EUROPEAN UNIVERSITY INSTITUTE
Department of Economics

Essays on Financial Intermediation

Paul Schure

Thesis submitted for assessment with a view to obtaining the degree of Doctor of the European University Institute

Florence
June 2000

EUROPEAN UNIVERSITY INSTITUTE
Department of Economics

Essays on Financial Intermediation

Paul Schure

The Thesis Committee consists of:

Prof. Franklin Allen, Wharton School
" Pierpaolo Battigalli, EUI, internal advisor
" Dave Cass, University of Pennsylvania, Supervisor
" Joseph Zeira, Hebrew University of Jerusalem

Contents

Introduction

Financial Intermediaries (FIs) mediate between investors, in the broad sense of the word, and those who search finance. An obvious example of a FI is a commercial or a savings bank that, among other things, mediates between depositors and firms and other agents that look for loans. But there are many examples of FIs, such as investment banks, mortgage banks, insurance companies, venture capitalists, mutual and money market funds, lottery organisers and rating agencies.

Many FIs engage in the important business of producing and processing information about the firms and other agents that apply for finance. I believe the activities of such FIs are fascinating for two reasons. First, sometimes large sums of money are involved and the decision of the FI can crucially affect important decisions such as a firm's investment decision. Second, the nature of acquiring and processing information makes it difficult for outsiders to assess how good a job the FI does.[1] Possible consequences are that investors in the FI do not know precisely the value of their assets and that the FI's clients do not know precisely what 'product' they are buying. Let me give an example. Consider a firm that attempts to raise finance to invest in a new plant or to develop a new product. One way to achieve this is to issue new shares on the stockmarket and in this case the firm normally approaches an investment bank. One of the tasks of the investment bank can be to price the new equity, another is to thoroughly verify the firm's conduct and

[1] In fact, the financial intermediation literature stresses that an important reason for having FIs is precisely because it is efficient to delegate the task of acquiring and processing information to a single agent or institution. So, if it were no problem for us to observe precisely what the FI did, there would be one reason less for the FI to exist!

books. The investment bank has more clients than the firm in question, namely the investors who consider buying the new equity. These investors can of course do their own research on the firm, but if all investors would choose to do so, it is not clear why the firm approached the investment bank in the first place. Normally, therefore, the investors rely to an important degree on the information provided by the investment bank when deciding to buy equity. These investors are thus in the awkward position to buy shares without knowing whether they have a good deal. Likewise, the firm can merely hope that the investment bank incurs enough effort to sell the shares for a good price.

The problems addressed in the first two chapters of the thesis are closely related to the picture sketched above. The FI engages in the opaque, but costly activities of screening and monitoring borrowers and enforcing contracts. Since outside investors find it difficult (i.e. costly) to assess the performance of the FI, the FI has the additional problem to convey that it does a good job. In Chapter 1 the wealth endowment of the FI serves as the signal that the FI is committed to its task. Chapter 2 shows that an even better strategy for the FI can be to invest in illiquid assets. In other words, holding illiquid assets and engaging in opaque activities can be complementary. However, the model shows this is only the case if the FI is financed with short-term hard claims, e.g. debt or deposit claims.[2] The final chapter of the thesis is based on joint work with Rien Wagenvoort in which we deal with an empirical issue on financial intermediation. We determine the efficiency of the European banking sector during the period 1993-1997 by adopting a so-called cost frontier method. Let me next introduce the Chapters that follow in little more detail.

[2] By the way, I just sketched that Chapter 1 and Chapter 2 deal with financial intermediation. I shall argue, however, that the models apply more generally to all firms that engage in opaque activities or offer difficult-to-assess products.

Why Wealth Generates More Wealth: Finance and Financial Intermediation

Chapter 1 deals with the relationship between wealth and finance. In particular I address the question: "What is the role of a wealthy investor when the remaining investors are poor?" The model assumes that investment projects require to be monitored by investors and that monitoring efforts are best delegated to a single agent. I show that the distribution of wealth among the agents has important consequences for finance. First, the wealthy investor is the most efficient delegated monitor. Second, if the wealthy investor indeed becomes delegated monitor, she can extract a surplus which cannot be competed away by the poor investors. Thus, a large wealth endowment can give an investor a degree of market power in the capital market. An interesting corollary of the findings above is that even though investing may be efficient (second best) none of the investors have enough wealth to become a FI.

Asset Liquidity, Short Term Debt and Managerial Effort: the Example of Banks

In Chapter 2 the focus is on the financing of firms that engage in opaque (i.e unverifiable) activities. I first show that such firms are inclined to deliver a bad 'product' if those in control bear a large share of the costs of their efforts while outside stakeholders receive a large share of the benefits of the efforts incurred. The model then shows that this 'free rider problem' is mitigated if the firm invests in illiquid assets. In the paper I study the leading example of commercial banks. The theory predicts that two features make a bank an efficient institution in screening and monitoring projects. First, the bank possesses illiquid long-term assets. Examples of these assets are the bank's loan portfolio or its brandname or reputation. Second, the bank is financed short-term so that it faces the risk of liquidation. In case of liquidation the future rents accruing from the illiquid assets are lost so that a bank with illiquid assets can make a credible commitment to incur a high amount of monitoring effort. I show that the open market is not able to make such a credible commitment.

Economies of Scale and Efficiency in European Banking: New Evidence

In Chapter 3 Rien Wagenvoort and I investigate the cost efficiency of 1974 credit institutions in the 15 European Union (EU) countries in the period 1993-1997. The sample period 1993-1997 is interesting because it immediately followed the implementation of the Second Banking Directive of the EU. This directive implied a large degree of deregulation of the EU banking sector. We estimate a Cobb-Douglas cost frontier which is augmented with dummies to allow banks with different legal structures or with different scales of operation to operate with different costs per unit of assets. We also incude time dummies to allow the cost frontier to shift over time. Our study shows that the most important source of inefficiency in the European banking sector is managerial inability to control costs. We do not find any major gains from economies of scale or technological progress. Finally, we find big efficiency differences between the EU countries in the sample period.

Acknowledgments

I started my Ph.D. at the European University Institute in September 1995. Apart from being at the EUI I spent 5 months at the European Investment Bank in Luxembourg, several brief periods at the University of Victoria in Canada, and my final year at the Hebrew University of Jerusalem. I would like to thank all these institutions for their kind hospitality. I also gratefully acknowledge the financial support I received from the Dutch National Scholarship Foundation (Nuffic), the European University Institute, the Pinhas Sapir Economic Policy Forum, and the TMR network FMRX-CT96-0028 of the European Union.

Very many people have had an influence on the contents of my Ph.D. dissertation. I would first of all like to thank my supervisors Pierpaolo Battigalli, David Cass and James Dow for their input. Although some say that having more than one supervisor is a serious disadvantage for a student, I have learned different lessons from each of them. Dave, my main supervisor, has taught me not to worry too much initially about the existing literature when having an interesting problem, and rather to spend my energy on solving it. He has also taught me to be precise in my mathematics; once he sat down next to me and we went through all the proofs in one of my earlier papers. Finally, our frank discussions about economic methodology in the beginning of my second year at the EUI have been very important. James has taught me to spend time to think through 'the idea' of a paper well, before actually starting to write it. He was also the one who has introduced me to the world of financial economics. Chapter 1 has benefited enormously from James' comments. I started working with Pierpaolo only one and a half years ago, but Chapters 1 and 2 would have looked completely different without his input. Strategic interaction between the agents plays an important role in corporate finance. Pierpaolo has forced me to make this explicit in my work. Apart from this, he has given me really fantastic comments on my papers. Finally, and probably without him knowing, Pierpaolo has been an important motivator while I was finishing my thesis.

Franklin Allen and Joseph Zeira are the two external members on my thesis committee. I would like to thank them very much for their comments on my thesis draft. Franklin's and Joseph's comments were not only of high quality, but the supportive nature of their comments also make me look forward to carry on my research after having completed my dissertation. On this occasion let me thank you, Joseph, for your kindness in receiving me at Hebrew. You have done more than can be expected to make me feel at ease at the Economics Department.

Let me now thank the other people that have been helpful to the thesis. Chapter 1 has benefited from the comments of Merwan Engineer, Marco Haan, Eugene Kandel, Rainer Kiefer, Yehuda Kotowitz, Eric Perée, Bert Smid, Rien Wagenvoort, Yishay Yafeh, participants of the 1998 'Student Forum' at the European University Institute, participants of the 'Theme E' seminar series at the University of Groningen, participants of Finance Group Seminars at the University of Amsterdam, and participants of the XII World Congress of the International Economic Association. Regarding Chapter 2, I would like to thank Merwan Engineer, Eugene Kandel, Rainer Kiefer and David Weil. Also thanks to the participants at the 1999 Annual Meeting of the Association of Southern European Economic Theorists (ASSET), and participants of seminars at the Finance Department at Tel Aviv University, the Economics Department at Hebrew University, the Norway School of Management, BI, the Bank of International Settlements (BIS) in Basel, the Economics Department and the Business School at the University of Victoria, and the Business School at Hebrew University. Regarding Chapter 3, I would first of all thank my co-author Rien Wagenvoort. Rien, not only did I learn a lot from our collaboration, it is also great fun working with you! Let me also thank Christopher Hurst, Søren Johansen, Dermot O'Brien, and seminar participants at Hebrew University.

I shall not mention explicitly my dear friends. Obviously, my friends have been very important to me during the last five years. However, I expect that they know who they are and I hope that they realise how grateful I am for having them around me. I would like to thank my parents, Jan en Corry, and my brothers, Frank, Joost en Bart, for their warm support. Finally, I would like to express very special thanks to Amy and Zoey Verdun.

Chapter 1

Why Wealth Generates More Wealth: Finance and Financial Intermediation

1.1 Introduction

Consider a one-period setting with investors and managers with an investment project. Let the project be indivisible and large so that several investors are needed to finance it. Assume that two frictions play a role when the investors and the managers engage in financial contracts. First, investors can only observe the project return by incurring a cost while managers observe it freely. In other words, we consider a *costly state verification* problem. Assume that monitoring expenses need to be incurred before the project returns are realised, i.e. *ex ante*. Also, assume that monitoring costs are duplicated when more investors monitor the managers. Costly bankruptcy is the other friction. If wished, costly bankruptcy can be interpreted as *ex post* costly state verification.[1] Finally, assume that the investors have the bargaining power when approaching the managers. Having all the

[1] Standard papers on the costly state verification problem are Townsend (1979) and Gale and Hellwig (1985). These papers address what I just referred to as ex post costly state verification.

bargaining power, the investors' problem is to finance the project so as to extract the maximum possible value.

Let the project be financed with either debt or equity. Pure debt or equity finance can never be optimal but, as we shall see, for the purpose of this chapter it is a convenient assumption. From an investor's point of view debt has an advantage and a disadvantage when compared to equity. The advantage is that the manager has the right incentives to repay whenever possible, so that the (ex ante) monitoring expenses need not be incurred. On the other hand there are two costs. First, if the project return of the manager is lower than the face value on the debt contract the costly liquidation procedure is invoked. Second, if the return exceeds the face value the manager obtains all production exceeding the face value. With an equity contract more value of the manager's project can be extracted but costly monitoring expenses have to be incurred. Namely, if the investors would not monitor, the manager would divert the project cash-flows.

The focus of this chapter is on the role of a wealthy investor when the remaining investors are 'poor'. I will show that in the framework sketched above the wealthy investor has access to investment opportunities which dominate the investment opportunities of the poor investors if her wealth endowment is sufficiently large. In particular she can become a financial intermediary (FI) taking debt from poor investors to supply equity in the project.

The intuition for this result is relatively straightforward. Focus on a single project and first consider equity finance. As monitoring expenses are duplicated with multiple investors, equity becomes more attractive if the fraction of the project which is financed is sufficiently large. By contrast, the rate of return on debt finance is constant in the fraction financed. In this setting a FI is an agent taking debt deposits from investors to buy a larger equity share in the project.[2] However, investors are reluctant to bring their funds to the FI as they recognize that, just like the manager, also the FI has an incentive to divert cash-flows. To convince the investors that they should deposit funds at the

[2]The model shows that the deposit contract can only be debt.

intermediary she needs to pledge a part of her endowment as *collateral* on the deposits. This collateral mitigates the agency problem attached to external finance. As sufficient collateral is needed, in some circumstances only the wealthy investor can successfully become a FI.

The closest empirical counterpart of the FI in this chapter is a European type of investment bank. These banks take debt to supply equity capital in firms. According to the model in this chapter this is profitable once the equity share taken is large enough. Namely, monitoring the firm (or its management) becomes cheaper when the shares taken are larger. Such a situation would apply if small shareholders face relatively higher costs to discipline management than large shareholders, e.g. due to 'voting costs'. More generally, costly expenses may have to be incurred to overcome the free-rider problem in monitoring management. In the conclusion to the chapter I will argue that the FI can also be interpreted as commercial banks or a 'wealthy' firm which takes over a 'poor' firm

What is the role of the FI in this model? And how does it relate to the existing financial intermediation literature? Hellwig (1991) and Bhattacharya and Thakor (1993) are two recent surveys on financial intermediation. Bhattacharya and Thakor divide the services of the financial intermediary into two categories: *brokerage* and *qualitative asset transformation* services. The FI acting as a broker brings together suppliers and users of capital without changing the nature of the claims being transacted. By contrast, when offering qualitative asset transformation services the FI offers asset and liability services which change the nature of the claims transacted. In this chapter the FI is a qualitative asset transformer as equity held by the FI and its deposits are of a different nature. The asset transformation services offered by the FI are twofold. First, the FI monitors the manager so she is a *delegated monitor* as in Diamond (1984, 1996). Second, as collateral is pledged on deposits depositors get some degree of repayment insurance. Turning to the second question above, what is the comparative advantage of the FI in providing these services? As Bhattacharya and Thakor mention, the bulk of the financial intermediation theory stresses the capacity of FIs to *interpret* and *re-use information* and states that the opportunities for *diversification* increases with their size. Early papers on this are e.g. Leland and Pyle (1977), Campbell and Kracaw (1980), Ramakrishnan and Thakor (1984) and Diamond (1984). Older work mentions that FIs reduce the number of transactions needed or have an exclusive beneficial relationship with the government.[3] More recently the call to view the FI as a broker becomes stronger. Allen and Santomero (1998) demonstrate that brokerage services form an increasingly important part of the bank's business and conjecture that this is due to participation costs in financial markets.

The model in this chapter is in the spirit of the informational explanations of the FI. However, the results do not hinge on the FI having any special technology in dealing with information, or limited participation. The FI has the same skills to reduce risks, gather information, or produce contracts as the remaining investors. The mere fact that the FI has a large amount of capital which is not subject to agency problems drives the results. The amount of wealth forms a limit to participate in financial intermediation business. As the FI is really just a wealthy investor it is nontrivial to embed this paper in the existing financial intermediation theory.

The model introduced in this chapter is inspired by Diamond (1984, 1996). Diamond takes the same basic setup as in this chapter, but he assumes that there are many identical risky investment projects with independently distributed returns. He studies the role of diversification when all agents are risk neutral. He shows that an agent which can diversify the return risks of these projects becomes a FI. So, using the terminology of Bhattacharya and Thakor (1993), in Diamond pooling independent risks is the qualitative asset transformation service that the FI offers. In this chapter the FI and the other investors have the same diversification skills.[4]

Campbell and Kracaw (1980) and Holmstrom and Tirole (1997) are related as they

[3] An early example of the latter feature is Tobin (1963) which mentions that the government provides guarantees on the liabilities of the FI, regulates the banking sector and offers last-resort lending.

[4] I am convinced that risk diversification as in Diamond (1984, 1996) is a very important feature of FIs in reality. Leaving away this aspect in this chapter is therefore a matter of focus.

address the role of capital of the FI. Campbell and Kracaw view the FI as a producer of information about firms' investment projects. They show that FI needs a sufficient amount of capital to convince the market of the reliability of this information. Holmstrom and Tirole focus on the role of wealth in attracting capital for firms in general, and in particular FIs. They assume that external funds are subject to moral hazard making wealth a necessary requirement for attracting outside funds. The impact of a decrease in wealth in their model is shown to be broadly consistent with the evidence on the credit crunch of the late 1980s and early 1990s in the United States and Scandinavia. In Holmstrom and Tirole FIs can uniquely restrict the set of the firms' feasible actions; the FI is not a normal investor.

Hart and Moore (1995) address in their paper the role of 'hard debt' in constraining a firm's management. As I shall argue in the conclusion, the debt contract introduced in this chapter is comparable to 'hard debt' in Hart and Moore, and the equity contract to their 'soft debt' contract.

Finally, the model in this chapter is related to a vast branch of literature that focuses on the lender–borrower relationship in the presence of financial market frictions. Stiglitz and Weiss (1981), Bester (1987) and Williamson (1987) are examples of papers that explicitly clarify the role of the borrower's collateral when attracting funds.

The remainder of the chapter is structured as follows. In the next section I present some evidence from the early banking period that shows that, broadly spoken, until a hundred or so years ago wealthy agents played an important role like the FI of this chapter. The following section presents the model and also discusses some critical assumptions. I show that under some conditions an equilibrium with financial intermediation emerges and give a simple numerical example of the model. In the concluding section I discuss the interesting results and some caveats of the model.

1.2 Evidence from the Early Banking Literature

A crucial assumption of the model is that agents with small endowments face serious agency costs when bringing together their funds. As a consequence, wealthy agents will be the most efficient way to channel funds from small lenders to investment projects. Thus, the main prediction of the model is that, all other things equal, wealthy agents become financial intermediaries. Arguably, the scope of the model is limited in present day Western economies with relatively well-developed accounting standards and legal systems to enforce contracts.[5] However, this is less clear for some developing countries with underdeveloped financial and legal institutions.[6] In this section I try to find stylized evidence for this hypothesis from yet another source: the early banking period. Needless to say, it can be expected that the agency problems considered played a bigger role in the past than at present. The model predicts that, all other things equal, the early bankers were wealthy agents, or wealthy families. A subprediction is that, all other things equal, wealthy agents have acquired their wealth from third sources before becoming bankers.

It is not easy to find evidence for the hypotheses raised above. When writing on banking, economic historians have normally focused on the operational side and the investment portfolio of the early banks, and not so much on managerial issues and ownership structures. Also, while it becomes clear that historically the financial scene more than once was dominated by families, besides public institutions, it is not clear how these families acquired their wealth.[7]

Having said this, there are some very interesting works worth mentioning. Goldsmith (1987) presents a comparative study on premodern financial systems ranging from Periclean Athens to the Netherlands during the mid 17th century. Regarding Augustan

[5] However, see e.g. Jensen (1986), Shleifer and Vishny (1989) and Hart and Moore (1994, 1995). Also notice that evidence from the venture capital industry could form an interesting challenge for this statement.

[6] See e.g. La Porta et al. (1997a, 1997b).

[7] See e.g. De Roover (1966) in his 'The Rise and Decine of The Medici Bank 1397-1494'. The book is excellent, but unfortunately it mainly deals with the operations and investments of the Medici bank and its historical background

Rome he concludes "[p.44] The most important suppliers of external capital thus were not financial institutions but private wealthy citizen, and the largest of them, it has been claimed, was the emperor." It also becomes clear that in many of the societies he studies there is no banking system present at all! Notice that this is by no means inconsistent with the theory. Kindleberger (1993) makes an interesting observation in his *Financial History of Western Europe* regarding the source of bankers' wealth: "[p.176, regarding Western Europe in the 16th-19th century] Conventional wisdom has it that fortunes made in commerce tend to go at the next stage into industry. Such seems rarely to be the case. The tendency is rather to move into land or finance or both." A few pages later, Kindleberger presents the interesting cases of the Morrison family and Sir John Ellerman who became rich in trade after which they moved into banking and rapidly accumulated more wealth. He concludes: "[p.185] Before 1880 industrialists were not only less rich, they were less singled out for elevation to the peerage than bankers or civil servants.."

The most interesting remarks I found in an edited volume on finance and financiers in European history in the period 1880-1960. Daunton (1992) reviews the literature that states that the financial elite of the City of London was not only rich, but also formed a close-knit network by for instance intermarriage.[8] From Plessis (1992) and Augustine (1992) it becomes clear that this was no different in France and Germany. Interestingly, Augustine (1992) finds in her unique dataset of all 502 German entrepreneurs in industry, trade and commerce, and banking that owned 6 million German marks or more, that 27 percent were bankers and that two-third of the bankers in her study were themselves sons of bankers.

The evidence above makes clear that, compared to today, in the early banking period wealthy families played a relatively important role in channeling funds from lenders to borrowers. Also, families acquired their wealth more than once in other trades before proceeding into the banking business. Finally, nonfinancial ties may have played a role in

[8]Although, the authors seem to disagree whether this elite was a financial elite or, rather, a business elite.

resolving agency problems between families in the business community. Unfortunately, the evidence had nothing to say about the precise role of the banker. It is unclear whether the banker provided funds only, or also had to actively monitor the loan portfolio.

1.3 When a Wealthy Consumer becomes a Financial Intermediary

Setup

Consider a one-period model with two dates of interest, time 0 and time 1. Distinguish the following agents: a 'wealthy' consumer, 'poor' consumers, and managers. There are countably infinite poor consumers and managers. All agents maximize expected lifetime consumption of 'goods', so they are risk-neutral and time 1 consumption is not discounted. Let the wealthy consumers be endowed with e_0^W goods at time 0, and all poor consumers with $e_0^P > 0$ goods. Time 1 endowments are zero. Assume for simplicity that the endowment of the wealthy consumer is a multiple of the endowment of the poor consumers: $e_0^W = ne_0^P$, for some $n \in \{1, 2, ...\}$. Managers have no goods endowment but have access to an indivisible *production technology.*

The projects described by this production technology require an input of $i_0 > 0$ goods in period 0 and yield in the next period a random production amount of $\tilde{x}_1$. Let the distribution of the production amount be representable by a cumulative distribution function F. Projects are assumed to be large in the sense that no consumer has enough wealth to finance it

$$e_0^P \leq e_0^W < i_0$$

Finally, project returns are assumed to be perfectly correlated: in what follows the issue of risk diversification plays no role. Apart from the production technology there is a *storage technology.* If at time 0 goods are stored this yields the same amount of goods at time 1. Let us assume without loss of generality that consumers will only store goods if this increases their expected consumption of goods. As we will see later, the consumer which becomes a financial intermediary stores goods.

With respect to the production technology we make the following assumptions. First, average production exceeds the investment amount

$$E[\tilde{x}_1] > i_0 \tag{1.1}$$

Second, production is risky in the sense that production may turn out to be lower than the amount invested. In fact, we make a slightly stronger assumption

$$\Pr\{\varepsilon \leq \tilde{x}_1 < i_0\} > 0, \text{ for some } 0 < \varepsilon < i_0 \tag{1.2}$$

Third, for simplicity, assume that $m > 1$ poor consumers can just finance a project:

$$me_0^P = i_0 \tag{1.3}$$

Since production increases the expected amount of goods (equation 1.1) and consumers are risk-neutral, production is first-best. Therefore in a perfect world the consumers and managers would engage in financial contracts at time 0. Thus, the manager would attract the necessary funds to invest. Then, at time 1, managers would pay back part of the amount produced. However, we assume the presence of *frictions* in this economy.

