) auctions.

To sum up, based on the theoretical literature on auction it is evident that number of bidders plays a role, or to put it differently, that market design matters for the price to be paid. A procurer may expect to pay less the thicker the market is. This result is particularly clear in the literature on auction with entry. It suggests that the particular design of a trading procedure may influence the number of bidders that want to participate in the first place. It is this latter line of thought that is at the core of this study. A bargaining theory with incomplete information and with the possibility of an endogenous number of trading partners has not yet been developed in the same stringent framework as what auction theory can offer. Still, the basic point in this analysis is that we want to test whether trading procedures influence the "thickness" of a market and whether market structure matters for prices.

Thus, we ask three main questions in this study:

- Do auctions involve more potential suppliers compared to negotiations?
- Does the number of suppliers (or the endogenous market structure) matter for prices?
- Do auctions result in lower prices to be paid by the public procurer compared to negotiations?

One should keep in mind that treating auctions and bargaining as two widely different trading procedures might be dubious. In his review of game-theoretic analyses of trading processes, Wilson (1987) proposes and illustrates that the way to interpret complex markets is to see them as a hierarchy of embedded bargaining and auction games with endogenous processes of signalling and competitive pressure. Often, as Milgrom (1989) points out, it is necessary to combine bargaining with bidding to support efficient trade. In many instances it would be foolish to invite bids from all comers and just take the lowest bid. Before the final bids can be made, there may be rounds of negotiations over the specifications and then another round to determine which of the potential bidders are qualified to produce a product that meets specifications. Then, the evaluation of the bid may take into account design or quality differences, service capabilities, the ability to deliver on time, and perhaps the need to maintain multiple sources of supply, as well as price.

Another dynamic issue is the challenges of administering post-auction contracts on behalf of several hospitals, as can be the case in this study. The procurer initially writes a contract on price and quality with the winning supplier. The supplier then locally introduces samples of products of a different quality or with other specifications and (most likely) at a higher price level than the prices specified in the winning contract. The local units may prefer these product variants and start ordering them. The auction can as such be used by suppliers as a mean of creating an entry to new local markets and the monopoly position ex post can be used to manipulate the local purchasers both in terms of quality and prices. Of course, such tactics should be easy to detect if the procurer actually pays the bills. However, this is not always the case. For instance, for the units in this study, the expenses are usually incurred and paid for by the hospitals themselves and locally they may not have the incentives to benchmark the price they pay against the winning contract. Locally, they even get the quality they want.

The bargaining contracts in our sample are basically a result of a process in which the procurer invites a number of suppliers to participate in the bargaining process. Keep in mind that in the same way as suppliers may choose not to participate in an advertised competitive bidding process organised for instance as a first-price sealed bid auction, not all of the suppliers invited to a bargaining process necessarily participate in the negotiations. Perhaps the procurer is too small to make the bargaining process interesting compared to bargaining processes offered by other procurers. Still, one may expect that it is more likely that auctions are open to potentially more suppliers compared to negotiations. On the other hand, on theoretical grounds it does not necessarily follow that negotiations always will have fewer participants compared to auctions.

In the bargaining processes studied here, the procurer may demand samples of products and the products are subsequently distributed to hospitals for testing. Users' verdicts are collected and their experiences discussed in so-called user-boards. The procurer then uses this information to short-list suppliers. Thus, at the next stage of the game quality is known but the reservation prices to suppliers and procurers are private information. These features of the process are not necessarily different from what a procurer could do as part of the preparation for a first-price sealed bid auction. In our case, where the basket consists of common and much-used products, such parity between auctions and negotiations is not a strong assumption. Rather, in this study the main difference between auctions and negotiations is the way the prices are determined. In the case of negotiations, prices are settled after bargaining between the procurer and the short-listed suppliers while in the auctions prices are determined according to the strategy the suppliers choose in response to the first-price sealed bid auction.

3. DATA

This study is based on data collected from hospitals and counties (i.e. the hospital owners) throughout Norway during a period of four months in 2000 asking for information about a set of items much used at hospitals (see Table A1 in the appendix). We ran a pilot survey during the autumn of 1999 and the experiences we made led us to change the composition and number of items 'in the basket', and to make changes in the formulation of some of the questions. We also limited the number of questions asked (see Table A1 in the appendix for a copy of the final questionnaire). The items asked for in the pilot survey turned out to have either a large hospital bias or to be too vaguely described making it difficult for the respondents to provide consistent information.

The final questionnaire was sent to officials responsible for procurement of articles used in medical and surgical treatment. The final survey covered 14 different familiar products at most hospitals giving us a total number of contracts of 216.[10] After considerable follow-up effort only two procurement units (representing approximately 28 contracts) out of 24 procurers did not answer the query.

[10] We received questionnaires covering 249 contracts. Due to missing data and concern about the quality of some data observation, the sample was reduced to 229. Suggestions that the large price differentials concerning product 8, central line catheter, must reflect different product variants, we decided to drop these observations altogether, reducing the sample to 216 observations.

We received information on date of completion of contracts making it possible to compare contracts in terms of year 2000 prices. The variable age of contract is used to control for time effect on prices. We also learned the duration of the contracts. The size of the order in terms of units is known although we have some reason to believe that the accuracy of the figures could have been better for some of the contracts. The basic problem seems to be that neither the central procurer nor the hospitals covered by the same contract knows exactly how much that is used during say a year. This has led to frustration among suppliers. They argue that the specified quantities in the contracts are not met.

Furthermore, we collected data showing whether the contracts is made on the basis of auction or negotiations and whether procurement is centralised to county level or decentralised to the hospitals. It turned out that the majority of the counties had centralised procurement functions. This means for instance that a county with five hospitals is represented in our data with one contract per product.

Approximately 15% of the 216 collected contracts, were based on bargaining outcomes and 85% were based on auction in the EU/EEA area. Not all of the procurers used the whole basket of goods about which we had chosen to question them. The number of contracts per product varies from 11 to 20.

Price comparisons are challenging for several reasons. We are in particular aware of the importance of quality differentials. Care was taken asking for data on well-specified products (close description of the product and posting producer's name). Still, we learned that for some products other producers or slightly different products were used. Thus, the price data (per unit prices) we received may reflect quality differences not captured by the survey. For some products prices, varies relatively much. Central line catheter (product 8) have price differentials of approximately 400 percent between minimum and maximum prices (see table A4 in the appendix) indicating that our product specification may not be precise enough or that procurers have given data concerning a different types of catheter. Product 8 was dropped from the study. However, based on interviews made as part of the data collection and follow-up work in connection with the pilot study, some of our informants were not surprised to learn about relatively large price differentials. Some argued that price differentials have been declining over time.

Another point is that if two hospitals have chosen to employ different products in the same medical treatment, price differentials basically mean that some hospitals are using more expensive products even if cheaper and quite as useful products are available. On the other hand, changing from one product to another even when they are very close in terms of attributes may cause additional costs in terms of change of medical routines and necessary instructions of medical personnel. Organisational issues like these can create lock-in effects. Normally contracts last for 24 months with an option of prolonging it by an additional 36 months. If a product is on the shelves, the purchaser may then have strong incentives to keep them there for a longer period than the initial contract states. Knowing this, suppliers could possibly bid harder for the contracts. Most of our products are not of such a character, though.

Last but not least, we collected data on the number of suppliers that had taken part in the auction or negotiation. Tables A3 and A4 in the appendix give summary statistics of the main variables.

4. ECONOMETRIC MODELS

To answer whether auctions involve more potential suppliers compared to negotiations; whether market structure matters for prices and whether auctions result in lower prices compared to negotiations, we need first to address some methodological issues.

Here, the aim is to test the hypotheses drawn from the literature on auctions and negotiations against the data available. The analyses are constrained by the limited number of contracts (observations) available per product. The number of contracts per product varies between 11 to 20, leaving product wise regressions an unattractive approach. Pooling of data is the natural way to proceed but we should still be aware of possible pitfalls. The dummy variable approach (a dummy variable per product) implies that the intercepts/fixed effects may vary across products but not the slopes. Predicting prices in a single equation model using product dummies (and other explanatory variables) implies that we assume the price flexibility to be homogenous across products. We should keep this in mind when interpreting our results.

Estimating a price equation using number of suppliers as one of the exogenous (explanatory) variables should be critically examined. There is, as discussed in section 2, ample support in the literature for treating the market structure as endogenous in the sense that choice of trading procedure influences the number of potential suppliers.

Assuming that there is a simultaneous relationship between prices and number of suppliers, OLS is inappropriate for the estimation of an equation in a system of simultaneous equations (here a system of two equations: a supplier equation and a price equation). On the other hand, OLS can be applied appropriately even in the context of a simultaneous equation model if the model is a recursive one. In other words, if the number of suppliers participating in an auction or in a bargaining process is the cause, and the dependent variable, the price, is the effect. In a recursive system, OLS can be applied to each equation separately. Assuming that the trading process is a one shot game and that neither procurers nor suppliers have any memory of former trading experience to update their beliefs, the recursive approach is applicable. We could also argue that the recursive model is appropriate in cases where yesterday's contracts have no bearing on today's contracts. Let this approach be termed Model 1.