In particular, assume that consumers cannot directly observe the project returns whereas the managers can. As a result, at time 1 the manager has an incentive to default on any contractual arrangement made at time 0, diverting the project cash-flows. In this way managers can consume more of the project return than is specified in the contract. I assume that investors can offset this in two ways. First, the project can be monitored. By monitoring the project a lender gathers evidence on the project returns, but it costs k units of goods. Thus, the lender can be delegated monitor as in Diamond (1984, 1996). Assume (i) that the decision to monitor has to be made *ex ante*, i.e. at time 0 and (ii) that consumers cannot communicate monitoring outcomes among each other. Assumption (i) says that investors have to precommit on monitoring. An interpretation of this as well as of assumption (ii) is given below. A second way to enforce payments to the consumers at time 1 is by liquidating the project when the manager defaults. I assume (i) that managers do not have the chance to divert cash-flows when the project is liquidated and (ii) that liquidation is costly.[9] Costly liquidation can be interpreted as *ex post* costly state verification.[10] For simplicity I take the extreme case that *all* assets are destroyed when the project is liquidated. Thus, in case of liquidation neither the manager nor the consumers get any return.[11]

Let me give an interpretation of the assumptions given above. Assume that in the economy sketched above courts can enforce the contracts. Courts are assumed to function at no cost when the investor who appeals has gathered evidence on how much is produced. But in order to do so monitoring costs k are made. Court appeals can also be made when no evidence is gathered, but this is costly. In fact, gathering the evidence *ex post* is assumed to be so costly that all the assets of the manager at time 1 are destroyed. Why do the monitoring costs k have to be duplicated with multiple investors? Investors that have incurred the monitoring costs will propose a settlement to the manager at time 1, rather than actually stepping to court. In this way a larger part of the produced amount can be extracted at the expense of the investors that decided not to gather evidence. Investors cannot hire an agent gathering the evidence (say an accountant) eiher: any such agent

[9]Assumption (i) is in the spirit of the incomplete contract literature [see e.g. Aghion and Bolton (1992) or Hart and Moore (1994, 1995)]. This can best be seen in framework with three relevant dates. Date 0 is the investment date. On date 1 production yields 'assets' worth of $\tilde{x}_1$. 'Assets' are not marketable but can be transformed into marketable assets ('cash') by date 2. While production requires managerial expertise, both investors and managers can transform assets into cash. It is assumed that courts cannot verify the amount of assets or cash the agents hold, but that contracts can be written on (i) transfers of assets or cash and (ii) ownership of the project. Ownership entails the right to liquidate the project whenever wished. As assets are nonmarketable the liquidation value of the project at time 1 is zero. If the project of the main text is seen in this way it becomes clear that credible liquidation threats motivate the manager to hand over assets to investors whenever possible. This credible threat is obtained by monitoring the project.

[10]See e.g. Gale and Hellwig (1985), Winton (1995) and Schure (1998).

[11]Clearly, this assumption is stronger than needed, but relaxing it complicates the analysis without offering additional insight in the question raised in this paper. I do need to assume that on expectation bankrupcy costs exceed the monitoring costs k. If this were not the case monitoring at time 0 is never interesting.

would be bribed by the manager. Summarising, when the evidence is available investors have a credible threat to step to court, but in equilibrium they will never do so.[12]

I assume that at time 0 two possible financial contracts can be concluded between the consumers and the managers: *debt* contracts and *equity* contracts. Debt contracts are characterized by three variables: the *loan* amount, a *collateral* amount, and a *repayment requirement* or face value. Of course, as managers have no wealth endowment, the amount of collateral can only be taken equal to zero when a manager is the borrowing party. In case the manager does not come up with the repayment requirement at time 1 assume that the project is liquidated automatically.[13] As all assets are destroyed in case of liquidation, the manager will come up with the face value whenever possible. Therefore, with debt no monitoring costs need be incurred. Equity contracts entail the right to some fraction of the project returns. Because with equity managerial incentives are absent, equity holders need to monitor. If not, their return becomes zero. It can be concluded that, from the investors' point of view, debt has both an advantage and a disadvantage when compared to equity. The advantage is that investors economize on monitoring costs, the disadvantage is that costly liquidation can occur and that less wealth is extracted if project returns turn out to be high.[14]

In the sequel I maintain the assumption that consumers have all the bargaining power in the contracting process with the managers.[15] Also, assume for simplicity that projects are either all-equity or all-debt financed. Although more efficient, a mixed finance structure is not allowed for. Finally, I assume that the contracts offered are 'symmetric',

[12] As we shall see below, contracts for which investors monitor are equity contracts. Therefore, 'gathering the evidence' can for example be interpreted as 'taking a seat in the supervisory board of a firm'.

[13] In other words, debt is 'hard debt'. We do not allow for renegotiations, even though this might be optimal *ex post*, i.e. after the manager has defaulted. In other words, the presence of some precommitment device is assumed (for example, courts that have an interest in liquidating the project when the manager defaults).

[14] Gale and Hellwig (1985), Winton (1995) and Schure (1998) have shown that if costly liquidation is the only friction debt dominates equity. In this model this result does not hold as costly liquidation is prevented with an equity contract.

[15] The assumption would be appropriate when the managers compete for scarce funds of consumers. Although the assumption is much stronger than needed I do need to assume that consumers have some degree of bargaining power when contracting.

i.e. (i) the returns are proportional to the amount invested, and (ii) when monitoring takes place all creditors monitor. Again this assumption is made for simplicity. Winton (1995) shows that symmetric debt contracts are dominated by debt contracts with varying seniority.

The Benchmark Case

In the benchmark case the projects of the managers are financed *directly*, i.e. the consumers approach the managers directly to finance the project. In the next subsection we will see under what circumstances the benchmark case applies.

Take a consumer, say C^i, with endowment e_0^i who considers financing a fraction, say s^i, of a manager's project. Let $s^i i_0 \leq e_0^i$ so that financing this fraction is compatible with the consumer's budget constraint. A debt contract with repayment requirement $s^i r_1$ will yield on expectation $s^i r_1 \Pr\{\tilde{x}_1 \geq r_1\}$. Therefore, the optimal debt arrangement is to set the repayment requirement to $s^i r_1^*$ where r_1^* is given by:

$$r_1^* = \arg\max_{r_1}\{r_1 \Pr\{\tilde{x}_1 \geq r_1\}\}$$

Buying equity and incurring the monitoring costs yields an expected utility of

$$s^i E[\tilde{x}_1] - k \tag{1.4}$$

As has become clear in the preceding subsection, when not monitoring the return on equity becomes zero and this is never optimal. Investing either in debt or in equity is interesting if

$$\max\{s^i r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\}, s^i E[\tilde{x}_1] - k\} \geq s^i i_0 \tag{1.5}$$

In connection to these equations I now introduce the following definition.

Definition 1 *"Debt (finance) is viable" if $r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\} \geq i_0$. In this case direct debt yields higher expected consumption than consuming directly. "Monitoring is viable" for a consumer i with endowment e_0^i and equity share s^i if*

$$e_0^i - s^i i_0 + s^i E[\tilde{x}_1] - k \geq 0$$

Monitoring is viable if with the amount of goods left over after investing and the expected return on equity the monitoring expenses can be incurred.[16] *"Equity (finance) is viable" for an equity share* s^i *if* $s^i E[\tilde{x}_1] - k \geq s^i i_0$. *In this case equity yields higher expected consumption than consuming directly. Note that if equity finance is viable then so is monitoring the manager. "Debt is (weakly) better than equity" for share* $s^i > 0$ *if*

$$s^i r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\} \geq s^i E[\tilde{x}_1] - k$$

"Equity is (weakly) better than debt" for share $s^i > 0$ *if*

$$s^i E[\tilde{x}_1] - k \geq s^i r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\}$$

"Direct finance is viable" when shares s^i, $i = 0, 1, 2, ...$ *can be chosen such that (i)* $\sum_i s^i = 1$ *and (ii) for all investors* i *with* $s^i > 0$ *debt or equity is viable.*

A few important remarks can be made. First, the mean return rate on debt is $\frac{r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\}}{i_0}$ and this quantity does not depend on the share invested. Instead, the mean return rate on equity $\frac{s^i E[\tilde{x}_1] - k}{s^i i_0}$ increases with the share. This is because monitoring costs are independent of the share whereas the project return is proportional to the share. It follows immediately that for small enough shares debt is better than equity. Second, because projects are large and have perfectly correlated returns, and the rate of return on both equity and debt is non-decreasing in the fraction s^i, we can assume without loss of generality that consumers invest in a single project. Therefore, as of now consider a single project and a manager M, say. Third, despite of the fact that M has no bargaining power when contracting she will attain a positive surplus in the case of debt. Returns exceeding the face value of the debt contract are kept. To put this differently, with debt finance the manager retains the residual cash-flows, so has 'inside equity'.[17] In particu-

[16] *Note that even if monitoring is viable it could be the case that for some realisations of* $\tilde{x}_1$ *the monitoring expenses cannot be paid for. Yet, if this equation holds a perfect, risk-neutral (short-term) capital market at time* 1 *would be prepared to finance the monitoring expenses, if needed.*

[17] This result hinges on the fact that at time 0 the manager has no wealth and hence cannot make any side-payments to the investors.

lar, when the manager's project is completely financed with debt the project returns are divided as follows:

$$E[c_1^C] = r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\} \tag{1.6}$$

$$E[c_1^M] = E[\tilde{x}_1 - r_1^* | \tilde{x}_1 \geq r_1^*] \Pr\{\tilde{x}_1 \geq r_1^*\} \tag{1.7}$$

$$E[w_1] = E[\tilde{x}_1 | \tilde{x}_1 < r_1^*] \Pr\{\tilde{x}_1 < r_1^*\} \tag{1.8}$$

Here c_1^C and c_1^M represent the consumption of the consumers and the manager, respectively. The amount wasted by costly liquidation is w_1. We have $E[w_1] > 0$: (on expectation) resources will be lost. Namely, for debt to be viable we must have $r_1^* \geq i_0$. Then, from equation 1.2 we get

$$\Pr\{\tilde{x}_1 < r_1^*\} \geq \Pr\{\varepsilon \leq \tilde{x}_1 < r_1^*\} \geq \Pr\{\varepsilon \leq \tilde{x}_1 < i_0\} > 0$$

From this it also follows that $E[\tilde{x}_1 | \tilde{x}_1 < r_1^*] > 0$.

If the manager's project is all-equity financed, the entire production amount is extracted from the manager. In that case we obtain

$$E[c_1^C] = E[\tilde{x}_1] - k \sum_i I_{\{s^i > 0\}} \tag{1.9}$$

$$E[c_1^M] = 0 \qquad E[w_1] = k \sum_i I_{\{s^i > 0\}} \tag{1.10}$$

Here I_A stands for the indicator function which becomes one on a set A and is zero elsewhere, and w_1 represents the amount wasted by costly monitoring.

Having introduced the model setup I will now list the assumptions that will be maintained throughout the paper. First, I assume that direct finance is viable. Second, with direct finance, debt is better than equity for all the consumers

$$\frac{e_0^P}{i_0} E[\tilde{x}_1] - k \leq \frac{e_0^W}{i_0} E[\tilde{x}_1] - k \leq \frac{e_0^W}{i_0} \left(r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\}\right) \tag{1.11}$$

As m poor consumers are needed to finance the project (equation 1.3) their share becomes $s^P = \frac{1}{m}$. Third, equity would be better than debt if it were supplied by a single consumer.

$$E[\tilde{x}_1] - k > r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\} \tag{1.12}$$

Without this assumption there would have been no reason to have introduced equity contracts in the first place; debt would have been better than equity, irrespective of the share.

These assumptions result in what I will refer to in the rest of the paper as the benchmark case. In the benchmark case the managers' projects are viable and they are financed with debt directly sold to the consumers. All poor consumers buy a share of $s^P = \frac{1}{m}$ of the debt of a single project, whereas the wealthy consumer takes a share of say $s^W = \frac{n}{m} < 1$.

Financial Intermediation

I will here investigate whether there is scope for a *financial intermediary* in this economy. A financial intermediary (FI) is just an opportunistic consumer who seeks to increase expected consumption by becoming an entrepreneur borrowing money from other consumers so as to finance a larger fraction of the project. I focus on the particular case where one consumer becomes big enough fund a fraction of 1.[18] Assumption 1.12 shows that this intermediary will buy the equity of the manager. As equity requires monitoring the intermediary becomes a *delegated monitor* as in Diamond (1984). An alternative interpretation would be that the intermediary is an entrepreneur that takes over the project (including M's expertise) and spends k to restructure it.

Consumers depositing goods at the intermediary face an *identical* agency problem as when they approach the firm directly: at time 1 the FI prefers to divert assets rather than paying the depositors. Therefore, the intermediary needs to be incentivised appropriately by monitoring or a debt contract.

We will see that whether or not a consumer can successfully become a FI depends her endowment. Namely, by pledging funds as *collateral* when taking deposits, the intermediary can alleviate the agency problem of external finance. Since the wealthy consumer has the largest endowment, I study without loss of generality whether the wealthy consumer

[18] This strategy need not be optimal. The FI might want to consider financing more than one project. I make the assumption for simplicity and because investigating other possibilities is not the central issue in this paper.

can become intermediary.[19] However, in principle, also poor consumers could become intermediary.

Consider the wealthy consumer, C^W say, who has managed to take i_0 deposits from m consumers. From assumption 1.12 we know that C^W finances M with equity. Suppose that C^W pledges a *collateral* b_0 on the deposits (i.e. an amount of goods which is invested in the storage technology at time 0). At time 1 she then has an amount of goods of

$$\tilde{x}_1^{FI} \equiv \tilde{x}_1 + b_0 - k \tag{1.13}$$

Will the deposit contract issued by the FI described above be a debt or an equity contract? For C^W to become a FI two conditions must be met. First, the FI must be able to attract the necessary funds from the other consumers. Second, intermediation must be profitable. From this second condition it is easily seen that the deposit contract must be a debt contract. Namely, if the deposit contract were equity then at time 1 the FI would be stripped of all her assets, including the collateral amount. Clearly then financial intermediation is not profitable for C^W. If financial intermediation is viable then the deposit contract must be debt.

Poor consumers who deposit e_0^P require back on expectation at least what they would get with direct finance. Let ρ_1^d represent the repayment requirement on deposits. The participation constraint of the poor consumers becomes

$$s^P \rho_1^d \Pr\{\tilde{x}_1^{FI} \geq \rho_1^d\} \geq s^P r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\} \tag{1.14}$$

I will show next that this inequality implies that the FI will have to pledge collateral on deposits and that this collateral amount at least suffices to finance the monitoring costs

$$b_0 \geq k \tag{1.15}$$

Suppose instead that $b_0 < k$. In this case equation 1.13 implies

$$\Pr\{\tilde{x}_1^{FI} < \tilde{x}_1\} = 1$$

[19] As from now I will use the terms 'wealthy consumer' and '(financial) intermediary' interchangeably.

But then we must have

$$\rho_1^d \Pr\{\tilde{x}_1^{FI} \geq \rho_1^d\} < \rho_1^d \Pr\{\tilde{x}_1 \geq \rho_1^d\} \leq \max_{\rho_1^d} \rho_1^d \Pr\{\tilde{x}_1 \geq \rho_1^d\} \equiv r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\}$$

Consequently, the participation constraint of the poor consumers, equation 1.14, is violated. In words, poor consumers prefer lending directly to M over accepting the deposit contract. Note that from the FI's budget constraint ($b_0 < e_0^0$) and equation 1.15 we get that for intermediation to be viable the endowment of the wealthy consumer should exceed the monitoring costs

$$e_0^W \geq k \tag{1.16}$$

The individual rationality constraint of the C^W is the other necessary constraint for intermediation to be viable. If production $\tilde{x}_1$ is such that $\tilde{x}_1^{FI} \geq \rho_1^d$ the intermediary meets the repayment requirement on deposits at time 1. In that case intermediation yields

$$\tilde{x}_1^{FI} - \rho_1^d = \tilde{x}_1 + b_0 - k - \rho_1^d$$

If $\tilde{x}_1 + b_0 - k < \rho_1^d$ the intermediary is liquidated and all goods are lost. Therefore, the individual rationality constraint of the C^W becomes:

$$E[\tilde{x}_1 + b_0 - k - \rho_1^d | \tilde{x}_1^{FI} \geq \rho_1^d] \Pr(\tilde{x}_1^{FI} \geq \rho_1^d) \geq \frac{b_0}{i_0} r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\} \tag{1.17}$$

The wealthy consumer invests b_0 to become FI if the expected yield exceeds the yield of investing it directly in the manager's project.

In order to verify whether intermediation is viable we can focus on an easier version of constraint 1.17. When making poor consumers just indifferent between investing directly in the project and investing in deposits 1.14 holds with equality. Equation 1.17 then reduces to

$$E[\tilde{x}_1 + b_0 - k | \tilde{x}_1^{FI} \geq \rho_1^d] \Pr(\tilde{x}_1^{FI} \geq \rho_1^d) \geq \left(1 + \frac{b_0}{i_0}\right) r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\} \tag{1.18}$$

This version of C^W's individual rationality constraint would apply if she had all bargaining power when engaging in deposits contracts. It states that the expected return of the FI must exceed the opportunity costs of all the investors in the intermediary.

For the FI described above we took for granted that she attracted i_0 of deposits and pledged b_0 as collateral. In principle, however, the FI optimizes over these variables. In particular, it is not clear that the FI should store her endowment as collateral instead of investing it in the project. Nevertheless, I stick to this example because it is simple and it carries the main intuition by clarifying the role of wealth. We have come to the following result.

Proposition 2 *Financial intermediation is viable if there is a b_0 such that $e_0^W \geq b_0 \geq k$ and inequality 1.18 holds. The deposit interest rate in inequality 1.18 is determined implicitly by*

$$\rho_1^d \Pr\{\tilde{x}_1 \geq \rho_1^d + k - b_0\} = r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\}$$

Proposition 4 shows that there is scope for a FI if two conditions are satisfied. I show next that even if there is scope for a FI the economy may be worse of with a FI than without a FI. As all agents are risk-neutral let the 'performance' of the economy be measured by the expected aggregate consumption level. Notice that expected aggregate consumption is maximised when the amount of the goods lost due to monitoring and costly liquidation is minimized.

Equation 1.8 gives the expected amount of goods lost in the benchmark case. In the case of intermediation the expected loss arises from (i) monitoring costs and (ii) costly liquidation of the FI. Denoting the amount of goods lost with a FI by w_1^{FI} we therefore have

$$E[w_1^{FI}] = k + E[\tilde{x}_1 + b_0 - k |\, \tilde{x}_1^{FI} < \rho_1^d] \Pr\{\tilde{x}_1^{FI} < \rho_1^d\}$$

Financial intermediation enhances aggregate consumption if

$$k + E[\tilde{x}_1 + b_0 - k |\, \tilde{x}_1^{FI} < \rho_1^d] \Pr\{\tilde{x}_1^{FI} < \rho_1^d\} < E[\tilde{x}_1 |\, \tilde{x}_1 < r_1^*] \Pr\{\tilde{x}_1 < r_1^*\} \qquad (1.19)$$

The deposit interest rate (ρ_1^d) and the collateral amount (b_0) in equation 1.19 will in general depend on the division of the bargaining power between the FI and the poor consumers when concluding deposit contracts.

Does financial intermediation indeed enhance aggregate consumption? It need not. Suppose for example that $e_0^W = k$. Then from equation 1.15 it follows that depositors require that the FI pledges her entire endowment as collateral and from 1.14 we get that the deposit rate becomes $\rho_1^d = r_1^*$. Assume also that in this inequality 1.18 holds so that financial intermediation is indeed viable.[20] Then we get

$$E[w_1^{FI}] = k + E[\tilde{x}_1 | \tilde{x}_1 < r_1^*] \Pr\{\tilde{x}_1 < r_1^*\} > E[\tilde{x}_1 | \tilde{x}_1 < r_1^*] \Pr\{\tilde{x}_1 < r_1^*\}$$

Financial intermediation is *harmful.*

When a Financial Intermediary Emerges: An Example

Consider the economy introduced in the beginning of Section 2.2 Let the poor consumers have an endowment of 1/50 and C^W 3/25. Let the production amount of the managers' projects be distributed as

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{4}x & 0 \leq x < \frac{4}{5} \\ \frac{1}{5} + 4(x - \frac{4}{5}) & \frac{4}{5} \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Assume that the investment amount is $i_0 = 16/25$ and that the monitoring cost is $k = 1/25$ goods. From F above we can compute that expected production amount equals $E[\tilde{x}_1] = 4/5$. Production is first best as $E[\tilde{x}_1] = 4/5 > 16/25 = i_0$.

Let us first see which financial contracts are viable in this economy. For a debt contract the return requirement would be optimally set to $r_1^* = \arg\max\{r_1(1 - F(r_1))\} = 4/5$. With $r_1^* = 4/5$ the expected repayment on debt becomes $4/5 * 4/5 = 16/25$. As this

[20] In this case with $b_0 = k$ we have that 1.18 holds if and only if in the default state the manager extracts a surplus which is higher that the monitoring costs.

amount exactly equals the investment amount debt is just viable. For equity to be viable we must have that $s^i E[\tilde{x}_1] - k \geq s^i i_0$. This implies that the shares s^i must minimally be

$$s^i \geq \frac{k}{E[\tilde{x}_1] - i_0} = \frac{1/25}{4/25} = 1/4$$

In order to buy a share of 1/4 an amount of $1/4 * 16/25 = 4/25$ is needed. Not even C^W can afford that, so that with direct finance indeed all consumers buy debt of the manager.

In this economy the wealthy consumer, C^W, can successfully become a FI. Assume for simplicity that C^W has all the bargaining power when negotiating with the poor consumers, such that these are pushed to a level of expected consumption which they achieve which direct finance. C^W borrows from 32 poor consumers to collect 16/25 goods and buys an equity share of 1 in a project. Assume she stores the 2/25 of funds which remain after investment as collateral. At time 1 C^W will have an amount of funds of

$$\tilde{x}_1^{FI} = \tilde{x}_1 + b_0 - k = \tilde{x}_1 + 2/25$$

The face value on the deposits can be computed using equation 1.14. The deposit rate solves

$$\rho_1^d \Pr\{\tilde{x}_1^{FI} \geq \rho_1^d\} = \rho_1^d[1 - F(\rho_1^d - 2/25)] = \rho_1^d \left(1 - 1/4(\rho_1^d - 2/25)\right) = 16/25$$

The solution, $\rho_1^d \approx 39/50$, just makes the poor consumers indifferent between lending to a manager directly or investing in deposits. With a deposit rate of $\rho_1^d = 39/50$ intermediation can be computed to yield on expectation

$$E[\tilde{x}_1^{FI} - \rho_1^d \big| \tilde{x}_1^{FI} \geq \rho_1^d] \Pr(\tilde{x}_1^{FI} \geq \rho_1^d) = E[\tilde{x}_1 - \frac{35}{50} \big| \tilde{x}_1 \geq \frac{35}{50}] \Pr(\tilde{x}_1 \geq \frac{35}{50}) >$$

$$E[\tilde{x}_1 - \frac{35}{50} \big| \tilde{x}_1 \geq \frac{4}{5}] \Pr(\tilde{x}_1 \geq \frac{4}{5}) = \left(\frac{9}{10} - \frac{35}{50}\right) \frac{4}{5} = \frac{4}{25}$$

This amount exceeds $\frac{3}{25}$ $(= \frac{b_0}{i_0} r_1^* \Pr\{\tilde{x}_1 \geq r_1^*\})$ so that equation 1.17 is satisfied. The wealthy consumer will become a FI in this economy.

Financial intermediation is viable, but does it also enhance aggregate consumption? With direct finance on expectation 2/25 goods are lost since

$$E[\tilde{x}_1 | \tilde{x}_1 < \frac{4}{5}] \Pr\{\tilde{x}_1 < \frac{4}{5}\} = \frac{2}{5}\frac{1}{5} = \frac{2}{25}$$

With intermediation the expected loss becomes:

$$k + E[\tilde{x}_1 + b_0 - k | \tilde{x}_1^{FI} < \rho_1^d] \Pr\{\tilde{x}_1^{FI} < \rho_1^d\} = \frac{1}{25} + \left(\frac{2}{25} + \frac{1}{2}\frac{35}{50}\right)\frac{1}{4}\frac{35}{50}$$

This can be seen to be greater than 2/25. In this example intermediation is not beneficial.