By emphasising the points made in the literature on auctions with entry, the residuals in the supplier and price equations cannot be expected to be uncorrelated and we need another estimation approach. If procurer and suppliers, based on former experiences, have updated prior beliefs attached to the expected traded price, the IV/2SLS approach is a possible specification. We term this approach Model 2. More specifically, assume that potential suppliers of the product learn yesterday's winning price. In such a situation prices and number of suppliers may well be treated as interdependent variables: Yesterday's price influence today's expected reservation price, which again influence the number of suppliers that will participate. The number of participants determines the degree of competition and finally (expected) price. If this is the more realistic description of procurement at hospitals and centrally at county level, the residuals in the supplier and price equation cannot be expected to be uncorrelated. Thus, estimating each equation with ordinary least squares will give inconsistent estimates.

The main problem with this last trading process - from an empirical point of view - is that we do not have time series in the sense that we can track the prices paid by the same procurer over time for the same products. We do have information about (winning) prices paid over time for the same product, though. Given that counties or hospitals know these prices, we can model prices and number of suppliers as interdependent variables with greater confidence.

A possible critique against Model 1 (OLS) and Model 2 (IV/2SLS) is that the choice of trading procedure is endogenous too. Based on supply side and/or demand side information (experience from bargaining processes, former prices achieved through either trading procedures, former number of bidders, etc.), a procurer may decide to use a particular trading procedure for a particular product. One advantage of the IV/2SLS regression is that it is possible to estimate a single equation of a multiple-equation system without specifying the functional form of the remaining equations. Thus, the price equation is estimated treating both the choice of trading procedure (METHOD) and the number of suppliers (SUP) as endogenous variables (Model 3). Method and suppliers, as will be shown below, are instrumented in the same way as number of suppliers is modelled in Model 2.

The econometric specification of the models presented above needs to be described in more detail. The following variables will be used:

$PRICE_{ij}$ is a dependent variable measuring the per unit price (in 1999 prices).
SUP_{ij} measures the number of suppliers participating in the contract specific auction/negotiation.
$METHOD_{ij}$ is a dummy variable with value 1 if auction and value 0 if negotiation.

ORG_{ij} is a dummy variable with value 1 if the purchase is organized centrally and with value 0 if the purchase is organized at hospital level.

$ORDER_{ij}$ denotes the size of the contract measured in product units.

$UNITS_{ij}$ denotes number of hospitals covered by the contract.

$DURAT_{ij}$ is the duration of the contract, measured in months.

AGE_{ij} is the time difference measured in months between the year 2000 and the year the contract was written.

$METSUP_{ij}$ is the interaction term between method and number of suppliers.

PRODj are dummy variables for j-1 products, i.e. 12 product dummies.

Variables marked $\hat{}$ are estimated explanatory variables.

υ, η, ε, v and ρ are disturbance terms.

The indices i and j denote procurer (i = 1,..,20) and product (j = 1,..,13) respectively. Thus, the subscript ij is related to a specific contract made by procurer i for the supply of good j.

Model 1: OLS regressions of supplier and price equation

$$SUP_{ij} = \alpha_0 + \alpha_1 METHOD_{ij} + \alpha_2 ORDER_{ij} + \alpha_3 AGE_{ij} + \alpha_4 DURAT_{ij} + \alpha_5 ORG_{ij} + \alpha_6 UNITS_{ij} + \sum_{7}^{19} \alpha_j PROD_j + \eta$$

$$PRICE_{ij} = \beta_0 + \beta_1 SUP_{ij} + \beta_2 METHOD_{ij} + \beta_3 METSUP_{ij} + \sum_{4}^{14} \beta_j PROD_{ij} + \upsilon$$

Model 1 is based on a two-equation specification but where the model is a recursive one, i.e. it is assumed that there is an unidirectional cause-and-effect between suppliers and prices, and that υ and η are uncorrelated disturbances. Under these circumstances both equations can be consistently estimated by ordinary least squares.

The model may help us to answer (i) whether auctions give more suppliers than negotiations; (ii) whether market structure matters for prices and (iii) whether auctions result in lower prices to be paid by the public procurer compared to negotiations. The first question can be answered through the sign and size of the estimated coefficient attached to METHOD. The second question can be determined through the sign of the variable SUP. The latter question is answered by the estimated coefficients for variable METHOD and the interaction effect between choice of method and number of suppliers, METSUP. The interaction term is useful since it allows us to capture whether auctions and negotiations may have a separate effect on prices also through separate effects of number of suppliers that participate.