1.4 Interpretation and Discussion of the Model

This paper has introduced a model with two types of agents: consumers/investors with wealth, and managers with a large investment project. Consumers face an agency problem when engaging in financial contracts with the managers. As a result first-best solutions get beyond reach of the agents. In some circumstances the financial markets close up making investment and production impossible. However, in the benchmark case the manager takes debt from a group of investors to finance the project. A wealthy investor may have a better option. She can become intermediary providing the equity of a manager's project while taking debt from 'poor' investors. Whether intermediation will be viable depends crucially on the amount of wealth of the intermediary. Even when intermediation will occur in the equilibrium it may lead to a higher loss of resources.

I will next interpret the concepts and the results of the model into real-world phenomena. I will also discuss several shortcomings of the model.

As was mentioned in the introduction, maybe the most obvious transcription of the financial intermediary of the model is a European investment bank which takes debt to provide equity in firms. However, with a somewhat broader definition of equity contracts also other institutions, such as for example a commercial bank, can be seen as the FI. Namely, in more general terms the equity contract of the model is some contract which allows the investor to extract more value from the manager than a debt contract at the cost of ex ante monitoring. Note, however, that in the model we endowed the investors with *all* the bargaining power when contracting. With this assumption the equity contract of the model indeed corresponds to a real-world equity contract. When, instead, also the borrower has some bargaining power the equity contract could become a debt contract, so that the analogy with a commercial bank applies. Also, changing the setting of the problem slightly by assuming only ex post costly state verification, it can be shown using the result of Gale and Hellwig (1985) that all contracts become debt.

This last, small change suggested above leads to the following discussion. Let us call the debt contract of the model D^H and the equity contract D^S. How should we interpret D^S and D^H? In case the repayment requirement on D^H is not met, the project is liquidated unconditionally. Renegotiation is not admissible so D^H is a *hard* debt contract. After a default on D^S, instead, monitoring follows and all the assets of the project flow to the debt holders. Why would not everyone opt for contract D^S in this case? The reason is again that with multiple investors the monitoring costs have to be duplicated, e.g. because information on the monitoring outcome is not reliable. Hence, with many investors precommitting to liquidation by choosing contract D^H is cheaper. Interpreting D^H in this way, D^S can be viewed as 'soft debt'. Soft debt works as follows. In case of default debt holders *renegotiate* the debt contract. Assume that in the renegotiation game the investors have all the bargaining power.[21] When renegotiating all investors incur costs k, but they get the entire production amount. In the above, the crucial assumption is again the feature that monitoring costs are duplicated with multiple investors. This would be the case if investors face costs to coordinate the renegotiation offer ('voting costs'). In reality debt renegotiations are indeed more costly when there are multiple claim holders. With duplication, poor investors will buy hard debt, and wealthy investors soft debt. Moreover, if there are many poor investors which buy debt D^H they have a credible commitment to invoke the bankruptcy procedure at time 1, if needed. The intuition

[21] This would be the case if the renegotiation game is as follows. After a default debt holders 'monitor' and propose a take-it-or-leave-it offer for repayment. The manager can either accept the settlement, or reject, in which case liquidation follows and all assets are destroyed after all.

sketched above is closely related to Berglöf and Von Thadden (1994), Hart and Moore (1995) and Bolton and Scharfstein (1996).

How to interpret the collateral amount of the FI? Take the example of an investment bank or a commercial bank. The most obvious interpretation of the collateral amount is the bank's capital. Another possibility would be the bank's risk capital. But maybe the most appealing interpretation is that collateral equals the amount of bank capital plus the discounted value of all future profits. Namely, in case of bankruptcy of a bank not only all bank capital is lost but also the 'option on future profits'.[22] However, let us stick for a moment to the interpretation of collateral as the bank's capital. The model argues that the amount of bank capital must be substantial so that the creditors of the bank have a sufficiently high chance that their loans and deposits are repaid. This seems to be in accordance with what practitioners claim is the role of capital for the FI.[23] Typically, however, the amount of bank capital is *small,* instead of 'substantial', when compared to a bank's asset portfolio. Often at least 90 percent of a bank's liabilities are debt. Several remarks can be made with respect to this. First, in reality the asset portfolio of a bank is *diversified.* Therefore, the bank's total assets bear less risk than the assets of a typical company. With a well-diversified portfolio the amount of capital can be 'small'.[24] In fact, bank capital is often claimed to be large enough to cope with the fluctuations in the bank's asset return. The size of the bank capital as recommended by the Basle Accord is precisely established so as to capture these fluctuations.[25] Second, often the size of bank capital reported in the balance sheet underestimates its true size —banks have hidden reserves. Moreover, as was argued above, it seems appealing to include discounted future

[22] And regarding the occurrence of bankruptcy, with efficient capital markets the bank should still be able to attract capital if the capital base becomes negative, but is dominated by discouted future profits.

[23] For example, Chris Matten, an executive director of the Swiss bank Corporation, argues that the size of bank capital should be set exactly according to this rule (see Matten (1996)).

[24] In Diamond (1984) the amount of capital needed can even approach zero. However, this relies on the assumption that the returns of managers' projects are independently distributed. So there is no aggregate risk.

[25] In the 1988 Basle Accord the representatives of the central banks and supervisory authorities of the G10 countries set guidelines for the amount of bank capital which is considered to be sufficient for a stable banking system. For a clear brief account of this see Matten (1996).

profits when interpreting the collateral amount. With these modifications the collateral amount comes close to the market capitalisation of the equity of the bank.[26]

There is a second and maybe more fundamental problem with interpreting the intermediary's collateral. In reality, the owners of the intermediary's capital typically are not the intermediary's managers: ownership and control are separated. Therefore, it should be questioned why the provision of the intermediary's capital is not subject to the same agency problem as the creditors face in the model. To put this criticism differently, the model seems only to apply to entrepreneurial firms and not to public companies with dispersed shareholders. Yet, large banks typically fall in the latter category. When ownership and control are separated, the model only makes sense if the incentives of the intermediary's management are for some reason sufficiently in line with the shareholders' objectives. Jensen (1986) argues that this is often not the case, but that a large debt burden alleviates the problem. The high leverage of a typical bank could therefore offer one explanation why the objectives of the managers and the shareholders of an intermediary are more or less in line with each other. The argument is not complete but worth consideration.

In the model the manager/borrower does not have a wealth endowment. In reality firms do hold wealth. What would be the consequences of this? Applying the logic of the model, wealth mitigates the agency problem attached to external funds. At some point it may be that the firm has enough wealth to use as collateral so that the services of the financial intermediary are no longer needed. The firm can then directly address investors. In reality wealthy firms with a well-established track-record can indeed apply for funds in the open market. It is also argued in, for example, the credit channel literature that notably small borrowers rely on bank loans.[27] Therefore also this prediction seems broadly consistent with reality.

---

[26]Consider the following example. ING, a Dutch bank and insurance company, reports that its balance sheet total at 31/12/1997 was DFl 620.4 billion (DFl 2 is about 1 US dollar). At the same time, its equity is reported to be DFl 46.1 billion. However, the stock market at 31/12/1997 valued ING's equity to be DFl 88.086 billion, at a share price of DFl 85.40 (ING 1997 Annual Report).

[27]Hubbard (1994) and Bernanke and Gertler (1995) are good reviews on the credit channel literature.

Allowing the borrowing firm to have wealth opens up an even more interesting discussion. When firms hold wealth there is in principle no reason why they cannot play the role of FI themselves. Firms with a high amount of wealth could in principle borrow funds and take equity in other firms. Such 'FIs' take over companies (projects). monitor, i.e. *restructure,* the targets at cost k and thus extract more value than the capital market. It is easy to conceive that certain firms can be very efficient in this activity. For example, it is quite likely that firms operating in a specific sector need to incur lower costs k to reorganise firms in the same business than agents which are not familiar with the particularities of the sector. If all this were the case we would predict that wealthy firms are prone to take over poor firms in the same sector. Interestingly. the reason for a takeover of the kind described above, is simply rent seeking: there is no reason to assume that the wealthier firm manages the target better than other investors would do, or that synergies are involved.

1.5 References

Aghion, P. and P. Bolton (1992), 'An Incomplete Contracting Approach to Financial Contracting,' *Review of Economic Studies,* **59**, pp. 473-494.

Allen, F. and A.M. Santomero (1998), 'The Theory of Financial Intermediation,' *Journal of Banking and Finance,* **21**, pp. 1461-1485.

Augustine, D.L. (1992) 'The banker in German society,' in *Finance and Financiers in European History, 1880-1960,* Y. Cassis, ed.

Berglöf, E. and E-L.Von Thadden (1994) 'Short-term versus Long-term Interests: Capital Structure with Multiple Investors,' *Quarterly Journal of Economics,* **109**, pp. 1055-1084.

Bernanke, B. and M. Gertler (1995), 'Inside the Black Box: the Credit Channel of Monetary Policy Transmission,' NBER Working Paper 5146, June.

Bester (1987) 'The Role of Collateral in Credit Markets with Imperfect Information,' *European Economic Review,* **31**, pp. 887-899.

Bhattacharya, S. and A.V. Thakor (1993) 'Contemporary Banking Theory,' *Journal of Financial Intermediation,* **3**, pp. 2-50.

Bolton, P. and D.S. Scharfstein (1996) 'Optimal Debt Structure and the Number of Creditors,' *Journal of Political Economy,* **104**, pp. 1-25.

Campbell, T.S. and W.A. Kracaw (1980) 'Information Production, Market Signalling, and the Theory of Financial Intermediation,' *Journal of Finance,* **35**, pp. 863-882.

Daunton, M.J. (1992) 'Financial elites and British society, 1880-1950,' in *Finance and Financiers in European History, 1880-1960,* Y. Cassis, ed.

De Roover, R. (1966) *The Rise and Decline of The Medici Bank 1397-1494,* New York, W.W.Norton & Company Inc.

Diamond, D. (1984) 'Financial Intermediation and Delegated Monitoring,' *Review of Economic Studies,* **51**, pp. 393-414.

Diamond, D. (1996) 'Financial Intermediation as Delegated Monitoring: A Simple Example,' *Federal Reserve bank of Richmond Economic Quarterly*, **82-3**, pp. 51-66.

Gale, D. and M. Hellwig (1985) 'Incentive-Compatible Debt Contracts I: The One Period Problem,' *Review of Economic Studies*, **52**, 647-664.

Goldsmith, R.W. (1987) *Premodern Financial Systems, A Historical Comparative Study,* First Edition, Cambridge University Press.

Hart, O. and J. Moore (1994) 'A Theory of Debt Based on the Inalienability of of Human Capital,' *Quarterly Journal od Economics*, **109**, pp. 841-880.

Hart, O. and J. Moore (1995) 'Debt and Seniority: An Analysis of the Role of Hard Claims in Constraining Management,' *American Economic Review*, **85**, pp. 567-585.

Hellwig, M. (1991) 'Banking, Intermediation and Corporate Finance,' in Giovannini, A. and C. Mayer, eds., *European Financial Integration.* Cambridge: Cambridge university Press.

Holmstrom, B. and J. Tirole (1997) 'Financial Intermediation, Loanable Funds and the Real Sector,' *Quarterly Journal of Economics*, **112-3**, pp. 663-691.

Hubbard, R.G. (1994) 'Is there a Credit Channel for Monetary Policy?,' NBER Working Paper 4977, December.

Jensen, M. (1986) 'Agency Costs of free Cash Flow, Corporate finance and Takeovers' *American Economic Review, Papers and Proceedings*, **76-2**, pp. 323-329.

Kindleberger, C.P. (1993) *A Financial History of Europe,* Second Edition, New York, Oxford University Press, Inc.

Leland, H.E. and D.H. Pyle (1977) 'Informational Asymmetries, Financial Structure, and Financial Intermediation,' *Journal of Finance*, **32**, pp. 371-387.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer and R.W. Vishny (1997a) 'Trust in Large Organisations,' *American Economic Review, Papers and Proceedings*, **87**, pp. 333-338.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer and R.W. Vishny (1997b) 'Legal Determinants of External Finance,' *Journal of Finance*, **52-3**, pp. 1131-1150.

Matten, C. (1996) *Managing Bank Capital: Capital Allocation and Performance Measurement*, John Wiley & Sons Ltd., West Sussex, 1996.

Plessis, A. (1992) 'Bankers in french Society, 1860s-1960s,' in *Finance and Financiers in European History, 1880-1960,* Y. Cassis, ed.

Ramakrishnan, R.T.S. and A.V. Thakor (1984), 'Informational Reliability and the Theory of Financial Intermediation,' *Review of Economic Studies,* **51**, pp. 415-432.

Schure, P. (1998) 'The optimal Loan Contract with Costly State Verification,' Mimeo, European University Institute, Florence.

Stiglitz, J.E. and A. Weiss (1981) 'Credit Rationing in Markets with Imperfect Information,' *American Economic Review*, **71**, pp. 393-411.

Shleifer, A. and R.W. Vishny (1989) 'Management Entrenchment: The Case of Manager Specific Assets,' *Journal of Financial Economics*, **25**, pp. 123-140.

Tobin, J. (1963) 'Commercial Banks as Creators of "Money",' *Banking and Monetary Studies*, edited by D. Carlson for the Comptroller of the Currency, Homewood, Illinois: R.D. Irwin, pp. 408-419. Also in Lewis, M.K., ed., *Financial Intermediaries* Edward Elgar Publishing Ltd., Hants 1995.

Townsend, R.M. (1979) 'Optimal Contracts and Competitive markets with Costly State Verification,' *Journal of Economic theory*, **21**, pp. 265-293.

Williamson, S.D. (1987) 'Costly Monitoring, Loan Contracts, and Equilibrium Credit Rationing.' *Quarterly Journal of Economics*, **102**, pp. 135-146.

Winton, A. (1995) 'Costly State verification and Multiple Investors: The Role of Seniority,' *Review of Financial Studies*, **8**, pp. 91-123.

Chapter 2

Asset Liquiditity, Short Term Debt and Managerial Effort: The Example of Banks

2.1 Introduction

Many firms engage in activities the quality of which is difficult to observe by outside stake holders such as the outside equity holders or the creditors. Examples of such firms are startup firms in the hightech industry, startup internet firms, and also many "monitoring firms" such as management consultants, auditing companies, rating agencies, investment banks, commercial banks, finance companies and venture capital funds. Often it takes years for the outside world to fully realise the consequences of the actions of those who control these firms.

The focus of this paper is with the financing of these firms. Due to their opaque activities the financing issue can be a difficult one. This problem becomes apparent when those in control —think of the managers or inside equity holders— largely bear the 'costs' of their effort in managing the firm, while a large share of the revenues flows to the outside stakeholders. In such a situation a free-rider problem arises and those in control exert less effort than is efficient (see e.g. Jensen and Meckling, 1976). The financing problem occurs since investors realise the presence of this free rider problem *ex ante*. i.e. when deciding whether to invest in the firm. Once the free rider problem becomes bad enough the firm will not find enough external finance to undertake its activities.

In this paper I show that the firms described above can mitigate the financing problem by combining their opaque activities with investments in *illiquid* assets, i.e. assets with a low market value and debt capacity when compared to the future surplus which they generate. In other words I show that effort and holding illiquid assets can be complementary. Interestingly, though, the theory shows that effort and illiquid assets only actually become complementary when they are financed with short-term debt.

The theory does not argue that firms that engage in opaque activities that impose large 'costs' to those in control should *always* invest in illiquid assets. Rather, the theory argues that when other mechanisms fail to bring managerial effort to an acceptable level, then investments in illiquid assets and a short-term debt structure could do the job. For example, in case of a startup internet firm it could well be that the prospect of a huge surplus when successful leaves enough scope to subject management to high-powered incentive schemes while still leaving outside investors with enough surplus. However, as I will argue below, in case of a commercial bank this is not so clear. For this example, the theory may explain why commercial banks traditionally combined the activities of screening and monitoring projects with the holding of an illiquid loan portfolio and a short-term debt structure.

One could argue that costly managerial effort is important for most firms, so that the theory should apply to many other industries. However, the amount and the quality of the effort which is exerted is normally more easy to assess, for example, simply by studying the financial statements of the firm in question. In such cases, when needed, investors can take appropriate action swiftly. Moreover, in the extreme case where effort is directly observable and enforceable, financial contracts could simply stipulate the amount of effort that management exert. By contrast, in case of the firms that are relevant to this study the amount and the quality of the effort exerted is initially not an (easily) observable variable. In addition it is assumed that at the future moment when the consequences of the effort exerted today become apparent, the outsiders cannot sufficiently punish those that were responsible.[1] Summarising, effort is initially unobservable and remains at all times unverifiable (by courts).

The leading example of the theory shall be a (commercial) bank. Banks exert effort in screening and monitoring projects and firms. The screening and monitoring activities are costly but normally do not lead to the huge surplus that may be observed in other industries and is typically a small percentage of the amount of finance that the monitored companies receive. In addition, by nature, the screening and monitoring activities are opaque. It often takes considerable time before the results of the banks' monitoring effort of today become clear. A notable example where bad monitoring was not detected for a long time was the 1997 Asian financial crisis. Many have claimed the crisis was the result of a misallocation of capital by local banks that went unnoticed for many years. Two predictions of the theory apply to banks. First, banks typically have a considerable amount of illiquid investments and, second, they depend to a large extent on short-term debt finance. I shall argue below that two important illiquid assets of a bank are its portfolio of illiquid loans and its reputational capital.

Sketch of the Model

The model distinguishes a risky and large *venture*.[2] The expected surplus of the venture depends on the (monitoring) effort exerted by the *investors* that finance the venture. Effort has to be exerted after the investment date and is assumed to be costly and unverifiable. Unverifiability implies that no enforceable contracts can stipulate the amount of effort that has to be exerted. The venture is large in that more *investors* are

[1] For example because of a situation of lack of hard evidence, a legal environment that forbids certain forms of punishment, or limited liability of the responsibles.

[2] The term 'venture' is chosen merely to suggest that it is risky. The model also distinguishes a riskfree 'project'.

needed to bring together the investment amount. In addition, the model assumes that effort cannot efficiently be shared among the agents and in particular that it is best that a single agent exerts all effort, the *monitor*.[3] It is shown that this situation gives rise to a free-rider problem. The free-rider problem can be so acute that the investors, the "open capital market", may not be prepared to finance the venture.

Besides the investors there is an *entrepreneur* who has (the idea for) a risk-free, but illiquid *project*. Although the entrepreneur can in principle attract finance to reap the surplus of the project, there may be an even better option: the entrepreneur could take on the role of monitor of the venture. The model shows that the project can serve as a precommitment to exert high monitoring effort. The model also shows that the entrepreneur can only be a more successful monitor than the open capital market if she takes short-term debt to finance the project and the venture. An interesting interpretation is that the entrepreneur founds a firm with two assets and a fragile financial structure to engage in monitoring activities: a *bank*.

Summing up the results of the model, first, in some circumstances gains are made by combining the project and the venture in a single firm. Importantly, the synergy does not arise from risk diversification, as is the case in Diamond's (1984) model of the delegated monitor, but stems from the fact that the illiquid project improves the incentives to monitor the venture. Second, the firm must issue enough short-term debt to run a *liquidation risk*, i.e. the risk that investors (inefficiently) liquidate the firm. It is also shown that firm effort normally decreases when the project becomes more liquid. This is true even though a more liquid project gives the investors more value when they resort to liquidating the entrepreneur's firm.

The intuition behind these results is relatively straightforward. An illiquid project normally has a low *debt capacity* or external finance capacity compared to the amount of future rents it generates. As a result, those in control of the illiquid project have an

[3]Thus in the model the monitor is a delegated monitor as in Diamond (1984), an agent that monitors on behalf of a group of investors. But the monitor can also simply be seen as the agent that undertakes the venture.

incentive to stay in control. This incentive can be used to precommit to exert a high level of effort in monitoring the venture. The precommitment is only credible when a construction is used that allows outside investors to punish the entrepreneur by liquidating the project upon observing that monitoring effort proved insufficient.[4] An effective construction to achieve this is short-term debt finance. The short-term debt claims mature *after* investors observe a signal with respect to the outcome of the venture, but *before* the project assets mature. Upon receiving a negative signal with respect to the venture the following happens. The debt capacity of the entrepreneur's assets becomes insufficient to repay the creditors. The creditors obtain control of the entrepreneur's assets and liquidate the project. By liquidating the project the entrepreneur loses access to the project's future rents.[5]

Asset Liquidity versus Debt Capacity

Above I have argued that a necessary characteristic of the project is its illiquidity. However, it also became clear that the model assumes that the project has a limited debt capacity. I shall argue here that asset illiquidity and a limited debt capacity often stem from the same underlying feature, namely that those who own the project do not have access to its full future rents. I will then explain that in my model the illiquidity feature of the project is modeled by embedding it within an incomplete contracting framework. Finally, I shall argue that the project can well represent a bank's portfolio of illiquid assets or its reputational capital.

---

[4]Recall that liquidation indeed hurts the entrepreneur because she loses access to the future rents attached to the project.

[5]An important step in the argument is that the creditors liquidate the entrepreneur's assets upon receiving a negative signal regarding the outcome of the venture. There are two conditions for the liquidation threat to be credible. First, the debt capacity of the bank given a bad signal should be insufficient to meet the face value of the outstanding debt. In other words, there should be a default risk. Second, with a bad signal the investors get (on expectation) the most out of the bank when they liquidate the assets. In other words, there should be a liquidation risk. In the model these conditions follow both from the assumption that the project is illiquid. In addition, the entrepreneur has no external financial resources so as to be able to bribe the creditors when liquidation is imminent. By the way, in the conclusion to th paper I shall argue that the second assumption listed above may not be needed for the results.

Above I defined an asset to be *illiquid* when its market value is low compared to the value of the future rents it generates. There are two possible reasons why an asset can be illiquid. First, the potential buyers of the project have limited access to funds (see e.g. Diamond, 1997). The second reason for illiquidity is that the buyer cannot extract the full value of the future surplus for some reason. Normally this is because the buyer will not actually *control* the asset, and those in control have some bargaining power over the future surplus. By contrast, the *debt capacity* of an asset is the amount of external finance that those in control of the asset can attract on the basis of its future rents. Again, if those in control of the asset have some bargaining power over the future surplus the debt capacity is low. It is clear that a low debt capacity and illiquidity go normally hand in hand, although there could be a difference.

In my model the assumption of a low debt capacity is conveniently modelled by adopting an incomplete contracting framework. In particular, the project surplus is divided into two components, *cash* and *assets*, and the agent who controls the project has all the bargaining power over the cash component. Cash revenues are thus not verifiable as in e.g. Berglof and Von Thadden (1994) or Hart and Moore (1995).[6] In the model the contractual incompleteness described above not only causes the debt capacity of the project to be low, but also makes the project illiquid.[7]

In reality there could be several reasons why the debt capacity of a project is low.

[6]Because cash revenues are unverifiable the firm can *divert* the produced cash without having to fear prosecution by courts. Strictly spoken, what is termed 'cash' need not correspond in reality with a projects cash-flows, just as the assets need not be a project's physical assets. Nevertheless, the empirical counterpart of 'cash' is often seen as the produced 'cash-flows' (and other liquid assets) of a project. The idea is that cash-flows can often easily be diverted by the management. It is far too simple to say —and not a very convincing argument for applying the incomplete contract framework to modern economies with sophisticated accounting standards— that diverting cash-flows is simply 'taking the money and run'. A frequently heard interpretation of diversion is that managers invest 'free cash flow' in negative net present value projects (see e.g. Jensen, 1986). Other forms of diversion are discussed in i.a. Myers and Rajan (1998). For early work on incomplete contracts see Hart and Moore (1988) or Aghion and Bolton (1992).

[7]Namely, the investors, i.e. the potential buyers of the project, have limited wealth endowment by assumption. The maximum conceivable market value of the project thus becomes the debt capacity plus the wealth endowment of the wealthiest investor.