In the second model, an instrumental variable estimator is used. Here we assume that ε and ν are correlated. We regress SUP - the endogenous variable - on all the exogenous variables in the system. The estimated values of number of suppliers ($\hat{SUP}_{ij}$) are then used as best instrument in the price equation.

Model 2: IV/2SLS regression, endogenous sup and metsup

$$SUP_{ij} = \lambda_0 + \lambda_1 METHOD_{ij} + \lambda_2 ORDER_{ij} + \lambda_3 AGE_{ij} + \lambda_4 DURAT_{ij} + \lambda_5 ORG_{ij} + \lambda_6 UNITS_{ij} + \sum_{7}^{18} \lambda_j PROD_j + \varepsilon$$

$$PRICE_{ij} = \sigma_0 + \sigma_1 \hat{SUP}_{ij} + \sigma_2 METHOD_{ij} + \sigma_3 \hat{METSUP}_{ij} + \sum_{4}^{15} \sigma_j PROD_j + \nu$$

Finally, assuming that the choice of trading procedure is endogenous too, the price equation in model 3 is given by:

Model 3: IV/2SLS regression, endogenous sup, metsup and method

$$PRICE_{ij} = \pi_0 + \pi_1 \hat{SUP}_{ij} + \pi_2 \hat{METHOD}_{ij} + \pi_3 \hat{METSUP}_{ij} + \sum_{4}^{15} \pi_j PROD_j + \rho$$

where both number of suppliers and choice of method are estimated with the same model, using *ORDER, AGE, DURAT, ORG, UNITS* and *PROD*-dummies as instruments.

5. RESULTS

A log-linear specification (log of dependent variable, here prices) reduces heteroscedasticity, and it turns out here that it also serves the purpose of making the price equations in Model 1 and Model 2 pass the Ramsey test. In Table 1, the results for the price equations in Model 1 and Model 2 with a log transformation of the dependent variable are reported.[11]

Based on the OLS estimation of the supplier equation, Table 1 below shows that the market structure depends significantly on whether auctions or negotiations are used. Auctions appear to give a thicker market compared to negotiations. Furthermore, comparatively new contracts involve significantly fewer suppliers compared to older contracts. Duration of the contracts has no significant bearing on the number of suppliers. On the other hand, whether the procurement is organised centrally or decentralised to hospital level seems to matter. Decentralised trade attracts more suppliers than centralised. Likewise does number of hospitals covered by the contract: the more units, the more suppliers.

The supplier-equation does not pass the Cook-Weisberg test for heteroscedasticity, though, but does pass the Ramsey RESET test for omitted variables. White-corrected standard errors are used in the presence of heteroscedasticity but as Table 1 shows, without changing the main results: For the supplier equation, the t-values in the non-robust and the robust case are close to each other in value.

Let us first address the question of what is the effect on prices of the choice of trading procedure (METHOD). As pointed out in the introduction, some authors advocate using auctions arguing that the (expected) price is lower that way.

[11] As mentioned earlier, the specification of the price equation is somewhat unsatisfactory given that all products will have the same price flexibilities (the same slope but different intercepts). This assumption can be critized for not capturing supply and demand differences between at least some of the products.