First, more external finance would lead to adverse managerial incentives.[8] There are many possible examples of such adverse incentives. In the context of banking the literature often mentions risk-shifting incentives or 'gambling for resurrection'. A small variation is a situation in which investors have to incur certain costs when seizing the rents.[9] Second, it can be that the rents are control rents.[10] Third, there may be asymmetric information regarding the amount of future rents. The firm may have information indicating that rents will be high but cannot reveal this credibly to the market.[11] Fourth, the firm's management cannot commit its human capital to the firm.[12] Finally, it may be that regulation makes it impossible to borrow based on the full extent of the value. In the context of banking this can be important because the Basle guidelines present a limit on the amount that can be borrowed.[13]

A typical bank holds a substantial amount of illiquid assets and has a limited debt capacity (see any survey article on banking, e.g. Bhattacharya and Thakor, 1993). The limited debt capacity arises when more borrowing induces risk-shifting incentives or because investors may not know the true (high) value of the loan portfolio. Also, a bank's debt capacity may be limited simply because of regulation. A counter argument is that the possibility of securitization shortens the actual maturity of the loan portfolio and increases the liquidity. Though this will be true to a certain extent, securitization will only be successful if investors are convinced that they do not buy the bad loans ('lemons') from the bank. As a consequence a bank can therefore securitise only a part of the loan

[8]See e.g. Jensen and Meckling (1976) or Myers (1977). For some examples of adverse incentives in the contect of banking see e.g. Flannery (1994) or Calomiris and Kahn (1991).

[9]For instance because of *costly state verification* as described by Townsend (1979), Gale and Hellwig (1985) or Schure (1998).

[10]I.e. utility that the manager obtains when operating the firm and which cannot be transferred to outside investors. For a model that stresses the importance of control rents, see e.g. Aghion and Bolton (1989). See also Jensen (1986).

[11]See e.g. Diamond (1991).

[12]In Hart and Moore (1994) and Diamond and Rajan (1999) the project is specific to the human capital of the manager, but contracts cannot bind the manager to the project. By threatening to quit, the manager obtains bargaining power over the future surplus.

[13]The Basle Accord limits the amount a bank can borrow to a maximum part of the bank's book value. Notice that the book value of the assets often underestimates the market value which is the relevant variable for investors. The Basle guidelines can therefore be quite restrictive.

portfolio in one go, or must have built up a reputation to sell high quality assets.

This brings me to another illiquid bank asset that sometimes is not recognized as such, namely the bank's reputation (and its brandname). In a different context Boot, Greenbaum and Thakor (1993) stress the importance of reputational capital for banks. A bank's reputational capital and its brandname are clearly long-term assets that cannot be easily sold. Also, if investments in reputational capital are substantial they can normally not be financed with debt. This is because most book keeping standards do not allow for positive valuations of the bank's brand name or reputation. Debt finance would thus leads to a negative book value of the bank's equity. Clearly the Basle norms forbid this.

I have just argued that in the context of banking the project can well represent the bank's loan portfolio or its reputational capital. Thus, for a bank the model predicts that the bank's investments in illiquid loans and reputational capital induce the bank managers to exert high effort in screening and monitoring projects.

Literature and Setup

The theory presented in this paper shows that exerting effort and holding illiquid assets are complementary, but only when financed with short-term debt. Applied to banks the theory says that banks precommit to exert high monitoring effort by investing in illiquid projects and choosing a short-term debt finance. There is a new and growing strand in the banking literature that aims to explain why banks traditionally invested in long lived assets, while relying on short-term debt, often deposit, finance. To state the focus of this strand of the literature in the terminology of Bhattacharya and Thakor (1993)'s survey article, this new literature argues that there are synergies between certain *asset services* that banks banks traditionally offered and certain *liability services.* Specifically, this literature stresses that short-term debt finance, and in particular deposit finance, has positive incentive effects on those who control the bank, either the shareholders or the bank managers.

I shall relate my paper to this literature in the next section.[14] In Section 2.3 I present some data which sketch to what extent banks mismatch by borrowing short-term and investing long-term, one of the predictions of the theory. I find that, unsurprisingly, the banking sector is generally mismatched. I also show that almost every individual bank mismatches. Section 2.4 contains the model and gives a simple example. Section 2.5 discusses the results and concludes.

2.2 Relationship with the Banking Literature

Myers (1977) argues that firms, i.e. their owners, face future options to invest in risky investment opportunities. If such firms hold debt, investment opportunities with negative net present value (NPV), but with high risk may still be undertaken. The reason is that debt holders may carry a substantial amount of the losses when project outcomes are bad, whereas equity holders benefit more from the better outcomes. *Vice versa,* investment opportunities with positive NPV, but with low risk, may not be undertaken because too large a part of the benefits flows into the pockets of the debt holders. Flannery (1994) argues that the picture that Myers sketches is particularly alarming for banking firms. He asserts that the activities of bank managers are typically non-verifiable so that direct contractual solutions to the 'risk-shifting' problem of Myers are not feasible. However, Flannery argues, short-term debt finance remains a remedy. Short-term debt is repriced when rolled over. So if the firm adopts too risky projects this will translate in more expensive debt when the debt claims are rolled over. Thus, through the frequent repricing of short-term debt the shareholders pay the price of engaging in too risky projects. My model is close in spirit to Flannery who mentions monitoring, project selection, asset

[14] Notice, though, that of course this banking literature is inspired strongly by studies that deal with firms' debt maturity choice more generally. The relevant seminal article for the banking literature I focus on is Myers (1977). Not unrelatedly, Diamond (1991) focuses on the firm's maturity choice when those in control are better informed on the firm's prospects than the investors. There is also literature that focuses on tax effects when explaining the mismatch of a firm (See e.g. Brick and Ravid, 1985). For an empirical study on the debt maturity choice, see Barclay and Smith (1995).

substitution, etc., as the bank's main activities. However, I study the interesting limit case of Flannery in which repricing short-term debt would *not* provide sufficient incentives to undo the risk-shifting incentives.[15] My model shows that in such cases investments in illiquid assets can provide additional incentives to monitor. Flannery's prediction regarding the holding of illiquid assets are precisely be the opposite: banks should strive not to hold illiquid assets at all as this induces inefficient liquidation of assets when new short-term debt is priced too high. In fact Flannery sees the liquidation risk attached to issuing short-term debt as a downside because it leads to inefficient liquidation.[16] By contract, although in my model liquidation is *ex post* inefficient, it is a necessary because of the positive *ex ante* effects of the threat of liquidation.

Calomiris and Kahn (1991) and Qi (1997) argue that banks use deposit finance in order to create the right incentives for the depositors to monitor the work of the bank manager. The focus of Calomiris and Kahn is on explaining the form of the deposit contract. In their model the bank manager has opportunities to cheat on the investors by absconding part of the asset return of the bank. They show that with debt claims with a fixed face value, the bank manager has more incentives to abscond if the asset returns are low. Demandable debt finance may reduce the costs of absconding. Before the asset return is realised demandable debt holders receive a signal on the outcome and hence also the likelihood of future absconding of the bank manager. On the basis of this signal they may decide to demand their debt. If a substantial number of creditors demand their debt, the bank is liquidated. Liquidation is costly, but less costly than absconding. Also in Qi deposit withdrawals trigger liquidation and prevent bad action on part of the bank manager from happening. In Qi the bad action is allowing non-monitored loan applicants to borrow the bank's funds. Since the bank manager realises *ex ante* that no monitoring triggers bankruptcy, she will exert enough effort when monitoring the loan applicant. The

[15]In fact, the prediction of Flannery's analysis in my setting would be that the most wealthy investor would become banker. Repricing short-term debt works best for the most wealthy investor, because others are more often protected by limited liability.

[16]Notice, however, that in Flannery it is not clear why investors should ever liquidate the bank. Renegotiation of the debt contracts should always dominate actual liquidation.

most important difference between my model and those of Calomiris and Kahn and Qi is one of timing. In my model outsiders can only assess in *retrospect* that management engaged in bad action from the investors' point of view. As a consequence, outside investors *cannot* directly prevent bad action from happening. The outsiders' problem, then, becomes to design the structure of the bank so as to minimize the incentives for management to engage in bad action. Investments in illiquid investments and financing with short-term hard claims serve this purpose.

The same timing issue also marks the main difference between my paper and the intimately related paper of Diamond and Rajan (1999). Diamond and Rajan have a model in which they address how to finance a portfolio of illiquid loans that is best managed by a relationship lender. In their model loans are illiquid because the relationship lender cannot commit to employ her superior skills to extract the maximum possible return from the loan. Diamond and Rajan show that in this setting finance with demandable deposits serves as a precommitment of the relationship lender to employ her extraction skills. The argument is the following. Once the relationship lender is inclined not to use her extraction skills, depositors seize the loan to deal directly with the borrower (however, they retain the possibility to hire the relationship lender). In this way the relationship lender loses access to all her rents. Preventing this from happening she puts in her extraction skills in the first place. Again, in my model depositors do not have sufficient information to realise that the entrepreneur cheats and thus cannot withdraw and disintermediate the entrepreneur on time. In my model it is the threat to lose the illiquid asset that creates the bank's precommitment to behave diligently. The informational assumptions of my model is thus very mild: it is only expected from the depositors that they observe whether the illiquid asset is still there.

Myers and Rajan (1998) explicitly address why deposit taking banks, or more generally firms with liquid liabilities, traditionally ran a liquidation risk by investing in illiquid, often long-term assets. They show that, paradoxically, if the bank held only liquid assets short-term depositors would have less protection to receive back their funds. The reason is their that bank managers have greater ease to abscond the liquid assets than illiquid assets. Thus, liquid investments have an advantage —namely that they yield more when liquidation is needed— but also a disadvantage —namely that managers abscond more often. My model can be seen as complementary to Myers and Rajan in that it points to a second possible advantage of investments in illiquid assets. The advantage in my model is that illiquid assets give the banker a comparative advantage to screen and monitor projects.

With respect to banks it is often heard that maturity transformation —i.e. borrowing short-term ('issuing deposits') and lending long-term— is 'the nature of a bank's business'. Since the term structure of credit is normally upward sloping the bank makes money. Therefore, it is argued, it is obvious that banks are mismatched. I believe there are several reasons why such an explanation is incomplete. First, if the reasoning above presented the complete picture, it is not clear why banks combine the activities of maturity transformation and monitoring projects. As an alternative deposit banks could lend out their funds (long-term) to finance companies who specialise in the monitoring job. Second, and related, according to this line of thought we should observe banks to invest in long-term liquid assets and not in illiquid loans or reputational capital. In this way banks reap the benefits from the upward sloping term structure, while avoiding the costs of costly liquidation. Third, if enough banks mismatch —so that the market for mismatching becomes competitive— the upward-sloping term structure should merely represent a liquidation risk premium. We would then expect to observe a substantial number of banks who choose not to engage in the zero net present value business of mismatching. However, we shall see in the next section that not only the banking sector as a whole mismatches, but in fact almost all individual banks. Finally, if maturity transformation were the only explanation for observing the mismatch, it is not clear why, in case of a default, a bank should be liquidated. Also, it is a puzzle why many bank failures in history were due to fraud and mismanagement,[17] rather than 'bad luck' (i.e.

[17]See Calomiris and Kahn (1991) and Myers and Rajan (1998).

Mismatch in 1997	# Banks
-0.5	1
-0.45	0
-0.4	0
-0.35	0
-0.3	0
-0.25	5
-0.2	2
-0.15	2
-0.1	3
-0.05	6
0	16
0.05	18
0.1	16
0.15	17
0.2	15
0.25	20
0.3	32
0.35	40
0.4	53
0.45	44
0.5	42
0.55	34
0.6	30
0.65	23
0.7	30
0.75	28
0.8	11
0.85	5
0.9	1
0.95	4
1	0
Number of Banks	506
Missing Observations	8

Figure 2-2: Histogram of the maturity mismatch of assets and liabilities: EU banks.[a]

Mismatch in 1997	# Banks
-0.5	0
-0.45	0
-0.4	0
-0.35	0
-0.3	1
-0.25	1
-0.2	1
-0.15	3
-0.1	1
-0.05	4
0	4
0.05	1
0.1	6
0.15	6
0.2	4
0.25	8
0.3	26
0.35	28
0.4	60
0.45	58
0.5	94
0.55	81
0.6	39
0.65	20
0.7	10
0.75	3
0.8	0
0.85	0
0.9	0
0.95	0
1	0
Number of Banks	459
Missing Observations	0

Figure 2-3: Histogram of the maturity mismatch of assets and liabilities: US/Can banks.

bad investment outcomes or excessive liquidity needs of the depositors).

2.3 Liquidity Risk: The Example of Banks

Figure 2.1 and 2.2 give an idea to what extent banks are mismatched. In particular, the figures contain histograms of a 'mismatch ratio' that was computed on the basis of balance sheet data of banks in the European Union (EU) and in the US and Canada (US/Can), respectively. The data that was used to compute the mismatch ratio's was taken from *BankScope* of Bureau van Dijk Electronic Publishing. BankScope is a dataset with financial data of banks over the world. The coverage for the EU and US/Can is very good. In BankScope I selected 'large' Commercial Banks, Savings and Cooperative Banks, Real Estate & Mortgage Banks and Investment Banks & Security Houses.[18] This left me with 506 EU banks and 459 US/Can banks.

The mismatch ratio is a rough measure that is computed by taking the amount of short-term finance of a bank, subtracting the short-term assets and dividing the result by the bank's balance sheet total. The short-term finance of a bank is computed by adding up the BankScope variables 'money market funding' and 'total deposits'.[19] The short-term assets were taken to be 'cash and due from banks' and 'total other earning assets'.[20] By construction the mismatch ratio is a number between minus one and one. It is minus one if the bank is entirely financed long-term, but all the assets are short-term. The ratio is one if the bank is entirely financed short-term, but only has long-term investments.

Figure 2.1 and 2.2 give the histograms of the mismatch ratio in the year 1997.[21] To

[18]The selected banks were also 'living banks' with reports in 'raw data format' and consolidation codes C1, C2, C* and U1. I downloaded the balance sheets of these banks for the years 1993-1998, but deleted afterwards all banks for which data was missing for three years or more. In this deletion procedure many North-American Investment Banks & Security Houses were deleted.

[19]Money market funding: e.g. certificates of deposit, commercial paper and debt securities.

[20]Total other earning assets: e.g. deposits with banks, due from central banks, due to other banks, total securities, T-bills, bonds, certificates of deposit, equity investments and other investments.

[21]The figures only give 1997 data. I also computed histograms for 1993-1996 and 1998 but they all looked very similar. I picked the year 1997 as in 1998 many observations were missing. I have also computed a mismatch ratio where it was assumed that in every year a fifth of the long-term assets

explain the histograms by an example, the number 42 at a mismatch ratio of 0.5 in Figure 2.1 means that in case of 42 EU banks about half[22] of the balance-sheet consists of long-term investments which are short-term financed. In Figure 2.1 it becomes clear that EU banks are *en masse* mismatched. Only 69 of the 498 banks for which we have an observation, have a mismatch ratio of less than 0.10. We also see that a typical bank in the EU has a mismatch ratio 0.30 and 0.60. In the US/Can the picture is even more pronounced: just 22 of the 459 banks match more or less the maturities of their assets and liabilities, and roughly 50 percent of the assets of a typical bank are long-term assets which are short-term financed.

I would like to draw two conclusions from the evidence presented above. First, the banking sector is mismatched to a great extent and, second, almost all individual banks mismatch.

2.4 The Model

Introduction

Consider a competitive financial market with risk-neutral, small *investors*[23] and let for simplicity the competitive expected rate of return be zero. Time is divided into two periods with relevant dates date 0, date 1 and date 2. There is also a risk-neutral *entrepreneur*, who owns the idea for a *project* but has no funds endowment. The project requires an investment amount at date 0 and yields a (non-random) return at date 2. I assume the project is illiquid, i.e. separating the assets from the entrepreneur and selling

and liabilities mature. It is not shown because it looked very similar to the histograms in the figures. Finally, I tried to compute a ratio that expresses the amount of non-demandable assets which is financed by demandable debt. This attempt failed, however, due to lack of appropriate data.

[22]To be precise, between 45 and 50 percent of the balance sheet.

[23]As we will see below investors are small in that their wealth endowment is small compared to the amount that the investment opportunities require. Shortly the meaning of 'small investor' becomes more precise.

them at date 1 yields less value than waiting till maturity at date 2.[24] I also assume that the project return can be decomposed in a *cash* component and an *asset* component. This distinction becomes important below when assuming that contracts are incomplete.

There is also a *venture* which yields a *risky* asset return, but generates no cash. The venture is big in that no single investor is able to finance alone the investment amount needed. However, every group of investors that collects the investment amount is free to invest in the venture.[25] I assume that the return to the venture depends on (monitoring) *effort* of the investors. I also assume that effort implies disutility and that, if more investors monitor, they fully duplicate their efforts.[26] This last assumption implies that monitoring will be delegated to a single agent: the lead investor, or *monitor*. In summary, the monitor is a delegated monitor as in Diamond (1984) or Winton (1995).

The venture is the object of study in the model. In particular, the model studies whether the venture is undertaken, and, if so, how is it financed and who is the monitor. The results will show that in some circumstances the entrepreneur is the most efficient monitor. Using the project, the entrepreneur can precommit to exert a high level of monitoring effort, whereas the market cannot. I show that a necessary condition for this is that the monitor takes short-term debt finance so as to run the risk to be liquidated.

Financial contracts are incomplete in two distinct ways. First, contracts cannot be conditioned on the monitoring effort which is exerted by the monitor, i.e. monitoring is not verifiable. A reason why monitoring effort is not verifiable could be, for example, that outsiders do not observe the monitoring effort exerted. Another possibility is that effort is observable, but cannot be enforced by 'courts' due to lack of evidence. Second, contracts cannot be conditioned on the borrower's cash return. The borrower can divert cash-flows

[24] Possible reasons for the illiquidity assumption are listed later on when a detailed description of the project is given.

[25] As more groups of investors could be interested in undertaking the venture, it may be useful to think of a situation with many identical ventures.

[26] Full duplication of monitoring is an extreme assumption, but it is merely made for simplicity and could be relaxed. However, I do need to assume that the total monitoring cost increases in the number of monitoring investors. This would be true if investors face costs to coordinate their efforts ('voting costs').

without the fear of being prosecuted by a court. These two sources of incompleteness leave us with two relevant contracting variables, first, the amount of assets held by a borrower and, second, the *ownership* of the borrower's assets. Ownership of an asset gives the right to *control* it, which here simply means the option to liquidate the asset.

Both sources of contractual incompleteness potentially lead to problems. Starting with the latter; since cash can be diverted by the borrower, she cannot credibly promise to repay the investors up to the full amount of rents an investment project generates. This could make the debt capacity of the investment project lower than the investment amount, in which case investors do not finance the project, even though the total rents may exceed the investment amount. In the model, this problem is only relevant to the project since the venture generates only assets, and no cash. By contrast, the monitoring feature only plays a role for the venture because effort is not an input in the project. The problem with monitoring is that it results in a classical free-rider problem: The monitor fully bears the cost (i.e. disutility) of monitoring, while all investors share in the benefits.

In the model financial contracts can have all conceivable forms, but it can be shown that in our incomplete contracting framework we can focus on *debt contracts* without loss of generality. Since investment takes place at date 0, the most general debt contract has the form (L_0, D_1, D_2), where L_0 is the amount borrowed at date 0 and D_1 and D_2 are the stipulated repayments at date 1 and date 2, respectively. If a repayment D_i is not met, the borrower loses ownership of her assets to the lender.[27] In the model I will assume for simplicity that contracts are either *short-term* —i.e. of the form $(L_0, D_1, 0)$ or $(0, L_1, D_2)$; or *long-term* —i.e. of the form $(L_0, 0, D_2)$. I also assume $L_0 \geq 0$, $D_1 \geq 0$ and $D_2 \geq 0$.

I next present the details of the venture. I then study whether *open market finance* is viable, i.e. I assess in which case the investors get together to undertake the venture. We shall see that the venture is viable when the monitoring investor is wealthy enough. After

[27]The debt contract could also have sophisticated entries such as the seniority of the claim D_2 with respect to other possible claims. However, in the model these issues are not important.

that I give a detailed description of the project of the entrepreneur. In the main part of the section I derive whether the entrepreneur can take on the role of monitoring the venture. It turns out that in some circumstances the entrepreneur only undertakes the project. However, in other circumstances the entrepreneur also takes on the venture to become monitor. It cannot be the case that the entrepreneur only invests in the venture. Interestingly, if the entrepreneur monitors the venture she must take short-term debt and run the risk to be liquidated. Phrased in other words; the project and the venture are complementary, but only if they are financed with short-term debt claims. The section is concluded with an example of the model. In this example the market is not interested in investing in the venture, while the entrepreneur is.

The Venture

The venture is embedded in the simplest possible setting which encompasses its two relevant characteristics, namely that it is risky and that the return depends on monitoring effort.[28]

The venture requires an investment amount of I_0 at date 0. At date 2 the venture matures. The return is random: with probability $Q(e)$ the venture becomes a *success* and yields a (verifiable) return of A_2 assets; with probability $1 - Q(e)$ the venture ***fails*** in which case nothing is produced. The probability of a success is a function of the effort level e of the monitor. I assume throughout the paper that $Q(.)$ is a continuous function

$$Q \in C^0(\Re_+, [0,1]) \tag{2.1}$$

I shall also assume that there is a threshold monitoring level $\underline{e} \geq 0$ below which the success probability of the venture is zero. When effort exceeds the threshold $\underline{e}$, monitoring strictly

[28] In an earlier version of the paper I adopted a much more complex setting. A *manager* managed the venture and contracting between the investors and the manager was incomplete. While this setting was certainly more consistent —notice that contracting between the banker and the investors is also embedded in an incomplete contracting framework— it did not add to the understanding of the paper while complicating matters considerably. I have thus chosen this simpler setting.

increases the success probability, but at a decreasing rate.

$$Q(e) = 0 \qquad 0 \leq e \leq \underline{e} \tag{2.2}$$

$$Q'(e) > 0 \text{ and } Q''(e) < 0 \qquad e > \underline{e} \tag{2.3}$$

In the next few subsections I maintain assumptions 2.2—2.3. For the example of the model that concludes the section it comes handy to modify 2.2—2.3 slightly. In particular, I shall assume there that there is a monitoring level $\overline{e} > \underline{e}$ after which the success probability stays constant

$$Q(e) = Q(\overline{e}) \qquad e \geq \overline{e} \tag{2.4}$$

Finally, in this paper I will focus only the nontrivial case that the venture is undertaken in a first best world.

$$\text{There is an } e \text{ such that } Q(e)A_2 - e > I_0 \tag{2.5}$$

At date 1, the investors observe a signal about the outcome of the venture. In the simple setting that I adopt, the signal is perfect. In other words, investors costlessly observe at date 1 what the date-2 return of the venture will be. The assumption that the signal is perfect is made for simplicity and could be easily generalised. We shall see that the signal becomes important when the entrepreneur becomes the monitor of the venture. Figure 2.1 summarises the relevant aspects of the venture.

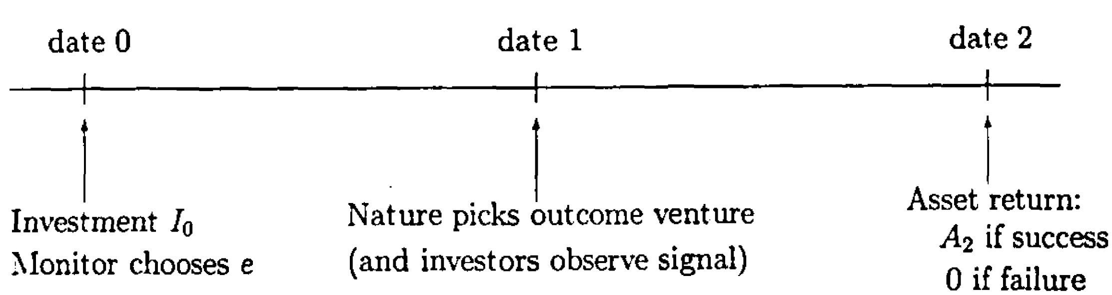


Figure 2-1: The venture

Open Market Finance

Consider a generic investor, I^M say, with wealth endowment fI_0. Here f stands for the fraction of the venture that I^M alone can finance. We assume that investors are small

$$f < 1$$

I^M thinks about offering a debt contract $(L_0, 0, D_2)$ to her fellow investors, I^N say, in order to undertake the venture.[29] A natural restriction on the loan amount is $L_0 \geq (1-f)I_0$. However, in this model there is no reason for overborrowing, so assume $L_0 = (1-f)I_0$ without loss of generality. Also assume without loss of generality that $D_2 \leq A_2$. The contracting game looks as follows.

Date 0

1. I^M offers $(L_0, 0, D_2)$, $D_2 \leq A_2$.
2. I^N accepts or rejects $(L_0, 0, D_2)$. If I^N rejects the game ends and the payoffs of I^N and I^M become $(0, 0)$. The entries of this *payoff vector* represent I^N's and I^M's payoff, respectively. If I^N accepts the contract, the project is started.
3. I^M chooses e.

Date 1

Nature picks the date-2 outcome of the venture. The players observe the (perfect) signal.

Date 2

The game ends. In case the venture is a success the payoffs become $(D_2 - L_0, A_2 - D_2 + L_0 - I_0 - e)$. In case the venture failed we get $(-L_0, L_0 - I_0 - e)$.

[29] Two clarifying remarks. First, I focus here on long-term contracts only because in the next financing game the difference between short-term contracts and long-term contracts is immaterial. Second, I treated the group of other investors as a single party. This can be done because the contracting game below is such that their incentives are perfectly aligned.

The equilibrium concept I adopt is subgame perfect Nash in pure strategies. We find the equilibrium by backward induction and we are interested in whether in equilibrium the venture is undertaken.

Definition 3 *"The open capital market is willing to undertake the venture" if there is at least one investor I^M with wealth endowment fI_0 such that in equilibrium of the contracting game above I^M offers a contract $(L_0^*, 0, D_2^*)$, say, which I^N subsequently accepts (so that the venture is undertaken).*

In stage 3 of date 0 I^M chooses effort e so as to maximise expected payoff given the contract $(L_0, 0, D_2)$.

$$e \in \arg\max_{\tilde{e}} \{Q(\tilde{e})(A_2 - D_2) - \tilde{e} + L_0 - I_0\} \tag{2.6}$$

Moving to stage 2, I^N accepts the contract $(L_0, 0, D_2)$ when receiving the market rate of interest, or more on the loan amount L_0

$$Q(e)D_2 - L_0 \geq 0 \tag{2.7}$$

In stage 1 of date 0 I^M chooses the contract so as to maximise expected profit, i.e.

$$Q(e)(A_2 - D_2) - e + L_0 - I_0 \tag{2.8}$$

However, I^M only considers contracts that deliver weakly more payoff than not offering a contract at all (or one that is trivially rejected such as $(L, 0, 0)$). Therefore

$$Q(e)(A_2 - D_2) - e + L_0 - I_0 \geq 0 \tag{2.9}$$

We have proved the following lemma.

Lemma 4 *Let $L_0 = (1 - f)I_0$. The open market is willing to undertake the venture if and only if there is an investor, I^M with a wealth endowment fI_0, such that Program P below has a feasible solution*

$$\max_{e, D_2} Q(e)(A_2 - D_2) - e + L_0 - I_0 \qquad \textit{(Program P)}$$

$$s.t\ Q(e)(A_2 - D_2) - e + L_0 - I_0 \geq 0$$

$$e \in \arg\max_{\tilde{e}} \{Q(\tilde{e})(A_2 - D_2) - \tilde{e}\}$$

$$Q(e)D_2 - L_0 \geq 0 \qquad D_2 \leq A_2$$

The solution to Program P depends, in principle, on the monitor's wealth endowment as summarised by the fraction f. Specifically, the next proposition shows (i) that Program P has a unique solution if and only if the fraction f exceeds a certain threshold $\underline{f}$ and (ii) that if Program P has a solution then the effort I^M exerts is smaller than the efficient level of effort. It is also shown that I^M's effort increases in f. An obvious corollary of the proposition is that the wealthiest investor is the best monitor.

Proposition 5

(i) If there is a solution to Program P it is unique and it satisfies

$$Q'(e)(A_2 - D_2) - 1 = 0 \text{ and } Q(e)D_2 - L_0 = 0$$

(ii) Assume there is a solution to Program P for $f < 1$. Any feasible solution (e, D_2) of Program P, and in particular the optimum (e^*, D_2^*), satisfies $e < e^{FB}$ where e^{FB} uniquely solves $Q'(e^{FB})A_2 - 1 = 0$.

(iii) Assume Program P has a solution (e^*, D_2^*) and take A_2 fixed. In a small enough open neighbourhood of (e^*, D_2^*) we have that e and D_2 are implicitly defined as functions of the fraction f with

$$\frac{de^*}{df} > 0 \text{ and } \frac{dD_2^*}{df} < 0$$

(iv) There is a $\underline{f} < 1$ such that Program P has a solution if and only if $f \geq \underline{f}$.

Proof. See Appendix.

Corollary 6 *Let $\underline{f}$ be as defined in the proposition above. The open market is willing to finance the venture if and only if there an investor with wealth endowment $\overline{f}I_0$, say, such that $\overline{f} \geq \underline{f}$.*

In the remainder of the paper I focus on the interesting cases in which the incentives to monitor form the key factor that determine whether the venture will be financed or not. In particular, I assume

$$\text{There is no feasible pair } (e, D_2) \text{ s.t.} \tag{2.10}$$

$$e \in \arg\max_{\tilde{e}} \{Q(\tilde{e})(A_2 - D_2) - \tilde{e}\} \text{ and}$$

$$Q(e)D_2 - I_0 \geq 0$$

Notice that from Assumption 2.10 it trivially follows that Program P has no solution for $f = 0$. Combining this fact with Proposition 5 (iv) it follows that the threshold fraction $\underline{f}$ must be strictly greater than zero

$$\underline{f} > 0$$

Assumption 2.10 has an interesting intrepretation in what could be called the *debt capacity* of the venture, i.e. the maximum possible amount of funds that the investors are prepared to lend to the monitor. Specifically, Assumption 2.10 says that the debt capacity of the venture is smaller than the investment amount I_0.

The Project

The project requires an investment amount i_0 at date 0. At date 2 the project assets mature to deliver c_2 units of cash and a_2 units of assets. At date 1 the project assets can also be *liquidated*, i.e. separated from the entrepreneur and sold. Let the market value of the assets at date 1 be a_1. In other words a_1 represents the liquidation value of the project at date 1.[30] The details of the project are summarised in Figure 2.2. Notice that the surplus of the project is:

$a_2 + c_2 - i_0$	if the project is not liquidated
$a_1 - i_0$	if the project is liquidated (at date 1)

[30]Likewise, if the project had been embedded in a 3-period setting, instead of a 2-period setting, a_2 could have been interpreted as the liquidation value of the project at date 2.

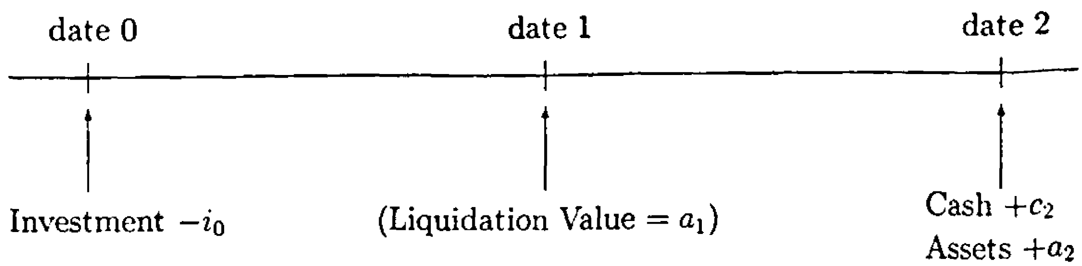


Figure 2-2: The cashflow of the project

I assume that the project assets are *illiquid*, i.e. the liquidation value of the project at date 1 is lower than the net present value of the cashflows

$$a_1 < a_2 + c_2 \tag{2.11}$$

Notice that the fraction $\beta \equiv \frac{a_1}{a_2+c_2}$ can be interpreted as a measure for liquidity. The higher β, the higher is the date 1 liquidation value compared to the NPV of the project.[31] I also assume

$$a_1 > a_2 \tag{2.12}$$

This assumption becomes important later on when the entrepreneur decides to set up a firm to invest in both the project and the venture. We shall see that with assumption 2.12 the short-term creditors of this firm turn out to have a credible threat to liquidate the firm upon receiving a bad signal with respect to the venture at date 1. Finally, I make two assumptions that limit the size of the project when compared to the venture

$$a_1 - i_0 < \bar{f} I_0 \tag{2.13}$$

$$a_2 + c_2 - i_0 < I_0 \tag{2.14}$$

[31] Notice that assuming the project to be illiquid is natural in the spirit of the model. Specifically, in the model the project has a low debt capacity because of contractual incompleteness. Hence, when the project is liquidated (i.e. sold) at date 1 the potential buyers (i.e. the investors) can maximally offer to pay their wealth amount plus the debt capacity of the project. To be precise, it is shown later that at date 1 the debt capacity of the project is a_2. Therefore, if the project were auctioned off at date 1 the maximum possible bid would be the wealth amount held by the wealthiest investor, $\bar{f} I_0$ say, plus a_2. Needless to say, we could well have that $\bar{f} I_0 + a_2 < a_2 + c_2$. In this case the project is illiquid.

In equation 2.13 $\overline{f}I_0$ is the wealth endowment of the wealthiest investor and we shall see later that a_1 is the debt capacity of the project. Therefore, the left-hand side represents the 'wealth equivalent' of the entrepreneur's idea for the project. The assumption says that the wealthiest investor has more (verifiable) wealth that the entrepreneur. The second assumption, which is merely made for simplicity, needs no explanation.

The Entrepreneur Game

In the next two subsections I will show that in some circumstances the entrepreneur sets up a firm with two assets —namely the project and the venture— and issues short-term debt. I have argued in the introduction that a possible interpretation of such a firm is a commercial bank. The present subsection presents the relevant financing game, the 'entrepreneur game'. The entrepreneur game is solved in the next subsection.

The desciption of the entrepreneur game involve a lot of details, so let me start with a broad overview. At date 0 the entrepreneur (E) first offers a debt contract, say k, to the investors. As becomes clear below, the incentives of the investors are perfectly aligned so they can be treated as a single agent (I^N). Regarding the contract we can assume without loss of generality that there is no overborrowing. Therefore, the loan amount becomes $L_0 = i_0$ if E only undertakes the project and $L_0 = i_0 + I_0$ if E undertakes both the project and the venture. It follows directly from equation 2.10 that undertaking the venture only is not an option. If E wishes not to invest, let the contract be $k = (0, 0, 0)$. In the second stage at date 0 I^N either accepts or rejects the contract offered by E. If I^N rejects, or if the contract is $k = (0, 0, 0)$, the option to invest is forgone so that I^N and E both attain zero payoff.

If the investors accept the contract, there are four different possible scenarios and each is represented by a particular subgame. In 'Subgame 1', only the project is financed, and with long-term debt. In Subgame 2 only the project is financed, but with short-term debt. In Subgame 3 both the project and the venture are financed with long-term debt. In Subgame 4 both the project and the venture are financed with short-term debt.

In Subgame 2 and Subgame 4, E offers a short-term debt contract $k = (L_0, D_1, 0)$. At date 1 when this debt claim matures, E can propose an alternative arrangement $k' = (L_0, D_1', D_2)$ instead of attempting to repay D_1. For example, when choosing $D_1' = 0$ E proposes to 'roll over' the outstanding debt claim D_1. In fact, it turns out that without loss of generality it can be assumed that E indeed proposes to roll over.[32] I^N can either accept the alternative arrangement, or reject it in which case E's assets are liquidated. In case of liquidation the game ends. I^N then receives the liquidation proceeds up to D_1, while E receives any possible leftovers.

At date 2 E is also allowed to try renegotiate the repayment requirement D_2. Specifically, E first proposes to repay P_2. The proposal P_2 has to satisfy E's budget requirement so we have $P_2 \leq a_2 + c_2$ in case the venture failed or was not undertaken and $P_2 \leq a_2 + c_2 + A_2$ if the venture was undertaken and became a success. If $P_2 = D_2$ the game ends. If, instead, $P_2 < D_2$ ('E defaults') I^N moves next. I^N can accept P_2, or she can reject it in which case E's assets are liquidated. In case of liquidation I^N gets the verifiable proceeds up to a maximum of D_2, while E receives any possible leftovers.

The solution concept I shall adopt is subgame perfect Nash in pure strategies. While solving the game I will keep an eye on two questions: When is E *willing* to undertake the venture?; And when does E *outperform* the open capital market in monitoring the venture?

Definition 7 *"E is willing to undertake the project and the venture" if, in equilibrium of the entrepreneur game, E offers a contract k^* with $L_0^* = i_0 + I_0$, which I^N subsequently accepts (so that both the project and venture are undertaken). "E is willing to issue long-term (short-term) debt to undertake the project and the venture" if the contract k^* above is a long-term (short-term) debt contract.*

Definition 8 *(outperformance) "E outperforms the open capital market (in monitoring the venture)" if (i) E is willing to undertake the venture and either (iia) The open capital*

[32]The alternative would be to liquidate (part of) E's assets, for example by securatising the venture assets or by selling the project assets. But, if there is a need to liquidate assets I^N will take the initiative.

market is not willing to undertake the venture, or (iib) The open capital market is willing to undertake the venture, but E exerts higher effort than any investor I^M. "The open capital market outperforms E" is defined similarly.

I shall also use the following definition

Definition 9 *(viability) "The project is viable" if there is a contract k with $L_0 = i_0$ such that k is accepted by I^N and in equilibrium of the relevant subgame (i.e. Subgame 1 if k is a long-term contract, and Subgame 2 if k is short-term) E receives nonnegative payoff.*

I will now give the details of the entrepreneur game. In the game I represent the payoffs of I^N and E by a 2×1 vector. The first entry of this *payoff vector* represents I^N's surplus and the second E's surplus. Of course the entries of the payoff vector add up to the surplus which is generated.

Date 0 E offers a contract to I^N.

Subgame 1 Contract offered: $k = (i_0, 0, D_2)$

Date 0

I^N accepts or rejects k. If I^N rejects k the game ends and the payoff vector becomes $(0,0)$.

Date 2

1. E offers $P_2 \leq a_2 + c_2$. If $P_2 = D_2$ the game ends with a payoff vector of $(D_2 - i_0, a_2 + c_2 - D_2)$.

2. (If $P_2 < D_2$: E defaults) I^N either accepts P_2, or rejects P_2 and liquidates the project. If I^N accepts, the payoffs are $(P_2 - i_0, a_2 + c_2 - P_2)$. If I^N rejects, she gets $\min\{D_2, a_2\}$ so that the payoffs are $(\min\{D_2, a_2\} - i_0, \max\{0, a_2 - D_2\} + c_2)$.

Subgame 2 Contract offered: $k = (i_0, D_1, 0)$

Date 0

I^N accepts or rejects k. If I^N rejects k the game ends and the payoff vector becomes $(0,0)$.

Date 1

1. E proposes to I^N the alternative schedule $k' = (i_0, 0, D_2)$.

2. I^N accepts k', or rejects k' and liquidates the project. In case of liquidation the game ends with payoffs $(\min\{D_2, a_1\} - i_0, \max\{0, a_1 - D_2\})$. If I^N accepts k', time proceeds to date 2.

Date 2

See Subgame 1, date 2

Subgame 3 Contract offered: $k = (i_0 + I_0, 0, D_2)$

Date 0

1. I^N accepts or rejects k. If I^N rejects k, the game ends with payoffs $(0,0)$.

2. E chooses e.

Date 1

Nature picks the date-2 outcome of the venture. The players observe the (perfect) signal.

Date 2, venture fails

1. E makes a repayment offer $P_2 \leq a_2 + c_2$. If $P_2 = D_2$ the game ends with payoffs $(D_2 - i_0 - I_0, a_2 + c_2 - D_2 - e)$.

2. (If $P_2 < D_2$: E defaults) I^N accepts P_2, or rejects P_2 and liquidates the project. If I^N accepts P_2 the payoffs are $(P_2 - i_0 - I_0, a_2 + c_2 - P_2 - e)$. If I^N rejects P_2 the payoffs are $(\min\{D_2, a_2\} - i_0 - I_0, \max\{0, a_2 - D_2\} + c_2 - e)$.

Date 2, venture succeeds

1. E makes a repayment offer $P_2 \leq a_2 + c_2 + A_2$. If $P_2 = D_2$ the game ends with payoffs $(D_2 - i_0 - I_0, a_2 + c_2 + A_2 - D_2 - e)$.

2. (If $P_2 < D_2$: E defaults) I^N accepts P_2, or rejects P_2 and liquidates E's assets. If I^N accepts P_2 the payoffs are $(P_2 - i_0 - I_0, a_2 + c_2 + A_2 - P_2 - e)$. If I^N rejects P_2 the payoffs are $(\min\{D_2, a_2 + A_2\} - i_0 - I_0, \max\{0, a_2 + A_2 - D_2\} + c_2 - e)$.

Subgame 4 Contract offered: $k = (i_0 + I_0, D_1, 0)$

Date 0

1. I^N accepts or rejects k. If I^N rejects k, the game ends and the payoffs are $(0, 0)$.

2. E chooses e.

Date 1

Nature picks the date-2 outcome of the venture. The players observe the (perfect) signal.

Subgame 4S (venture succeeds)

Date 1 (continued)

1. E offers to I^N the alternative arrangement $k^S = (i_0 + I_0, 0, D_2^S)$

2. I^N accepts k^S, or rejects k^S and liquidates E's assets. In case of liquidation the payoffs are $(\min\{D_1, a_1 + A_2\} - i_0 - I_0, \max\{0, a_1 + A_2 - D_1\} - e)$.

Date 2

1. E makes a repayment offer $P_2 \leq a_2 + c_2 + A_2$. If $P_2 = D_2^S$ the game ends with payoffs $(D_2^S - i_0 - I_0, a_2 + c_2 + A_2 - D_2^S - e)$.

2. (If $P_2 < D_2^S$: E defaults) I^N accepts P_2, or rejects P_2 and liquidates E. If I^N accepts P_2 the payoffs are $(P_1 - i_0 - I_0, a_2 + c_2 + A_2 - P_2 - e)$. If I^N rejects P_2 the payoffs are

$$(\min\{D_2^S, a_2 + A_2\} - i_0 - I_0, \max\{\min\{0, a_2 + A_2 - D_2^S\} + c_2 - e)$$

Subgame 4F (venture fails)

Date 1 (continued)

1. E offers to I^N the alternative arrangement $k^F = (i_0 + I_0, 0, D_2^F)$

2. I^N accepts k^F, or rejects k^F and liquidates E. If I^N rejects k^F the payoffs are $(\min\{D_1, a_1\} - i_0 - I_0, \max\{0, a_1 - D_1\} - e)$.

Date 2

1. E makes a repayment offer $P_2 \leq a_2 + c_2$. If $P_2 = D_2^F$ the game ends with payoffs $(D_2^F - i_0 - I_0, a_2 + c_2 - D_2^F - e)$.

2. (If $P_2 < D_2^F$: E defaults) I^N accepts P_2, or rejects P_2 and liquidates E. If I^N accepts P_2, the payoffs are $(P_2 - i_0 - I_0, a_2 + c_2 - P_2 - e)$. If I^N rejects P_2, the payoffs are $(\min\{D_2^F, a_2\} - i_0 - I_0, \max\{0, a_2 - D_2^F\} + c_2 - e)$.

When the Entrepreneur becomes a Banker

In this subsection I solve the entrepreneur game above. I first present a lemma that establishes what would happen if E decided to invest in the project only. After that I assess whether E is willing to take on both the project and the venture and, if so, whether she outperforms the open capital market. As it turns out E may be willing to issue *long-term* debt to undertake both the project and the venture, but is always outperformed by the open capital market. After that I show that in some circumstances E outperforms the open capital market by issuing *short-term* debt to undertake the project and the venture. In this last case case E effectively founds a firm with two assets and a short-term financial structure: a bank. We shall see that the bank necessarily has to run a liquidity risk in order to function well as monitor of the venture. In particular, if the signal of venture is bad the creditors liquidate the bank. Interestingly, we shall also see that normally an increase in the liquidity of the project has a negative impact on E's effort in monitoring the venture.

Lemma 10 *Project finance is viable if and only if $i_0 \leq a_1$. There are two possible equilibria. If $i_0 \leq a_2$ the project will be financed with either long-term debt, or with short-term debt which is rolled over at date 1. The equilibrium payoffs are $(0, a_2 + c_2 - i_0)$. If $a_2 < i_0 \leq a_1$ the project is financed with short-term debt and liquidated at date 1. In this* *case the equilibrium payoffs are* $(0, a_1 - i_0)$.

Proof. See appendix.

The intuition behind Lemma 10 is simple. In the game E has all the bargaining power, both when writing the contract k and at the repayment dates. As a consequence, in all possible scenarios E gets the entire surplus of the project and I^N gets nothing. The project is viable if E can credibly promise to repay (at least) the amount I^N invests at dtae 0, i.e. i_0. Promises D_i, $i = 1, 2$ become credible when they are backed up by the liquidation proceeds a_i, $i = 1, 2$. If $i_0 \leq a_2$ E can credibly promise $D_2 = i_0$. If $a_2 < i_0 \leq a_1$ E cannot credibly promise to repay $D_2 = i_0$ at date 2. However, a repayment promise $D_1 = i_0$ is credible. Thus, in the latter case the liquidation proceeds at date 1 are enough to make the project viable. Finally, if $i_0 > a_1$ I^N does not accept any contract k to finance the project, because E cannot credibly promise to repay enough.

Lemma 11 *Let the threshold fraction* $\underline{f}$ *be defined as in Proposition 5. Recall that from Assumption 2.10 it followed that* $\underline{f} > 0$. *If* $a_2 - i_0 < \underline{f} I_0$ *E is not willing to issue long-term debt to undertake the project and the venture. If* $a_2 - i_0 \geq \underline{f} I_0$ *E may be willing to issue long-term debt to undertake both the project and the venture, but in this case the open capital market outperforms E.*

Proof. See appendix.

The intuition behind Lemma 11 is the following. Suppose the project and the venture are financed with long-term debt. It can be shown that, in equilibrium, E reaps the cash revenues of the project, i.e. c_2, independent of the outcome of the venture. E therefore only risks losing the remaining part of the project surplus, i.e. $a_2 - i_0$, by combining the project with the venture. In a sense, therefore, $a_2 - i_0$ is the wealth E brings in. We can now use Proposition 5 (iv) to show that if $a_2 - i_0 < \underline{f} I_0$ E is not willing to issue long-term debt to undertake the project and the venture, and that if $a_2 - i_0 \geq \underline{f} I_0$ E prefers to issue long-term debt to undertake the project and the venture to undertaking only the project.

However, we also know from equations 2.12 and 2.13 that $a_2 - i_0 < a_1 - i_0 < \overline{f} I_0$. In words, there is at least one investor I^M who is wealthier than E. Proposition 5 (iii) shows that I^M outperforms E.