Table 1. Estimation results

	Model 1 Supplier equation Coef.	t-value	t-value[1]	Model 1 Price equation Coef.	t-value[1]	Model 2 Price equation Coef.	t-value[1]
Sup				-.0302821	-0.601	.2883821	0.271
Metsup				.0415557	0.857	-.1494646	-0.148
Method	.5562954	1.762 *	2.058 **	-.2407769	-1.152	.5190162	0.118
Order	-2.79e-07	-0.568	-0.543				
Units	.1967825	2.592 ***	2.130 **				
Age	.0238965	2.503 ***	2.325 **				
Durat	-.0097664	-0.650	-0.703				
Org	-.8443458	-2.954 ***	-3.117 ***				
prod_2	.7277053	1.270	1.030	-4.248223	-29.703 ***	-4.322601	-22.344 ***
prod_3	-2.10313	-4.090 ***	-4.322 ***	-1.472111	-15.004 ***	-1.159266	-4.098 ***
prod_4	1.924802	3.454 ***	2.660 ***	-3.409963	-42.730 ***	-3.67599	-9.042 ***
prod_5	.2992121	0.610	0.508	-3.993724	-62.377 ***	-4.028498	-30.608 ***
prod_6	-2.023306	-3.978 ***	-3.951 ***	-.4966163	-4.948 ***	-.1883423	-0.540
prod_7	.6430767	1.311	1.159	-2.764367	-23.730 ***	-2.837372	-17.588 ***
prod_9	.0846252	0.178	0.166	1.011695	13.610 ***	1.009616	8.420 ***
prod_10	-.5956969	-1.196	-1.036	-1.765611	-11.951 ***	-1.654181	-9.530 ***
prod_11	.0381256	0.078	0.061	-2.10269	-21.948 ***	-2.080513	-15.924 ***
prod_12	-2.157991	-4.303 ***	-4.166 ***	.0015112	0.018	.305068	1.113
prod_13	-.9674413	-1.945 **	-1.873 *	-.6474542	-9.668 ***	-.5057374	-3.583 ***
prod_14	-.9949508	-1.906 **	-1.957 **	-4.152184	-53.534 ***	-3.98886	-20.679 ***
Constant	4.332832	7.463 ***	6.407 ***	2.645865	11.601 ***	1.211255	0.257
N =216							
F(18,197) =		8.80	10.95				
F(16,212)=					1092,65		710,22
R^2 =	0,45			0,97		0,96	

[1] Based on a robust variance estimator.
***Significant at 1%. ** Significant at 5%. * Significant at 10%.

We do not find such an effect here. It is also interesting to note that number of suppliers does not matter for prices. As shown in Table 1, these results prevail when estimating the price equation using the IV/2SLS estimator. Neither of the key variables are significant. Taken together, the IV/2SLS results are surprising for at least two reasons:

- Prices appear not to be explained by differences in market structure/number of suppliers competing for a contract.
- Prices appear not to depend on the chosen trading procedure, auctions or negotiations.

The latter point is interesting in terms of the theoretical discussion of the pros and cons of auctions compared to negotiations, and in which much research gives credit to auctions. The first point is interesting in terms of the discussion on auctions with entry. The main argument here is that the rules governing a particular trade (auction) may influence the number of bidders, which again influences the (expected) price. In model 2 we find that number of suppliers does not have a significant bearing on prices. Thus, choice of trading procedure appears to not matter for prices.[12]

Which one of the models, Model 1 or Model 2, is the better one? We have performed the Hausman test (although a large sample test) to check inconsistency. Is there sufficient difference between the coefficients of the IV/2SLS regressions and the OLS to indicate that OLS would be less efficient, using the robust estimator? The results reported in Table 2 show that the null hypothesis – that the estimated coefficients do not systematically differ – cannot be rejected. Thus, we find support for the argument that number of potential suppliers in a trade does not have to be treated as endogenous. Note though that the result that the estimated coefficients do not systematically differ may very well be driven by the comparatively small differences of the dummy-coefficients. Keeping this in mind, Model 1 with White-corrected standard errors appears to be the most attractive model specification based on the Hausman test.

[12] The underlying supplier equation is the same in both Model 1 and Model 2. Thus, choice of trading procedure matter for markets structure but market structure itself does not have any impact on prices in either of the models.

Table 2. Hausman test Model 1 vs Model 2

	(b) Prior	(B) Current	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
sup	.2883821	-.0302821	.3186642	1.06332
metsup	-.1494646	.0415557	-.1910204	1.007371
method	.5190162	-.2407769	.7597931	4.406032
prod_2	-4.322601	-4.248223	-.0743787	.1302718
prod_3	-1.159266	-1.472111	.3128445	.2653603
prod_4	-3.67599	-3.409963	-.2660268	.3986217
prod_5	-4.028498	-3.993724	-.0347742	.1149959
prod_6	-.1883423	-.4966163	.3082741	.3342462
prod_7	-2.837372	-2.764367	-.0730049	.1116018
prod_9	1.009616	1.011695	-.0020791	.0940852
prod_10	-1.654181	-1.765611	.1114307	.0911203
prod_11	-2.080513	-2.10269	.0221775	.0888317
prod_12	.305068	.0015112	.3035568	.2612569
prod_13	-.5057374	-.6474542	.1417168	.1242488
prod_14	-3.98886	-4.152184	.1633233	.1766093

b = less efficient estimates obtained previously from ivreg.
B = more efficient estimates obtained from regress.
Test: Ho: difference in coefficients not systematic
chi2(15) = (b-B)'[(V_b-V_B)^(-1)](b-B) = 2.51
Prob>chi2 = 0.9999

Turning next to Model 3, recall that the main aim with this model is to allow for the choice of trading procedure to be endogenous, too. The basic idea is to try to capture the effect of the fact that procurers may choose to use negotiations or auctions depending on for instance the product to be purchased. Thus, treating METHOD as an endogenous variable gives the possibility to capture that the choice of method is in some way influenced by earlier experiences.

Note that all contracts in our sample have a value less than the EU threshold of ECU 200 000, meaning that procurers in principal are free to choose which trading procedure to employ, competitive bidding in the EU/EEA or negotiations. In practice, though, auctions in the EU/EEA are often used. Firstly, because multi-product contracts (more than one product traded at the same time) or the length of the contract may bring the total value above the threshold value, and, secondly, it is perceived to be good procurement policy to use EU/EEA regardless of the value of the order. On the other hand, quality dimensions like colour and material (as can be of importance for the users of hats, helmets, mask etc) have no bearing on the choice of trading procedure, according to procurement official we have spoken to.

Those opting for negotiations argue that they have a more flexible trading procedure, for instance possibility to incorporate 'local needs' in the contract, like warehousing, delivery, etc. Such dimensions, it is argued, are difficult to incorporate in a contract based on auctions in the EU/EEA.

The choice of trading procedure is modelled here as being dependent on the size of the order; the duration of the contract; the number of units covered by the contract; the manner in which procurement is organised; the age of the contract, in addition to the product dummies. Reporting only the price equation,

Table 3. Estimation results Model 3

	IV/2SLS Price Equation	
	Coef.	t-value[1]
Sup	.2898013	0.272
Metsup	-.1502052	-0.148
Method	.509987	0.114
prod_2	-4.321854	-22.084 ***
prod_3	-1.155125	-3.977 ***
prod_4	-3.675283	-8.936 ***
prod_5	-4.027865	-29.360 ***
prod_6	-.1850417	-0.527
prod_7	-2.836562	-17.597 ***
prod_9	1.011729	7.457 ***
prod_10	-1.650998	-9.208 ***
prod_11	-2.077766	-14.105 ***
prod_12	.3092107	1.083
prod_13	-.5016899	-2.984 ***
prod_14	-3.987263	-20.646 ***
Constant	1.213402	0.256
N=216		
R^2=	0,96	
F(15,200)=	710,30	

[1] Based on a robust variance estimator.
***Significant at 1%. ** Significant at 5%.

Table 3 show that none of the key variables are significant, as in the case of Model 2. Again based on the Hausman-test (Table 4 below), we conclude that Model 1 is a consistent specification. Adding that the price equation in Model 1 also passes the Ramsey test and the test for heteroscedasticity, we draw the conclusion that the log-transformed price equation in Model 1 performs better than the other specifications.

Table 4. Hausman test Model 1 vs Model 3

	(b) Prior	(B) Current	(b-B) Difference	S.E.
sup	.2898013	-.0302821	.3200834	1.064422
metsup	-.1502052	.0415557	-.1917609	1.010595
method	.509987	-.2407769	.7507639	4.476775
prod_2	-4.321854	-4.248223	-.0736312	.1335811
prod_3	-1.155125	-1.472111	.3169856	.2733688
prod_4	-3.675283	-3.409963	-.2653201	.4034771
prod_5	-4.027865	-3.993724	-.0341413	.1213308
prod_6	-.1850417	-.4966163	.3115746	.3363686
prod_7	-2.836562	-2.764367	-.0721949	.111415
prod_9	1.011729	1.011695	.0000346	,1134886
prod_10	-1.650998	-1.765611	.1146132	.1016106
prod_11	-2.077766	-2.10269	.0249245	.1118923
prod_12	.3092107	.0015112	.3076996	.273174
prod_13	-.5016899	-.6474542	.1457644	.1542089
prod_14	-3.987263	-4.152184	.164921	.1768643

b = less efficient estimates obtained previously from ivreg.
B = more efficient estimates obtained from regress.
Test: Ho: difference in coefficients not systematic
chi2(15) = (b-B)'[(V_b-V_B)^(-1)](b-B) = 2.45
Prob>chi2 = 0.9999

6. CONCLUDING REMARKS

Bargaining between two parties – a public procurer and a private provider – may be described as a trading procedure quite the opposite of the procurement regimes put forward by EU. Most envisaged bargaining processes would clearly score low on scales measuring openness, equal treatment and transparency. The main question in this study has been whether bargaining is less efficient then auctions in eliciting low prices.