The next lemma establishes in which case E is willing to issue short-term debt to undertake both the project and the venture. The lemma makes use of the following programming problem

$$\max_{e, D_1} Q(e)(a_2 + c_2 + A_2 - D_1) - e \qquad \text{(Program P')}$$

$$\text{s.t } D_1 \leq a_2 + A_2$$

$$e \in \arg\max_{\tilde{e}} \{Q(\tilde{e})(a_2 + c_2 + A_2 - D_1) - \tilde{e}\}$$

$$Q(e)D_1 + (1 - Q(e))a_1 \geq i_0 + I_0$$

$$Q(e)(a_2 + c_2 + A_2 - D_1) - e \geq V^E, \text{ where}$$

$$V^E = \begin{cases} 0 & \text{if } i_0 > a_1 \\ a_1 - i_0 & \text{if } a_2 < i_0 \leq a_1 \\ a_2 + c_2 - i_0 & \text{if } a_2 - \underline{f} I_0 < i_0 \leq a_2 \\ Q(\hat{e})(a_2 + A_2 - \widehat{D}_2) + c_2 - \hat{e} & \text{if } i_0 \leq a_2 - \underline{f} I_0 \end{cases}$$

Here V^E is E's opportunity cost of issuing short-term debt to undertake both the project and the venture. The fraction $\underline{f}$ in the definition of V^E is defined as in Proposition 5. The pair $(\hat{e}, \widehat{D}_2)$ in the definition of V^E is defined in the appendix with the proof of the lemma below. In particular, $(\hat{e}, \widehat{D}_2)$ represent the optimum effort level and debt amount if E issued long-term debt to undertake both the project and the venture. An intuitive explanation of Program P' follows the lemma.

Lemma 12 *If Program P' has a feasible solution E is willing to issue short-term debt to undertake the project and the venture. Suppose E indeed issues a short-term debt contract $k = (i_0 + I_0, D_1, 0)$. If, at date 1, the signal of the venture is bad, I^N liquidates E's assets. If the signal is good, E successfully prevents liquidation by proposing the alternative arrangement $k^S = (i_0 + I_0, 0, D_1)$.*

Proof. See appendix.

Lemma 12 suggests that E solves Program P' when deciding whether to issue short-term debt to undertake both the project and the venture. The objective function of Program P' is E's expected payoff. In case the signal is good (probability $Q(e)$) she retains all the project and venture returns minus the repayment requirement. If, instead, the signal is bad E's assets are liquidated. In this case I^N receives all the proceeds from liquidating E's assets, i.e. a_1, while E is left with nothing. The first restriction makes the repayment of D_1 of the alternative arrangement k^S credible. The second restriction gives E's effort choice. The third restriction is I^N's participation constraint. The last restriction is E's participation constraint: E is only willing to offer a contract k if it yields more than reaping the project surplus only, or issuing long-term debt to undertake the project and the venture.

The next proposition gives the solution to Program P' and derives some comparative statics results. It is first shown that an increase in the project returns, has a positive effect on effort. Likewise, an increase in a_1, the debt capacity of the project, has a positive effect on effort. However, interestingly, the proposition shows that this is merely a result of its implied change in E's 'wealth' $a_1 - i_0$. In particular, if i_0 changes at the same rate as a_1 E would exert *less* effort. Above we have defined $\beta = \frac{a_1}{a_2+c_2}$ as a measure for the liquidity of the project. The comparative statics sum up to the following interesting result. Once E's project becomes *more* liquid, while E's 'wealth' $a_1 - i_0$ is kept constant, E will exert *less* effort in monitoring the venture. In other words, all other things equal, E's effort will be higher with an illiquid project.

Proposition 13 *(Solution to Program P')*

(i) If there is a solution to Program P' it is unique and it satisfies

$$Q'(e)(a_2 + c_2 + A_2 - D_1) - 1 = 0$$

$$Q(e)(D_1 - a_1) + a_1 - i_0 - I_0 = 0$$

Assume Program P' has a solution (e^B, D_2^B).

(ii) E is willing to undertake both the project and the venture and would exert e^B effort in monitoring the venture.

(iii) e and D_2 are implicitly defined as functions of the project returns $a_2 + c_2$ in a small enough open neighbourhood of (e^B, D_2^B). In particular

$$\frac{de^B}{d(a_2 + c_2)} > 0 \text{ and } \frac{dD_2^B}{d(a_2 + c_2)} < 0$$

(iv) e and D_2 are implicitly defined as functions of the non-verifiable project returns a_1 in a small enough open neighbourhood of (e^B, D_2^B). In particular

$$\frac{de^B}{da_1} > 0 \text{ and } \frac{dD_2^B}{da_1} < 0, \quad \text{but}$$

$$\frac{de^B}{da_1} < 0\Big|_{(a_1 - i_0 = constant)} \quad \text{and} \quad \frac{dD_2^B}{da_1}\Big|_{(a_1 - i_0 = constant)} > 0$$

Proof. See appendix.

Proposition 14 *(Bank Finance versus Open Capital Market Finance)*
Let $\overline{f} I_0$ be the endowment of the wealthiest investor. E outperforms the open capital market iff Program P' has a solution (e^B, D_1^B) and either (a) Program P has no solution if $f = \overline{f}$ or (b) Program P with $f = \overline{f}$ has a solution (e^*, D_2^*) but $e^B > e^*$.

Proof. The sufficient part of the proof follows directly by definition of outperformance and Corollary 6. It remains to be shown that E outperforms the open capital market *only if* Program P' has a solution and either (a) or (b). Lemma 11 shows that E cannot outperform the open capital market by issuing long-term debt. The last possibility is that E outperforms the open capital market by issuing a short-term contract that *always* leads to liquidation at date 1. This follow easily from the assumption that in this case the 'wealth' that E brings in, i.e. $a_1 - i_0$, is smaller than $\overline{f} I_0$, the endowment of the wealthiest investor.[33]

[33]The relevant programming problem can be found in the footnotes of the proof of Lemma 12.

The next corollary translates the proposition above in the terminology of the banking literature. I use the term *bank* for the firm that E founds by issuing short-term debt to undertake both the project and the venture. I call the bank *fragile* if its creditors I^N (inefficiently) liquidate the bank at date 1 upon observing a bad signal with respect to the venture.

Corollary 15 *Only a fragile bank can potentially outperform the open capital market.*

Example

Below I present a simple example of the model. In this example the open market is not willing to undertake the venture, while E is willing to issue short-term debt to undertake both the project and the venture.

Let the venture require an investment amount of $I_0 = 5/4$. If the venture fails, no return is realised and in case of a success it yields $A_2 = 5/2$. Assume that the wealthiest investor has a fund endowment of 1/4 so that $\overline{f} = 1/5$. With respect to monitoring, let

$$Q(e) = \begin{cases} 0 & \text{if } 0 \leq e < \frac{1}{3} \\ \frac{3e-1}{2e} & \text{if } \frac{1}{3} \leq e < 1 \\ 1 & \text{if } e \geq 1 \end{cases}$$

I first show that for the parameters chosen above Program P does not have a solution, so that the open capital market is not willing to undertake the venture. From the second and the third restriction of Program P we find that $e = 1/2$ and $D_2 = 2$. For these values the first restriction is not satisfied

$$Q(e)(A_2 - D_2) - e + L_0 - I_0 = \frac{1}{2}\left(\frac{5}{2} - 2\right) - \frac{1}{2} + 1 - \frac{5}{4} < 0$$

Program P has no solution so that open market finance is not viable.

Assume with respect to the project that $i_0 = 1/8$, $a_2 = 1/8$ and $c_2 = 7/8$. Assume that, when the project assets are liquidated at date 1, a fourth of their value is recovered:

$a_2 = 1/4$. Notice that E's "wealth endowment" is $a_1 - i_0 = 1/8$, i.e. an amount smaller than $1/4$, the endowment of the wealthiest investor.

E considers two options, namely undertaking the project only or founding a banking firm, i.e. issuing short-term debt to undertake both the project and the venture. When only undertaking the project E successfully manages to issue long-term debt $(1/8, 0, 1/8)$ to reap the entire project surplus of $7/8$. Now let us consider bank finance, i.e. let us see whether Program P' has a solution (e^B, D_1^B). From Program P' we can shown that $e^B = 1$ and $D_1^B = 11/8$. It is readily verified that the bank yields E larger profits than undertaking only the project.

In this example the entrepreneur borrows an amount $i_0 + I_0 = 11/8$ and exerts just enough effort to make the debt claim risk-free. In particular, in the first stage of the entrepreneur game E offers the contract $k = (i_0 + I_0, D_1, 0) = (11/8, 11/8, 0)$ so that we enter Subgame 4. I^N accepts this contract, after which the entrepreneur exerts effort $e = 1$. As $Q(1) = 1$ the good state occurs for sure. At date 1 the entrepreneur successfully proposes the alternative arrangement $k^S = (11/8, 0, 11/8)$ to I^N. At date 2 E repays I^N and keeps the remainder of the project and venture returns, i.e. $5/2 + 1 - 11/8$.

Although, in this example, the debt claim becomes risk-free, this does not mean that long-term debt would work as well. The mechanism in the example works as follows. In case the short-term contract k is used I^N liquidates E if she observes a bad signal with respect to the venture. In this case E loses the entire surplus of the project. To prevent liquidation from happening E exerts high monitoring effort, i.e. $e = 1$. By contrast, with long-term debt E could always retain the cash revenues c_2 of the project, so that there are less incentives to exert effort. By going through the proof of Lemma 11 it can even been shown that E *cannot* issue long-term debt to undertake both the project and the venture. Thus, E really needs to set up as a fragile bank.

2.5 Discussion and Conclusion

In this paper I have argued that firms that engage in opaque, i.e. initially non-observable and non-enforceable, activities can have a comparative advantage in holding illiquid assets. Interestingly, I showed that the comparative advantage only materialises if the firm is financed with short-term hard claims. When holding short-term hard claims investors can trigger liquidation by liquidating the firm. Punishment by liquidation is severe in case the firm holds illiquid assets as future rents are destroyed. When the opaque activity takes the form of screening and monitoring investment projects the firm can be well compared with a commercial bank. Like the firm a commercial bank holds illiquid assets (illiquid loans and reputational capital) and is mainly financed with short-term hard claims (deposits and other forms of short-term debt).

To state the results in other words, the model provides a 'complete' theory of banking, i.e. it explains why banks combine the business of monitoring projects with holding illiquid assets and a short-term finance structure (a universal bank?). At present the banking literature has no such theory available. Notice that the theory relates to a very fundamental issue in intermediation theory, namely it presents a rationale why opaque activities are performed by firms rather than agents.

The theory could potentially have important policy consequences. For example, one prediction is that banks exist precisely *because* of their fragile financial structure and that taking away this risk would undermine the market for monitoring services. Hence, according to the theory 'narrow banking' proposals, which prescribe that banks match the maturity structures of their assets and liabilities, could be harmful.

A corollary of the theory is that, if investors have relatively small fund endowments, the bank will be a better monitor than the open capital market. The reason for this result is that the market cannot precommit to exert high monitoring effort. As a consequence, the rate of return on a venture becomes too low. The corollary offers a theoretical explanation why in practice firms without an established credit history ('small firms') depend on bank finance and have no access to the open capital market.[34] The theory could possible also help answering another very interesting open question: Why do banks generally hold their originated illiquid loans on the balance sheet. The alternative would be to securitise the complete portfolio while retaining responsibility for eventual losses resulting from default. The theory argues that the loan portfolio has an as of yet undetected function, namely to stimulate bankers to exert effort. Yet, the theory also shows that a banks reputational capital may serve the same function. Therefore, banks with a valuable reputation or brandname can have plenty of scope to securitise the loan portfolio.

Venture capital funds form a counter example of the model. Indeed, venture capital funds exert substantial effort in monitoring 'small firms', but they are typically financed with *long-term* capital. A possible reason for this may be that the surplus generated in case the ventures become successful is big enough to adopt high-powered incentive schemes, such as option schemes, that align the interest of the funds' management sufficiently with that of the outside investors. Also it may be the case that venture capital funds often have one, or a few, big shareholders who actively monitor the funds' management.

An important caveat of the theory is that it applies only to entrepreneurial firms. It is assumed that the entrepreneur, who controls the bank, is also the residual claimant. However, in reality we see that banks are typically publicly-held firms. Therefore, those in control —the bank managers— are not the residual claimants. In this case the very same free-rider problem as is addressed in the paper should arise. When ownership and control are separated, the model only makes sense if the incentives of those in control are for some reason sufficiently in line with the owners' objectives. However, Jensen (1986) argues that this is often not the case.[35] Let us argue in the spirit of the model what

[34] See the credit channel literature for evidence, e.g. Gertler and Gilchrist (1993).

[35] But even if managers had the same interests as the shareholders there would be a problem with the theory. Namely, in that case it would not be clear why monitoring firms should ever issue risky debt; all equity finance would be optimal.

could be possible resolutions to this criticism. The model shows that the incentives of the bank manager and its investors (i.e. all the stakeholders which hold risky claims) are better alined if the bank manager reaps the rents c_2. However, if a part c_2 of the project rents flows to the bank manager it is problematic to interpret the project as the bank's loan portfolio, or its reputational capital. Why would the bank manager, instead of the investors, benefit from these assets? In this case it seems best to interpret the rents c_2 as *control rents*, and the project as the entire bank. Indeed, Jensen's argument was based on the assumption that managers obtain a stream of private benefits when retaining control. Possibly also the incentives of the owners and managers are sufficiently aligned in reality so that the model applies to the differences of interest between shareholders and fixed claim holders. However, in this case the prediction of the theory would be that banks should be completely financed with equity claims!

2.6 References

Aghion, P. and P. Bolton (1989), 'The Financial Structure of the Firm and the Problem of Control,' *European Economic Review,* **33**, pp. 286-293.

Aghion, P. and P. Bolton (1992), 'An Incomplete Contracting Approach to Financial Contracting,' *Review of Economic Studies,* **59**, pp. 473-494.

Barclay, M.J. and C.W. Smith, Jr. (1995), 'The Maturity Structure of Corporate Debt,' *Journal of Finance,* **50**, pp. 609-631.

Berglöf, E. and E-L.Von Thadden (1994) 'Short-term versus Long-term Interests: Capital Structure with Multiple Investors,' *Quarterly Journal of Economics,* **109**, pp. 1055-1084.

Bhattacharya, S. and A.V. Thakor (1993) 'Contemporary Banking Theory,' *Journal of Financial Intermediation,* **3**, pp. 2-50.

Boot, A.W.A., S.I. Greenbaum and A.V. Thakor (1993) 'Reputation and Discretion in Financial Contracting,' *American Economic Review,* **83**, pp. 1165-1183.

Brick, I.E. and S.A. Ravid. (1985) 'On the Relevance of the Debt Maturity Structure,' *Journal of Finance,* 40, pp. 1423-1437.

Calomiris, C.W. and C.M. Kahn (1991) 'The Role of Demandable debt in Structuring Optimal Banking Arrangements,' *American Economic Review,* **81**, pp. 497-513.

Diamond, D.W. (1984) 'Financial Intermediation and Delegated Monitoring,' *Review of Economic Studies,* **51**, pp. 393-414.

Diamond, D.W. (1991) 'Debt Maturity Structure and Liquidity Risk,' *Quarterly Journal of Economics,* **106**, pp. 709-737.

Diamond, D.W. (1997) 'Liquidity, Banks and Markets,' *Journal of Political Economy,* **105**, pp. 928-956.

Diamond, D.W. and R.G. Rajan (1999) 'Liquidity Risk, Liquidity Creation and Financial Fragility: A Theory of Banking,' NBER Working Paper 7430, December 1999.

Flannery, M.J. (1994) 'Debt Maturity and the Deadweight Cost of Leverage: Optimally Financing Banking Firms,' *American Economic Review*, **84**, pp. 320-331.

Gale, D. and M. Hellwig (1985) 'Incentive-Compatible Debt Contracts I: The One Period Problem,' *Review of Economic Studies*, **52**, 647-664.

Gertler, M. and S. Gilchrist (1993) 'The Role of Credit Market Imperfections in the Monetary Transmission Mechanism: Arguments and Evidence,' *Scandinavian Journal of Economics*, **95**, pp. 43-64.

Hart, O. and J. Moore (1988) 'Incomplete Contracts and Renegotiation,' *Econometrica*, **56**, pp. 755-785.

Hart, O. and J. Moore (1994) 'A Theory of Debt Based on the Inalienability of Human Capital,' *Quarterly Journal of Economics*, **109**, pp. 841-880.

Hart, O. and J. Moore (1995) 'Debt and Seniority: An Analysis of the Role of Hard Claims in Constraining Management,' *American Economic Review*, **85**, pp. 567-585.

Jensen, M. (1986) 'Agency Costs of Free Cash Flow, Corporate Finance and Takeovers' *American Economic Review, Papers and Proceedings*, **76-2**, pp. 323-329.

Jensen, M. and W. Meckling (1976) 'Theory of the Firm: Managerial Behavior, Agency Costs and Capital Structure' *Journal of Financial Economics*, **3**, pp. 305-360.

Myers, S.C. (1977) 'Determinants of Corporate Borrowing,' *Journal of Financial Economics*, **5**, pp. 147-175.

Myers, S.C. and R.G. Rajan (1998) 'The Paradox of Liquidity,' *Quarterly Journal of Economics*, **119**, pp. 733-771.

Qi, J. (1997) 'Deposit Liquidity and Bank Monitoring,' *Journal of Financial Intermediation*, **7**, pp. 198-218.

Schure, P. 'The Optimal Loan Contract with Costly State Verification,' Unpublished, European University Institute, 1998.

Townsend, R.M. (1979) 'Optimal Contracts and Competitive Markets with Costly State Verification,' *Journal of Economic Theory*, **21**, pp. 265-293.

Winton, A. (1995) 'Costly State Verification and Multiple Investors: The Role of Seniority,' *Review of Financial Studies*, **8**, pp. 91-123.

2.7 Appendix

Proof of Proposition 5

Let me start with a number of remarks. First, the restriction $D_2 \leq A_2$ is redundant because it is implied by the first restriction. Second, if there is a solution to Program P it satisfies

$$Q(e)D_2 - L_0 = 0 \tag{2.15}$$

[If this were not the case, L_0 could be slightly increased to increase the maximand, while remaining in the feasible region.] Third, from assumptions 2.2—2.3 it easily follows that the second restriction of Program P holds if and only if

$$Q'(e)(A_2 - D_2) - 1 = 0 \tag{2.16}$$

After substituting 2.15 the maximand of Program P simplifies to

$$Q(e)A_2 - e - L_0$$

From 2.16 we see that the maximand increases in e for all $D_2 > 0$. Equation 2.15 shows that indeed we must have $D_2 > 0$, so that solutions to Program P must be pairs (e, D_2) that satisfy 2.15 and 2.16 and have the maximum possible e.

Statement (i). Let (e^0, D_2^0) and (e^1, D_2^1) be two optima. We have just seen that $e^0 = e^1$. But then it follows from e.g. 2.16 that $D_2^0 = D_2^1$. The solutions are therefore identical.

Statement (ii). From 2.5 we know that e^{FB} exists and from 2.2—2.3 that it is unique. From 2.15 and $f < 1$ it follows that $D_2 > 0$. From $D_2 > 0$ and 2.16 it follows that $e < e^{FB}$.

Statement (iii). Substitute equation 2.15 in equation 2.16 and note that $L_0 = (1-f)I_0$. We can apply the implicit function theorem to show that $e = e(f)$. From $e(f)$ one can derive that $\frac{de}{df}\Big|_{(e,D_2)=(e^*,D_2^*)} > 0$ and then via 2.16 that $\frac{dD_2}{df}\Big|_{(e,D_2)=(e^*,D_2^*)} < 0$. [The signs of these partial derivatives can be determined for instance by looking at the Kuhn-Tucker optimality conditions of Program P.] We can also see more directly that the sign of the partials must be as given in the lemma. Namely, any feasible change $\Delta f > 0$ lead to changes Δe^* and ΔD_2^*. From 2.16 we see that we must have either $\Delta e^* > 0$ and $\Delta D_2^* < 0$ or $\Delta e^* < 0$ and $\Delta D_2^* > 0$. It follows directly from Program P that $\Delta e^* < 0$ and $\Delta D_2^* > 0$ is impossible because $\Delta f > 0$ increases the feasible set of Program P so that the objective cannot decrease. We have seen above that the objective function is increasing in e.

Statement (iv) follows directly from (iii) and assumption 2.5.

Proof of Lemma 10

The proof follows directly from the equilibria to Subgame 1 and Subgame 2. The equilibrium of both the date 2 subgames is $P_2^* = \min\{D_2, a_2\}$ so the equilibrium payoffs are $(\min\{D_2, a_2\} - i_0, \max\{0, a_2 - D_2\} + c_2)$.

In Subgame 1 at date 0 I^N accepts a contract $(i_0, 0, D_2)$ iff $\min\{D_2, a_2\} \geq i_0$. Clearly, if $a_2 < i_0$ I^N rejects any offered contract. If, instead, $a_2 \geq i_0$ E chooses precisely $D_2 = i_0$ in the preceeding stage. Summarising, if $a_2 < i_0$ the equilibrium payoffs are $(0,0)$ and if $a_2 \geq i_0$ the equilibrium payoff are $(0, a_2 + c_2 - i_0)$.

Now turn to subgame 2 at date 1. In stage 2 I^N accepts the alternative schedule $k' = (i_0, 0, D_2)$ iff $P_2^* = \min\{D_2, a_2\} \geq D_1$. Clearly, if $a_2 < D_1$ I^N rejects any offered k'. If, instead, $a_2 \geq D_1$ E offers k' with precisely $D_2 = D_1$ at stage 1 of date 1. In fact, this gives payoffs $(D_1 - i_0, a_2 + c_2 - D_1)$, while offers k' with $D_2 < D_1$ are rejected in stage 2 of date 1 to give $(D_1 - i_0, a_1 - D_1)$. Summarising the situation at date 1, if $a_2 \geq D_1$ the payoffs become $(D_1 - i_0, a_2 + c_2 - D_1)$ and if $a_2 < D_1$ $(\min\{D_1, a_1\} - i_0, \max\{0, a_1 - D_1\})$. At date 0 I^N accepts contracts $k = (i_0, D_1, 0)$ iff $\min\{D_1, a_1\} - i_0 \geq 0$. Consequently, if $a_1 < i_0$ I^N rejects all conceivable k. If, instead, $a_1 \geq i_0$ E offers precisely $k = (i_0, i_0, 0)$ in the preceeding stage, which is subsequently accepted.

Summarising, the project is viable iff $a_1 \geq i_0$. The equilibrium payoffs are $(0, a_2 + c_2 - i_0)$ if $i_0 \leq a_2$, $(0, a_1 - i_0)$ if $a_2 < i_0 \leq a_1$, and $(0,0)$ if $a_1 < i_0$.

Proof of Lemma 11

I shall solve Subgame 3 and apply Proposition 5 to show (i) if $a_2 - i_0 < \underline{f} I_0$ long-term finance to undertake both the project and the venture is either not feasible or dominated by investing in the project only, (ii) if $a_2 - i_0 \geq \underline{f} I_0$ long-term finance to undertake both the project and the venture dominates undertaking the project only, and (iii) that in any case the open capital market outperforms E.

Consider date 2. We can assume here that $D_2 \geq i_0 + I_0$. Namely, if not I^N would have rejected $(i_0 + I_0, 0, D_2)$ in the first stage of Subgame 3. Assume first that the venture has failed. The equilibrium then becomes ($P_2^* = \min\{D_2, a_2\}$, accept). From equations 2.12, 2.13 and $D_2 \geq i_0 + I_0$ it follows that $P_2^* = \min\{D_2, a_2\} = a_2$. Thus, if the venture fails the equilibrium payoff are $(a_2 - i_0 - I_0, c_2 - e)$. If, by contrast, the venture is a success the date 2 equilibrium becomes ($P_2^* = \min\{D_2, a_2 + A_2\}$, accept). Notice that for any contract k we have $P_2^* \leq a_2 + A_2$. Therefore, we can assume without loss of generality that k is such that

$$D_2 \leq a_2 + A_2 \tag{2.17}$$

Using this, we see that in case the venture succeeds we get $P_2^* = D_2$, which is accepted by I^N, so that the date 2 equilibrium payoffs are $(D_2 - i_0 - I_0, a_2 + c_2 + A_2 - D_2 - e)$.

Now consider date 0. In stage 2 E chooses effort so as to maximise expected payoff

$$e \in \arg\max_{\tilde{e}} \{Q(\tilde{e})(a_2 + A_2 - D_2) + c_2 - \tilde{e}\} \tag{2.18}$$

In stage 1 I^N accepts the contract if and only if

$$Q(e)D_2 + (1 - Q(e))a_2 \geq i_0 + I_0 \tag{2.19}$$

E would only consider contracts $(i_0 + I_0, 0, D_2)$ that yield more then undertaking only the project, i.e.

$$Q(e)(a_2 + A_2 - D_2) + c_2 - e \geq V^{\text{Project}} \quad \text{where} \tag{2.20}$$

$$V^{\text{Project}} = \begin{cases} 0 & \text{if } i_0 > a_1 \\ a_1 - i_0 & \text{if } a_2 < i_0 \leq a_1 \\ a_2 + c_2 - i_0 & \text{if } i_0 \leq a_2 \end{cases}$$

Define $D_2' \equiv D_2 - a_2$ and substitute it in equations 2.17—2.20 to get

$$D_2' \leq A_2 \tag{2.21}$$

$$e \in \arg\max_{\tilde{e}} \{Q(\tilde{e})(A_2 - D_2') - \tilde{e}\}$$

$$Q(e)D_2' \geq i_0 - a_2 + I_0$$

$$Q(e)(A_2 - D_2') - e \geq V^{\text{Project}} - c_2$$

The above can be summarised as follows. If there is *no* pair (e, D_2') satisfying Restrictions R then E is not willing to issue long-term debt to undertake the project and the venture. If there is a pair (e, D_2') satisfying Restrictions R then E prefers issuing long-term debt to undertake both the project and the venture to undertaking only the project. Possibly, though, issuing *short-term* debt is even better for E (This possibility is studied later in the main text). Thus, if there is a pair (e, D_2') satisfying Restrictions R then E *may be* willing to issue long-term debt to undertake both the project and the venture. We can now easily prove the statements above.

Statement (i) and (ii). First, assume that $i_0 > a_2$. It follows directly from Assumption 2.10 that there is *no* pair (e, D_2') satisfying Restrictions R. Now assume $i_0 \leq a_2$. Notice that $V^{\text{Project}} - c_2 = a_2 - i_0$. Therefore, in this case Restrictions R reduce to the restrictions of Program P in the main text, with $a_2 - i_0 = fI_0$. We can apply Proposition 5 (iv) to show that, first, if $a_2 - i_0 < \underline{f}I_0$ there is *no* pair (e, D_2') satisfying Restrictions R, and, second, if $a_2 - i_0 \geq \underline{f}I_0$ there is a pair (e, D_2') satisfying Restrictions R.

Statement (iii) above follows from $a_2 - i_0 \underset{\text{see 2.12}}{<} a_1 - i_0 \underset{\text{see 2.13}}{\leq} \overline{f}I_0$ and Proposition 5 (iii).

Proof of Lemma 12

The statements in the lemma are simply proved by solving Subgame 4 by backward induction. First solve Subgame 4F. In Subgame 4F we can assume that[36]

$$D_1 \geq i_0 + I_0 > a_1 \tag{2.22}$$

The date-2 equilibrium is $\{P_2^* = \min\{D_2^F, a_2\}, \text{accept}\}$ and it leads to a payoff vector of $(P_2^* - i_0 - I_0, a_2 + c_2 - P_2^* - e)$. Notice that $P_2^* \leq a_2$. At date 1, stage 2 I^N rejects any alternative arrangement $k^F = (i_0 + I_0, 0, D_2^F)$. This is because rejection gives I^N a payoff of $\min\{D_1, a_1\} - i_0 - I_0 = a_1 - i_0 - I_0$ while we just saw that acceptance only yields $P_2^* - i_0 - I_0 \leq a_2 - i_0 - I_0$. Summarising, in case the project fails E is liquidated and the equilibrium payoffs of Subgame 4F are

$$(a_1 - i_0 - I_0, -e)$$

I next solve Subgame 4S. The equilibrium at date 2 is $\{P_2^* = \min\{D_2^S, a_2 + A_2\}, \text{accept}\}$. Notice that for any k^S we have $P_2^* \leq a_2 + A_2$. Therefore, we can assume without loss of generality that k^S is such that

$$D_2^S \leq a_2 + A_2 \tag{2.23}$$

Using this equation we obtain that the equilibrium payoffs for the date 2 subgame are $(D_2^S - i_0 - I_0, a_2 + c_2 + A_2 - D_2^S - e)$. Now turn to date 1 at stage 2. I^N accepts alternative arrangements k^S iff $P_2^* = D_2^S \geq \min\{D_1, a_1 + A_2\}$. Clearly, if $D_1 > a_2 + A_2$ I^N rejects all offered k^S. If, instead, $D_1 \leq a_2 + A_2$ E offers precisely $k^S = (i_0 + I_0, 0, D_1)$ in stage 1 of date 1.

Summarising, if contract k satisfies $D_1 \leq a_2 + A_2$ the equilibrium payoffs to Subgame 4S are

$$(D_1 - i_0 - I_0, a_2 + c_2 + A_2 - D_1 - e)$$

[36] The equation holds because in stage 1 at date 0 of Subgame 4 I^N only considers accepting contracts $(i_0 + I_0, D_1, 0)$ that satisfy $D_1 \geq i_0 + I_0$. The second inequality follows from 2.13.

If, instead, $D_1 > a_2 + A_2$ I^N rejects any of E's feasible offers so that the equilibrium payoffs become

$$(\min\{D_1, a_1 + A_2\} - i_0 - I_0, \max\{0, a_1 + A_2 - D_1\} - e)$$

Date 0, stage 2. E chooses effort to maximise her expected payoff. If $D_1 \leq a_2 + A_2$ we get[37]

$$e \in \arg\max_{\tilde{e}}\{Q(\tilde{e})(a_2 + c_2 + A_2 - D_1) - \tilde{e}\} \tag{2.24}$$

Date 0, stage 1. I^N accepts the contract $(i_0 + I_0, D_1, 0)$ if her expected payoffs exceed the investment amount. If $D_1 \leq a_1 + A_2$ we get[38]

$$Q(e)D_1 + (1 - Q(e))a_1 - i_0 - I_0 \geq 0 \tag{2.25}$$

E is willing to offer a contract $(i_0 + I_0, D_1, 0)$ if it yields more that the payoff from the best alternative, V^E say. So, if $D_1 \leq a_2 + A_2$ we must have[39]

$$Q(e)(a_2 + c_2 + A_2 - D_1) - e \geq V^E \tag{2.26}$$

E's alternative options are to undertake the project only or to undertake the project and the venture with long-term debt. Thus V^E is as defined in the main text. The tuple $(\widehat{e}, \widehat{D}_2)$ of the main text is the solution to

$$\max_{e, D_2} Q(e)(a_2 + A_2 - D_2) + c_2 - e$$

s.t. equations 2.17—2.20

[37]If $a_2 + A_2 < D_1 \leq a_1 + A_2$ we get $e \in \arg\max_e\{Q(e)(a_1 + A_2 - D_1) - e\}$. And, if $D_1 > a_1 + A_2$ we get $e = 0$.

[38]And, if $D_1 > a_1 + A_2$ we get $Q(e)A_2 + a_1 - i_0 - I_0 \geq 0$.

[39]If $a_2 + A_2 < D_1 \leq a_1 + A_2$ we have $Q(e)(a_1 + A_2 - D_1) - e \geq V^E$. And, if $D_1 > a_1 + A_2$ we have $0 \geq V^E$.

Proof of Proposition 13

Statement (i) is proved along the lines of the proof of Proposition 5 (i). Statement (ii) is a restatement of Lemma 12. Statement (iii) and (iv) are all directly derived from analysing

$$Q'(e)(a_2 + c_2 + A_2 - D_1) - 1 = 0$$

$$Q(e)(D_1 - a_1) + a_1 - i_0 - I_0 = 0$$

Let me proof statement (iii) as a example. From the equations we obtain

$$\begin{pmatrix} Q''(e)(a_2 + c_2 + A_2 - D_1) & -Q'(e) \\ Q'(e)(D_1 - a_1) & Q(e) \end{pmatrix} \begin{pmatrix} \frac{de^B}{d(a_2+c_2)} \\ \frac{dD_2^B}{d(a_2+c_2)} \end{pmatrix} = \begin{pmatrix} -Q'(e) \\ 0 \end{pmatrix}$$

From this equation statement (iii) follows (from optimality it follows that the determinant of the matrix above is negative near the optimum).

Chapter 3

Economies of Scale and Efficiency in European Banking: New Evidence

Joint with Rien Wagenvoort[0]

3.1 Introduction

The number of studies that evaluate the performance of European banks sink into insignificance beside the voluminous literature on US financial institutions. This paper partially fills this gap by investigating the cost efficiency of almost 2000 credit institutions across the fifteen European Union countries.

On the first of January 1993, the Second Banking Directive (1988) of the European Union and a number of the other key EU directives[1] related to the financial service industry were implemented. This heralded a new episode of deregulation, standardised capital requirements and changes in supervision rules and deposit-guarantee schemes. The *single*

[0]Rien Wagenvoort is an Economist at the European Investment Bank (EIB), Luxembourg.

[1]Namely, the Money Laundering Directive, the Own Funds Directive, the Solvency Ratio Directive, the Consolidated Supervision Directive, and the directive on Deposit-Guarantee Schemes. The Large Exposures Directive, the Capital Adequacy Directive and the Investment Services Directive came into force in 1994, 1995 and 1996 respectively. See, among others, Molyneux et al. (1996).

passport and *mutual recognition* have cleared the road for cross-border banking, while the introduction of the Euro on the first of January 1999 took away one of the last obstacles for a competitive and integrated banking market. The general belief among bankers and academics is that competition has significantly increased in this changing European banking environment. Indeed, the numerous cases of recent mergers and acquisitions in the financial world would indicate that bankers and insurers are trying to reshape their businesses into more profitable and lean (cost efficient) institutions in order to face national and global competitive pressures. Traditional income streams such as interest margins have dried up, whereas new sources of revenues such as brokerage services, investment banking products, risk management and portfolio management have become more and more important. Besides major changes in the regulatory environment, the banking industry has been affected by the availability of new computer technologies.

Given the broad picture sketched above, one may ask whether the performance of European credit institutions has improved over the five years following the implementation of the Second Banking Directive. In this paper we evaluate the performance of banks in this period by looking at cost efficiency. In particular, we analyse how production costs depend on the scale of operation, managerial efficiency (X-efficiency), technological progress, and the legal status of the institutions. For this purpose, we estimate a cost frontier.

To specify the cost model we choose the Cobb-Douglas function which is augmented with dummies in order to account for differences in average costs due to the time period, the bank's type (legal status) and its size. In response to the critique that the standard Cobb-Douglas and Translog cost functions are too restrictive to accurately measure economies of scale we include seven size dummies in the cost function.[2] This way of modelling gives sufficient flexibility with respect to economies of scale, and can for example include a U-shaped average cost curve.

[2]Therefore, recent stochastic cost frontier studies consider the Fourier Flexible functional form (see Altunbas et al. (1999), Altunbas, Goddard and Molyneux (1999), and Berger and Mester (1997)).

Our model is not suited to measure economies of scope. One of the reasons for this are the restrictions on the substitutability between inputs and outputs that are imposed on the functional form of our cost model. Therefore, we refrain from predicting what will be the economic gains of *universal banking*. In recent efficiency studies, however, only small increasing economies of scope were detected. See Berger, Hunter and Timme (1993), Berger and Humphrey (1997), Berger and Mester (1997), and Berger, Demsetz and Strahan (1998) for comprehensive surveys on empirical findings regarding the existence of scale and scope economies and X-efficiency of financial institutions.

A second issue which is not addressed in this paper is performance as measured by profitability. We focus on cost minimisation. From the duality theorem it follows that the technology of a firm can be described by the parameters of the cost function. However, optimising the level of output given the available resources does not necessarily lead to profit and revenue maximisation in economies that are characterised by, for instance, oligopolistic markets, asymmetric information and risk-averse individuals. In response to this, some recent articles (see, among others, Berger and Mester (1997), and Rogers (1998)) consider, besides the traditional cost function, also the profit and revenue frontiers and derive X-efficiency measures from these. Although these studies give useful insights in the differences in profitability of banks, a serious problem with these approaches is that market power may obscure the efficiency (in terms of productivity) results.[3]

This paper innovates with respect to traditional cost frontier analyses in three distinctive ways:

- First, a new econometric technique is employed to estimate the parameters of the cost function. A detailed exposition of the method, the so-called Recursive Thick Frontier Approach (RTFA), is given in Wagenvoort and Schure (1999). The tradi-

[3]An interesting related topic is whether high market concentration or high market shares is a result of better performance or whether it reflects monopoly power. This question is especially relevant for public policy considerations of anti-trust institutions. In this study we do not test this so-called *structure-conduct-performance* relationship (see, among many others, Berger (1995), Goldberg and Rai (1996) and Maudos (1998)).

tional econometric techniques for frontier models, namely the Stochastic Frontier Approach (SFA), the Thick Frontier Approach (TFA) and the Distribution Free Approach (DFA) (see Aigner, Lovell and Schmidt (1977), Berger and Humphrey (1992) and Berger (1993) respectively) have in common that they depend on *a priori* assumptions that are difficult to test. Our approach is based on the observation that if deviations of X-efficient companies from the frontier are random then one must observe for this group of banks that the probability of being located either above or below the frontier is equal to one half. This hypothesis can be tested for panel data sets but requires careful sorting of the full sample into a group of X-inefficient banks and a group of X-efficient banks.

- Second, we present a method to disentangle the effects of input prices on the average costs from the impact of other time-related effects such as technological progress and deregulation. As a consequence we are able to reveal shifts in the cost frontier over time.

- Third, our data set allows for a more general definition of X-efficiency than obtained in the usual studies of this type. In traditional cost studies, X-inefficiencies arise from a wasteful use of resources. Differences in performance can normally not be caused by inefficient acquisition of the inputs, since every bank is assigned a different input price vector based on the *actual* cost incurred. For example, the price of labour is defined as the bank's expenses on labour divided by its number of employees. By choosing input prices on the basis of the actual expenses of each individual bank, it is implicitly assumed that banks pay the "right" amount for their inputs. In other words, variation in input prices is ascribed to differences in the quality of inputs. By contrast, in our study we adopt the idea that differences in efficiency stem from both the wasting of resources and because managers acquire these resources inefficiently. In particular, input prices are, as far as possible, constructed from general price indices (for buildings, financial services, wages etc.)

instead of the actual expenses of a bank.

Our results on the efficiency of European banks can be summarised as follows: the computed X-inefficiencies are on average between 16 percent and 20 percent and they dominate by far the possible gains from choosing the right scale of operation. Although the savings bank sector can reduce the costs per unit of assets by roughly 6 percent by increasing size, significant scale effects are only found for small institutions (with total assets up to EUR 600 million). For the overall banking industry, economies of scale are negligible with respect to the cost reductions that can be achieved by improving the quality of its managers. For the full sample, technological progress could not be detected. However, the average costs of X-efficient savings banks significantly reduced (by 9 percent) during the sample period, possibly due to technological innovation. Substantial differences in X-efficiencies exist across Europe. In 1997, UK bankers were almost fully efficient, whereas Greek bankers were the most inefficient ones with X-inefficiencies exceeding, on average, the 55 percent level. A striking result however is that the cost dispersions in some European countries, i.e. Finland, Ireland, Italy, The Netherlands and the UK, were rapidly reduced.

These empirical findings are in accordance with earlier studies on US financial institutions (see, for instance, Berger, Hanweck and Humphrey (1987), McAllister and McManus (1993) and the review article of Berger and Humphrey (1997)) but contradict recent results on the scale efficiency of both American and European financial institutions. Hughes and Mester (1998), and Altunbas and Molyneux (1996) find positive economies of scale for a broader range of size classes for American banks and French and Italian banks respectively, including banks with total assets above USD 3 billion.[4]

The remainder of this paper is organised as follows: In the next section we introduce our cost model and embed it into the literature and methodology of bank efficiency studies. Since our adopted econometric method is new we devote an entire section to it

[4]However, Altunbas and Molyneux (1996) do not find significant sale economies for banks above the size of USD 100 million in Germany and Spain.

(Section 3). Section 4 introduces the data we use and presents some descriptive statistics. Section 5 contains the results, while Section 6 concludes. Appendix 1 and Appendix 2 contain detailed information on price data and the tables containing the regression results, respectively. In Appendix 3 we explain how to compute the variance of the parameter estimates of our cost model.

3.2 The Model

Background

Our study focuses on the production technology of banks and in particular on their *X-efficiency* —i.e. whether they use and acquire their available inputs efficiently— and *scale efficiency* —i.e. whether they produce the right amount of outputs. *Scope efficiency* —i.e. whether banks choose an efficient combination of outputs— is left to future research. The approach taken here and in other studies is to estimate a cost frontier, i.e., a function which gives the minimum costs to produce a specific output bundle given the input prices.

The cost approach in bank efficiency studies is natural: a bank normally has multiple outputs so that a statistical model of the production function would have multiple endogenous variables. Such a model would be difficult to estimate. By contrast, in the cost function the only endogenous variable is total costs. By the Duality Theorem[5], we know that, like the production function, a cost function summarises all the relevant information of a firm's technology.

Cost frontiers for banking industries are usually estimated by regressing total costs on a particular function with outputs and input prices as its arguments:

$$TC_{ti} = f_t(y_{ti,1}, ..., y_{ti,K}, p_{t1}, ..., p_{tJ}) + \varepsilon_{ti} \tag{3.1}$$

In equation 3.1 TC_{ti} represents the total cost of bank i in period t. K and J are the

[5]The Duality Theorem can be found in any standard micro textbook such as for example Varian (1992).

number of outputs of bank i and the number of its inputs, respectively. The amount of output of type k is denoted by $y_{ti,k}$, $k = 1, ..., K$; the price of input j by p_{tj}, $j = 1, ..., J$; and the disturbance term by ε_{ti}.

Bank cost efficiency studies may differ in the choice and definitions of the variables of equation 3.1, in the choice of the functional form $f_t(.)$, and in the estimation technique employed. The remainder of this section explains what functional form $f_t(.)$ and which arguments we choose in this study. Our estimation technique is explained in the next section.

Output variables. Bank efficiency studies have adopted different notions of what is meant by "the production of a bank".[6] We view the bank as a *producer of services* such as screening projects, monitoring borrowers, enforcing contracts, portfolio selection, hedging risks, providing brokerage services, keeping deposits and other claims liquid, providing repayment insurance, etc. In defining services as the banks' production we adopt what Berger and Humphrey (1992) call the *Value-Added Approach* in defining a bank's production.[7] We choose inputs to be all goods and services which are needed to generate the value added. For example, labour and office space are inputs as they are needed for the service production of a bank.

Measuring the service production of a bank poses a new problem. How to quantify the services provided? For example, how would one measure the services offered to account holders? Ideally one would like to have data on the number of transactions processed, the number of account statements sent to customers and the like. Unfortunately these data are not available. For other outputs, such as the 'amount' of contract enforcement and the 'amount' of risk hedged, the problems get even worse. In the Value-Added Approach these problems are by-passed by assuming that the amount of services produced are

[6]Berger and Humphrey (1992) distinguish three approaches to defining bank outputs. The *Asset Approach* defines the assets of a bank as outputs and the liabilities as inputs. The *User Cost Approach* treats assets or liabilities that increase the value of the banking firm as outputs, and the remaining assets and liabilities as inputs. The *Value-Added Approach* is explained below.

[7]Viewing banks as offering financial services is traditionally called the "intermediation approach" (see Sealey and Lindley (1977)).

proportional to the money-value of certain various variables of the balance sheet and the profit and loss account. In other words, the output variables in the statistical model are those that are recognised to imply service production. In the example above, the Euro value of deposits would be the obvious choice. Also loans would be considered an output because offering loans implies services such as screening projects, monitoring borrowers, enforcing contracts and diversifying risks. It is less clear that assets such as government bonds are production as they normally do not imply much additional work for the bank and therefore do not incur much additional costs. Some variables have both output and input characteristics. For example, deposits imply service production but they also provide funding of the bank's activities.[8]

We have defined 5 output variables: customer deposits, loans, equity investments, off-balance sheet items, and other services. A detailed description of the outputs and why we chose them is found in Section 4.

Input prices. Like in many bank efficiency studies we distinguish the following inputs: labour, funds (capital) and 'buildings and computers' (fixed assets). Our dataset allows us to measure the input prices differently from most related studies. In particular, the input prices are, as far as possible, constructed from general price indices. For details, see again Section 4.

Other efficiency studies base the input prices of each bank on the *actual* cost incurred. To illustrate this, the wage rate is normally defined as the total wage bill of the bank over the number of employees, the price of funds as the ratio of the interest expenses and the total amount of funds used, and the price of a unit of housing as the ratio of depreciation and the average value of fixed assets. By choosing input prices on the basis of the actual expenses incurred, these studies impose that banks pay the "right" amount for their inputs, so that inefficiencies only arise from an inefficient use of resources. By

[8]It can be pointed out that variables on the balance sheet and profit and loss account often vastly overstate the actual flow of services produced. However, this is simply a matter of scaling which in a regression turns out to be immaterial. For example, in the case of loans the product of the total amount of loans and its concomitant parameter estimate will still correctly represent the contribution of loan services to total costs.

contrast, in our study, by using general price indices, we adopt the idea that differences in efficiency stem from both an inefficient use of resources and because managers acquire these resources inefficiently.

Functional form. In related studies the functional form $f_t(.)$ is usually chosen to be a second-order Taylor approximation in logs of a general cost function, the so-called translog cost function. Thus a fairly general specification of a bank's technology is provided. Some recent cost frontier studies using a translog cost function are Altunbas and Molyneux (1996), Berger (1995), Berger and Hannan (1998), Goldberg and Ray (1996), Lang and Welzel (1996), Hughes and Mester (1998), Maudos (1998), Mester (1996), Rogers (1998), and Vander Vennet (1996). Other studies choose the Cobb-Douglas function [Cooper (1980), Fanjul and Maravall (1985)] or the Fourier Flexible form [Altunbas (1997), DeYoung and Hasan (1998)] to model the banks' technology. A disadvantage of the Cobb-Douglas specification with respect to the other two is that it implies a stronger restriction on the set of technologies which can be borne out by the data. For example, the elasticity of substitution between inputs is restricted to equal one for a Cobb-Douglas specification. Furthermore, it imposes restrictions on the substitutability between outputs and it does not allow for U-shaped average cost curves. For these reasons, the standard Cobb-Douglas functional form is less suitable to measure for instance economies of scope or scale.

Despite its disadvantages we choose the Cobb-Douglas specification in this study, but we augment it with a number of dummies so that for example scale economies can be captured. There are two reasons for this choice. First, taking the adjusted R^2 as a measure, the translog cost function specification does not explain our data better than the Cobb-Douglas specification, but the regression results of the translog specification are far more difficult to interpret.[9] Second, input prices changed considerably over the considered time period. Therefore it appears to be useful to distinguish the effect of

[9] As an example, using the translog cost function we obtained one significantly negative price coefficient and two which exceeded one.

price developments on costs and other effects such as technological progress. We found an appealing solution for this problem, but one which is not suitable for the translog specification.