The main conclusion from our study is that auctions do not appear to have an upper hand compared to negotiations in terms of eliciting low prices. We do find that choice of trading procedure appears to have a significant impact on number of suppliers participating in the trading process. However, we do not find evidence suggesting that number of suppliers matters for prices. Of course, more studies are needed to make any final judgement of the relative merits of the two trading procedures, in the market for medical and surgical articles and on other markets. One important challenge would be to limit measurement errors due to quality differentials. The experience made through this project is that much follow-up work is needed to minimize these errors.

It is customary to argue that compared to running an auction, negotiations are more likely to be a relative expensive way of doing business. It then follows that bargaining represents a trading procedure that should be expected to involve fewer competitors than auctions and our results suggest that negotiations do give thinner markets.

Focusing only on per unit prices as we have done here may be a too narrow-minded approach, though. Both bargaining and auctions are likely to involve some form of ex ante and ex post costs, e.g. preparation costs and contract administration costs. Bargaining costs – likely to increase with the number of potential suppliers involved - must be compared to preparation costs associated with auctions. Still, many will argue that auctions represent a more cost efficient way of conducting business. In that case, our results give support for choosing auctions rather than negotiations. If bargaining is more expensive on average than auctions, our price study has underestimated the positive total effects of using auctions.

A possible argument in favor of bargaining, an issue that we have not dealt with in the empirical study, is the potential for determining the most preferred quality before any procurement is made. To the extent that a procurer wants to learn the quality of the goods before purchasing, the bargaining process represented in our study may be better than a first-price sealed-bid auction based on a more or less good description of the goods to be procured. Auction may well have a higher probability of missing quality targets set by the users of medical and surgical articles. Still, procurers using auction have the possibilities of testing products before tendering. In consequence, neither in this respect is there any reason to expect great differences between the two trading procedures.

Appendix

Table A1. Copy of Questionnaire

Supplier's product no.:
Name of product:
Manufactured by:

If a different supplier/manufacturer:

Supplier's product no.:
Name of product:
Manufactured by:

If the product is not in the 'basket', tick here: [ ]

A. The Contract:

1. *Size of contract in volume (no. of units):*	.... units
2. *Price per unit (excl. VAT):*	..................NOK/unit
3. *Date of contract (month/year):*	........./................
4. *Duration of contract (in months):*	....................months

B. Method:

1. Tender in EU/EEA:	[ ]
2. Tender in Norway:	[ ]
3. Negotiations with suppliers:	[ ]
- numbers of suppliers:	
4. Purchase without competition:	[ ]
5. *Other (specify):*	[ ]

C. Organization:

1. Centralized (the county's procurement unit):	[ ]
2. Decentralized (hospital's procurement unit)	[ ]
3. Centralized (county) + cooperation with other counties:	[ ]
4. Decentralized (hospital) + cooperation with other hospitals:	[ ]
5. Decentralized to hospital wards:	[ ]
6. *Other (specify):*	[ ]

D. Market conditions:

1. *No. of potential suppliers (as percieved before purchase)*	....
2. *No. of suppliers participating:*	....
3. *Wholesaler in Norway:*	Yes [ ] No [ ]
4. *Suppliers in Norway:*	Yes [ ] No [ ]
5. *Sales representative in Norway:*	Yes [ ] No [ ]
6. Other:	Yes [ ] No [ ]

Table A2. The 'basket' of products

Product1:	Diathermy plate
Product2:	Compress
Product3:	Compress, surgery
Product4:	Gloves, disposable
Product5:	Syringe
Product6:	Plaster cast (natural)
Product7:	Hat, disposable
Product8:	Central line catheter
Product9:	Gown, surgery
Product10:	Hat ('helmet')
Product11:	Mask
Product12:	Urethric catheter
Product13:	Cannula, infusion
Product14:	Cannula, injection

Table A3. Summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Number of suppliers	249	4.650602	1.830083	1	12
Number of units ordered	233	107656.2	258559	24	2000000
Age of contract	249	21.84337	13.7035	0	48
Duration of contract	249	26.52209	8.833616	12	48
Number of hospitals covered by the contract	249	2.64257	1.817465	1	7

Table A4. Summary statistics – prices (in NOK)
(based on the 229 contracts used in the regressions)

	Number of contracts	*Mean*	*Std. Dev.*	*Min.*	*Max*
Product 1	18	12.20278	2.717349	9.02	17.04
Product 2	11	.1845455	.0709033	.08	.31
Product 3	15	2.745333	.8209303	1.47	3.60
Product 4	19	.4031579	.1038189	.31	.77
Product 5	19	.2210526	.0366507	.19	.36
Product 6	15	7.356	1.892205	3.78	10.2
Product 7	17	.8223529	.3936294	.49	1.77
Product 8	13	97.14923	46.05194	35	186.12
Product 9	20	33.1695	8.043932	24.13	58.58
Product 10	17	2.352353	1.622539	1.04	3.95
Product 11	18	1.508333	.4532659	.79	2.14
Product 12	16	11.81875	3.12969	9.4	23.31
Product 13	17	6.117059	.7124672	5	7.33
Product 14	14	.1885714	.0389984	.15	.28

REFERENCES

Arnold, M.A. and S.A. Lippman (1998), "Posted Prices versus Bargaining in Markets with Asymmetric Information", *Economic Inquiry*, XXXVI, 450-457.

Bester, H. (1993), "Bargaining versus Price Competition in Markets with Quality Uncertainty", *American Economic Review*, 83, 278-288.

Bulow, J. and P. Klemperer (1996), "Auctions vs Negotiations", *American Economic Review*, 86, 180-94.

Cresta, J.-P. (1991), "Bargaining", in Laffont, J.-J. and Moreaux, M. (eds.): *Dynamics, Incomplete Information and Industrial Economics*. Basil Blackwell: Oxford.

Engelbrecht-Wiggans, R. (1993), "Optimal Auctions Revisited", *Games and Economic Behavior*, 5, 227-239.

French, K.R. and R. McCormick (1984), "Scaled Bids, Sunk Costs, and the Process of Competition", *Journal of Business*, 57, 417-441.

Fudenberg, D. and J. Tirole (1983), "Sequential Bargaining with Incomplete Information", *Review of Economic Studies*, 50, 221-247.

Fudenberg, D. and J. Tirole (1991), *Game Theory*, The MIT Press: London.

Kennan, J. and R. Wilson (1993), "Bargaining with Private Information", *Journal of Economic Literature*, XXXI, 45-104.

Kjerstad, E. and S. Vagstad (2000), "Procurement Auctions with Entry of Bidders", *International Journal of Industrial Organization*, 18 (8), 1243-57.

Klemperer, P. (1999), "Auction Theory: A Guide to the Literature", *Journal of Economic Surveys*, Vol. 13, 3, 227-286.

Laffont, J.-J. and J. Tirole (1987), "Auctioning Incentive Contracts", *Journal of Political Economy*, 95, 921-937.

Levin, D. and J.L. Smith (1994), "Equilibrium in Auctions with Entry", *American Economic Review*, 84, 595-99.

McAfee, P. and J. McMillan (1987a), "Auctions with Entry", *Economic Letters*, 23, 343-347.

McAfee, R.P. and J. McMillan (1987b), "Auctions and Bidding", *Journal of Economic Literature*, XXV, 699-738.

McAfee, R.P. and J. McMillan (1996), "Competition and Game Theory", *Journal of Marketing Research*, XXXIII, 263-267.

Milgrom, P. (1987), "Auction Theory", in Truman F. Bewley (ed.), *Advances in Economic Theory, Fifth World Congress*. Cambridge University Press: Cambridge.

Milgrom, P. (1989), "Auctions and Bidding: A Primer", *The Journal of Economic Perspectives*, 3, 3-22.

Osborne, M.J. and A. Rubinstein (1990), *Bargaining and Markets*, Academic Press: San Diego.

Rasmusen, E. (1994), *Games and Information, Second Edition*, Blackwell: Cambridge.

Riordan, M.H. and D.E. Sappington (1987), "Awarding Monopoly Franchises", *The American Economic Review*, 77, 375-387.

Rubinstein, A. (1982), "Perfect Equilibrium in a Bargaining Model", *Econometrica*, 50, 97-100.

NIS (1997), SAMDATA sykehus, rapport 6/97, SINTEF Unimed, Norsk institutt for sykehusforskning.

Samuelson, W.F. (1985), "Competitive Bidding with Entry Costs", *Economic Letters*, 17, 53-57.

Ståhl, I. (1972), *Bargaining Theory*, Stockholm School of Economics: Stockholm.

Tan, G. (1992), "Entry and R&D in Procurement Contracting", *Journal of Economic Theory*, 58, 41-60.

Vickrey, W. (1961), "Counterspeculation, Auctions, and Competitive Sealed Tenders", *Journal of Finance*, 16, 8-37.

Wilson, R. (1987), "Game-theoretic analyses of trading processes", in Bewley, T.F. (ed.) *Advances in Economic Theory, Fifth World Congress*. Cambridge University Press: Cambridge.