Before continuing the discussion on the specific specification we took to estimate model 3.1 we would like to address a point of criticism which applies to bank efficiency studies in general, but is potentially severe in our case. In bank efficiency studies it is implicitly assumed that the variables used to measure production can be compared between banks. For example, if two banks offer an equal amount of customer loans it is assumed that both banks produce an equal amount of services to generate customer loans and thus incur the same costs. Needless to say, this need not be reasonable as the customer loans portfolios of these two banks may be different in nature and therefore need a different treatment.[10] In our study this problem is potentially severe: We focus on banks in the European Union and, although the implementation of the Second Banking Directive on 1 January 1993 implied a considerable harmonisation of the EU banking laws, there are still large structural differences between banking markets in the EU member states.

Specification

The specific Cobb-Douglas cost model which we adopt in this study is the following:

$$\frac{TC_{ti}}{TA_{ti}} = \gamma_0 \left(\frac{y_{ti,1}}{TA_{ti}}\right)^{\beta_1} \ldots \left(\frac{y_{ti,5}}{TA_{ti}}\right)^{\beta_5} p_{t1}^{\alpha_1} p_{t2}^{\alpha_2} p_{t3}^{\alpha_3} \sigma_1^{s_{1,ti}} \ldots \sigma_7^{s_{7,ti}} \delta_1^{t_1} \ldots \delta_4^{t_4} \gamma_1^{d_{1i}} \gamma_2^{d_{2i}} \gamma_3^{d_{3i}} + \varepsilon_{ti} \quad (3.2)$$

In equation 3.2, TA_{ti} are the total assets of bank i in period t, $s_1, \ldots s_7$ seven *size dummies*, $t_1, \ldots, t_4$ four *time dummies*, and $d_{1i}, \ldots, d_{3i}$ the values of three type dummies of bank i. In equation 3.2 we have incorporated that in our study we have $K = 5$ output variables and $J = 3$ input prices. For future reference, let w be the vector of parameters to be estimated

$$w = (\gamma_0, \beta_1, \ldots, \beta_5, \alpha_1, \alpha_2, \alpha_3, \sigma_1, \ldots, \sigma_7, \delta_1, \ldots, \delta_4, \gamma_1, \gamma_2, \gamma_3)$$

[10]There have been attempts to account for this problem. For example, Mester (1996) includes the average volume of non-performing loans as a measure for the quality of the loan portfolio.

Some important features of equation 3.2 are explained next, namely (i) the fact that we have scaled total costs and the output variables by dividing through total assets, (ii) the use of the dummy variables, and (iii) how we tackle the problem of a likely dependency between the time dummies and our input prices.

Scaling. We had two reasons to scale the total costs and output variables by the total amount of assets. The first reason can best be understood by looking at equation 3.1. First, in equation 3.1 it is likely that the regressors and the error term are dependent. In case a large bank, i.e. a bank with relatively high values of the output variables, is inefficient, it will in general deviate to a larger extent from the efficient frontier than a small inefficient bank. So, for an inefficient bank, the inefficiency component of will depend, among other things, on the bank's operating scale. In other words, the disturbances are not orthogonal to the regressors in model 3.1 and we cannot get consistent estimates without resorting to instrumental variable estimation or appropriately scaling the model. We chose the latter solution, dividing the left-hand side and the right-hand side of equation 3.1 by total assets, which, we think, is an appropriate proxy for size.

A second advantage of scaling arises from the fact that all output variables are in nominal Euro terms and thus are possibly subject to inflation during the sample period. A cost function requires the outputs to be in physical units, rather than nominal values. Without scaling our nominal output proxies are not comparable between different years. For example, if deposits of a bank in a specific year are twice as high as in the preceding year, but inflation has been 100 percent, it is unlikely that the amount of deposit services have also doubled. In this case the amount of deposits should be deflated and in our study the deflator is total assets. We do not know of any panel data study that addresses this important issue.

Dummies. A second important feature of model 3.2 is that the cost model is augmented with dummies to allow the cost frontier to be different for different sizes of banks, between different years (for example due to technological progress), and between different types of banks.

There are two reasons for including size dummies in model 3.2. First, scaling may not necessarily fully solve the problem of non-orthogonality of the error terms. Inefficiency may depend on size in a non-linear way. In this case, including size dummies further reduces the under-estimation of the unknown parameters. Second, by inserting enough size dummies we allow for a fairly general form of the average cost curve that for example allows for a U-shape. Thus, the Cobb-Douglas function with size dummies can exhibit decreasing, increasing or constant returns to scale in any order.[11] We divided our sample of into 8 non-overlapping size groups and, correspondingly, included 7 size dummies.[12] To explain the dummies by an example, the size dummy for the smallest group of banks s_1 takes on the value $s_{1,ti} = 1$ if bank i has a total assets amount smaller than EUR 100 million in year t ($TA_{ti} \leq$ EUR 100 million) and becomes 0 otherwise.

We also included 4 time dummies to allow for the possibility that the efficient frontier is different in each of the five years of our sample period. This is appropriate in a panel data setting because the frontier may change over time due to technological progress, changes in regulation or other structural developments. In our study we have included dummies for the years 1994, 1995, 1996 and 1997, so that the reference year is 1993.

Finally, we included three type dummies, distinguishing between four different types of banks. These type dummies are inspired by the difference in legal status of the banks. Details on the types are found in the data section. Different types of banks can have different cost structures. For instance banks may operate on different market segments

[11] A third, practical advantage of the size dummies is that economies of scale can directly be observed from the computed regression coefficients associated with the size dummies. In other efficiency studies, economies of scale are computed by differentiating the estimated cost function.

[12] Group 1: total assets $\leq$ 100 million ECU
Group 2: 100 million ECU $<$ total assets $\leq$ 300 million ECU
Group 3: 300 million ECU $<$ total assets $\leq$ 600 million ECU
Group 4: 600 million ECU $<$ total assets $\leq$ 1 billion ECU
Group 5: 1 billion ECU $<$ total assets $\leq$ 5 billion ECU
Group 6: 5 billion ECU $<$ total assets $\leq$ 10 billion ECU
Group 7: 10 billion ECU $<$ total assets $\leq$ 50 billion ECU
Group 8: 50 billion ECU $<$ total assets.
In the remainder of the text we will write EUR instead of ECU, although the basket of currencies that constituted the ECU is slightly different from the one that constituted the Euro.

which require a more costly or less costly treatment of the financial products offered. Also it may be that differences in the regulatory environment, faced by the various types of banks, affect their costs. We appreciate that a third possibility could be that some types of banks may simply be X-inefficient in which case it would not be appropriate to include type dummies.[13]

In equation 3.2 we have introduced the dummies such that all dummy parameters $(\sigma_1, ..., \sigma_7, \delta_1, ..., \delta_4, \gamma_1, \gamma_2, \gamma_3)$ become one if there are no economies of scale, no technological innovation or other structural changes, and equal cost structures across different types of institutions. By contrast, if the banks in size class j, say, have significantly lower (higher) costs than the banks in the reference class the parameter estimate of the respective size dummy, σ_j^*, will be significantly smaller (larger) than unity.[14] Likewise, a parameter estimate for a time dummy δ_k^* which is significantly smaller (larger) than unity indicates that in year k costs have generally been lower (higher) than costs in the reference year 1993, for example due to technological progress. Finally, parameter estimates of type dummies which are significantly smaller (larger) than unity show that efficient banks of the corresponding type have lower (higher) costs over total assets than efficient banks in the benchmark class.

Price changes versus changes in the frontier over time. The joint presence of input prices and time dummies in model 3.2 gives rise to a problem with the interpretation of the estimated parameters. Input prices can show general time patterns in which case the time dummies are correlated with the price variables. For example, for our data set, the price of funds (the interest rate) decreased over time whereas the price of buildings and labour steadily increased. As a consequence the time dummies will not only pick up

[13]Besides economic considerations, there are two econometric arguments for putting in type dummies. First, in the presence of substantial efficiency differences between the various types of banks, the number of banks on the cost frontier might become too small to generate reliable parameter estimates. Second, if it were the case that X-efficiency and types are indeed correlated then by omitting type dummies we obtain a correlation between the regressors and the error term (assuming that different types of banks have a different output mix). This invalidates the regression results.

[14]Here and in the rest of the paper we adopt the notation that parameters with superscript * represent estimated parameter values.

the effects of technological progress and the like, but also these price changes. We solved this problem by first regressing each input price on a constant and the time dummies. In other words, before estimating the model we performed the following three auxiliary regressions:

$$\ln p_{tj} = \eta_{0j} + t_1\eta_{1j} + t_2\eta_{2j} + t_3\eta_{3j} + t_4\eta_{4j} + dp_{tj},\ j = 1, ..., 3 \quad (3.3)$$

Here $\eta_{0j}, \eta_{1j}, \eta_{2j}, \eta_{3j}$ and η_{4j} are the unknown parameters of the constant and the four time dummies $t_1, ..., t_4$, respectively, and dp_{tj}, $t = 1, ..., T$ are the errors. These errors can be interpreted as the deviation of the prices from their time pattern. At first glance, regression 3.3 looks strange as the notation suggests that we have five observations and as many unknown parameters. Notice, however, that we have different price observations for each EU-15 country. Hence, the equation detects a general (EU-15) time pattern in each of the input prices.[15] Estimation of 3.3 yields:

$$dp^*_{tj} = \ln p_{tj} - \eta^*_{0j} - t_1\eta^*_{1j} - t_2\eta^*_{2j} - t_3\eta^*_{3j} - t_4\eta^*_{4j},\ j = 1, ..., 3 \quad (3.4)$$

By construction, the estimated deviations in the prices in equation 3.4 are orthogonal to the time dummies. Therefore, in model 2 we can separate price effects on total costs from time related effects such as technological progress by using the results of auxiliary equation 3.3.[16] In particular, substituting equations 3.3 in model 2 we obtain:[17]

$$\ln\left(\frac{TC_{ti}}{TA_{ti}}\right) = c + \beta_1 \ln\left(\frac{y_{ti,1}}{TA_{ti}}\right) + ... + \beta_5\left(\frac{y_{ti,5}}{TA_{ti}}\right) + \alpha_1 dp^*_{t1} + \alpha_2 dp^*_{t2} + \alpha_3 dp^*_{t3}$$
$$s_{1,ti}\kappa_1 + ... + s_{7,ti}\kappa_7 + t_1\lambda_1 + ... + t_4\lambda_4 + d_{1i}\pi_1 + d_{2i}\pi_2 + d_{3i}\pi_3 + \eta_{ti} \quad (3.5)$$

where

$$c = \ln\gamma_0 + \eta^*_{01}\alpha_1 + \eta^*_{02}\alpha_2 + \eta^*_{03}\alpha_3 \quad (3.6)$$

[15] A country subscript to the parameters of equation 3.3 was omitted for notational convenience.

[16] We have just sketched how we separated possible price effects from time-related effects. This, in fact, was an important consideration that let us to choose the more restrictive Cobb-Douglas specification. For the translog cost function the number of explanatory variables in the regression equation, after substituting the results of the auxiliary regressions, explodes.

[17] In order to let the logarithms be well-defined we added 1 to TC_{ti} and $y_{ti,j}$, $j = 1, ..., 5$ for all banks and all periods.

$$\kappa_k = \ln \sigma_k, \; k = 1, ..., 7 \tag{3.7}$$

$$\lambda_1 = \ln \delta_1 + \eta^*_{11}\alpha_1 + \eta^*_{12}\alpha_2 + \eta^*_{13}\alpha_3 \tag{3.8}$$

$$\lambda_2 = \ln \delta_2 + \eta^*_{21}\alpha_1 + \eta^*_{22}\alpha_2 + \eta^*_{23}\alpha_3 \tag{3.9}$$

$$\lambda_3 = \ln \delta_3 + \eta^*_{31}\alpha_1 + \eta^*_{32}\alpha_2 + \eta^*_{33}\alpha_3 \tag{3.10}$$

$$\lambda_4 = \ln \delta_4 + \eta^*_{41}\alpha_1 + \eta^*_{42}\alpha_2 + \eta^*_{43}\alpha_3 \tag{3.11}$$

$$\pi_l = \ln \gamma_l, \; l = 1, 2, 3 \tag{3.12}$$

η_{ti} is the new error term (3.13)

For any cost function the sum of the input price elasticities equals unity. We thus estimated model 3.5 under the restriction:

$$\sum_{j=1}^{3} \alpha_j = 1 \tag{3.14}$$

After estimating equation 3.5 the parameters of interest, i.e. vector w above, can be reconstructed using relationships 3.6—3.12. However, computing the variances of the parameters of interest proved more difficult. The interested reader can find the calculations of the variances in Appendix 3.

Efficiency

Once parameter estimates of the variables of interest are obtained, it is relatively straightforward to compute the X-efficiencies and size efficiencies of the banks in the sample. In particular, regarding X-efficiency, let us first define $TC_{ti}^{\min}$ as the estimated cost level of bank i in year t if it were on the efficient frontier:

$$TC_{ti}^{\min} = \left(\frac{TC_{ti}}{TA_{ti}}\right)^* TA_{ti}$$

A measure for X-efficiency is then given by the fraction $\frac{TC_{ti}^{min}}{TC_{ti}}$. X-*inefficiency* represents the distance of a particular firm to the efficient frontier, or

$$\text{X-ineff}_{ti} = \left(1 - \frac{TC_{ti}^{\min}}{TC_{ti}}\right) \tag{3.15}$$

Regarding size efficiency, let us define $\sigma^{\min}$ to be the value of the size dummy of banks in the size class with minimum costs, i.e. $\sigma^{\min} = \min\{1, \sigma_1, ..., \sigma_7\}$. Our measure of *size-inefficiency* is defined as:

$$\text{S-ineff}_{ti} = \left(1 - \frac{\sigma^{\min}}{\sigma_1^{s_{1,ti}} ... \sigma_7^{s_{7,ti}}}\right) \tag{3.16}$$

In this study we choose to apply formula 3.16 only to banks which are in a size class with *significantly* higher costs then the optimal size group. By contrast, when the respective size dummy turns out not to be significantly different from the optimal scale dummy, then we take

$$\text{S-ineff}_{ti} = 0$$

3.3 The Estimation Technique

Estimating the cost frontier, i.e. equation 3.5, requires non-standard regression techniques instead of OLS or its generalisations.[18] The reason for this is that we look at the minimum cost incurred instead of the average costs. In this paper we adopt an estimation method which takes into account that deviations from the frontier may emerge due to inefficiency but also due to other temporary bank specific reasons, such as for example re-organisation costs or simply bad luck or good luck. However, our method is different from other so-called *stochastic* or *thick* frontier approaches which depart from the same principle.[19]

[18]Instead of estimating a stochastic frontier one could also employ Data Envelopment Analyses in order to find the close fitting frontier which envelops all data points (see among others Charnes et al. (1994)).

[19]See, for instance, Berger and Humphrey (1992).

A thorough exposition of our new method, the so-called Recursive Thick Frontier Approach (RTFA), is given in Wagenvoort and Schure (1999). The intuition behind the method is best understood by comparing it to the stochastic frontier analysis (SFA), a standard technique introduced by Aigner, Lovell and Schmidt (1977). This method decomposes the disturbance of model 3.5 into two components. One component is assumed to represent the noise term (usually modelled by the normal distribution) and the other component reflects the inefficiency part (usually modelled by the half-normal or exponential distribution). The Maximum Likelihood procedure is used to estimate the model. Although it is common in efficiency studies to adopt SFA, it is open to three main criticisms. First, the outcomes of the study crucially depend on the *a priori* distributional assumptions made in SFA, and it remains unclear how these assumptions can be tested. Second, adding a half-normal or exponential component to the disturbances of the regression model is often not an adequate way of capturing inefficiency. For example, for our bank sample SFA gives parameter estimates which are close to the OLS estimates, suggesting that all deviations from the regression line are random and not due to differences in performance. At the same time it seems that many efficiency differences do exist as relatively many banks are located below the regression line throughout the sample period, whereas many others banks are always located above the regression line. It is highly unlikely that this can be put down to bad and good luck. Third, SFA is highly sensitive to outlying observations.

The method that we propose is less vulnerable to the criticisms mentioned above. Instead of making the usual distributional assumptions we use the observation that the probability of an efficient bank of being at either side of the cost frontier should be equal to one half. Our RTFA procedure works as follows. We start with the full sample of banks and estimate model 3.5, for example using OLS, as if all banks were equally X-efficient. A test statistic is then computed to evaluate whether the selected banks are randomly distributed at either side of the regression line. If this is not the case, i.e. if our test statistic rejects that on average the probability for a bank to be above or below the regression line is 0.5, we reduce the group of banks by eliminating those banks which are positioned relatively far above the regression line (i.e. the banks with relatively high costs given their output mix and the input prices they face). Then the second iteration begins, re-estimating model 3.5 on the basis of this reduced sample. The algorithm iterates until the test statistic does no longer reject the null hypothesis, i.e. when the largest possible group of relatively efficient firms has been identified.

Our method is only suitable for panel data. The time dimension of panel data enables to require information on the persistence of some banks of having lower cost than others. We fear that with only cross-sectional data it will always be difficult to distinguish between luck and efficiency.

RTFA is also less sensitive to outlying observations than SFA. First, the parameters of the cost frontier are estimated by considering only the observations associated with the X-efficient companies. Outliers in the observations associated with the X-inefficient banks thus cannot spoil the cost frontier regression. Evidently, outliers may also occur in the group of banks with relatively low costs. We therefore employ the one-sided trimmed least squares estimator when estimating model 3.5 for the group of X-efficient banks (see Wagenvoort and Schure (1999) for more details).[20]

3.4 The Data

Bank Selection

Our main data source is 'BankScope' of Bureau van Dijk, a data set with bank

[20] Naturally, just as with SFA there can be practical problems with RTFA. For example, our full sample of firms contains relatively many German savings banks, so it could happen that the cost frontier is solely determined by these institutions. Our regression results for the full sample of firms reveal that this problem does not occur. For the separate regression including only savings banks, however, German savings and cooperative institutions did put their stamp on the shape of the cost frontier. We solved this particular problem by doing the regression for a smaller sample of savings banks which included, besides all the savings banks in the other EU countries, only 150 (randomly chosen) German savings banks.

data from annual reports and rating agencies. Also we made use of the 1998 edition of 'Bank Profitability' of the OECD, the International Financial Statistics of the IMF (IFS), Datastream International, and the CRONOS data set of Eurostat. Below we will describe how we selected our data from BankScope, and how we defined the variables for the cost function estimation.

The focus of our study is on *credit institutions*, as defined in the two European Economic Community (EEC) Council Directives on the "business of credit institutions". Both these so-called *Banking Directives* define a credit institution as "an undertaking whose business is to receive deposits or other repayable funds from the public and to grant credits for its own account" (First banking Directive, 1977). To translate this in practical BankScope terms, we selected "Commercial Banks", "Savings Banks", "Cooperative Banks", "Real Estate/Mortgage Banks", "Medium & Long Term Credit Banks", and "Non Banking Credit Institutions". We will use the terms 'banks' and 'credit institutions' interchangeably.

We took banks in the European Union countries (EU-15) for which yearly data for the period 1993-1997 is available. We have focused on this period as on 1 January 1993 the Second Banking Directive (1988) came in force implying a large degree of deregulation in the European Union. Also relatively many banks are available in BankScope for this period.[21] We selected all consolidated statements, unconsolidated statements, and some so-called aggregate statements.[22] Applying the selection criteria above led to a first group of 2185 banks with data for 5 years.

We put considerable effort into cleaning this first selection of banks. First, some legal

[21] The bank set only includes 'living banks'. BankScope calls a bank 'living' when it continues to exist as a legal body. So a bank dies when it goes bankrupt or when its activities are brought in another bank. Ownership structure has little to do with whether a bank is a living bank. In case of a takeover a bank often continues to exist as a legal entity, and therefore remains a living bank. The focus on living banks might imply a sample selection bias. However, including banks that have died leads to missing observations in our sample which are very difficult to deal with econometrically.

[22] In BankScope terms, we selected statements with consolidation codes C1, C2, U1, U2 and A1. Aggregated statements are generated by BankScope by adding up the statements of a group of affiliated banks, which, however, have no financial links between them, nor form a legal entity.

entities appeared twice in the data set as BankScope publishes both their consolidated statement and their unconsolidated statement. Often this was no problem when the consolidated balance sheet contained many more assets –and thus is of a different nature– than the unconsolidated one. However, sometimes the consolidated and the unconsolidated statement looked similar, in which case more or less the same balance sheet had entered the first selection twice. In order to prevent this we eliminated banks for which the total assets on the unconsolidated statements exceeded 70 percent of the total assets on their consolidated balance sheet. We also removed banks which report zero or negative interest expenses or operating expenses in one or more years.[23] Third, we removed two banks for which we found that individual balance sheet items exceeded the balance sheet

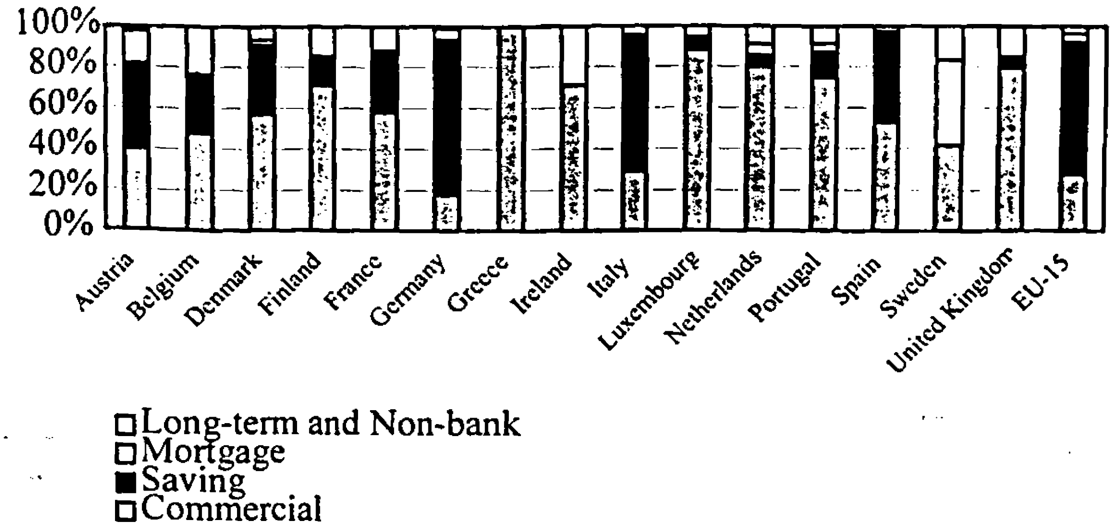
Table 3-1: Number of credit institutions in the EU-15 analysed in this study

Country	Commercial	Savings	Mortgage	Long-term and Non-bank	Total
Austria	20	21	8	1	50
Belgium	33	19	1	16	69
Denmark	47	28	2	5	82
Finland	5	1	0	1	7
France	171	86	3	35	295
Germany	156	673	49	8	886
Greece	17	0	0	0	17
Ireland	5	0	0	2	7
Italy	57	129	0	8	194
Luxembourg	86	5	1	5	97
Netherlands	28	2	2	3	35
Portugal	18	3	1	2	24
Spain	66	55	1	3	125
Sweden	5	0	5	2	12
United Kingdom	59	3	1	11	74
EU-15	773	1025	74	102	1974
Source: BankScope					

totals (erroneous data). Fourth, we left out some outlier observations. In particular, we omitted banks reporting an amount of 'off-balance sheet items' of more than twice

[23]Sometimes the data suggest that in a particular year interest expenses may indeed have been zero. In this case we left the bank in the sample.

Figure 3-1: The legal status of banks in the EU (percentages)



the balance sheet total, and a bank which reported the regular operating and interest expenses to be more than $2\frac{1}{2}$ times the balance sheet total in a particular year. While we recognise that the data for these few banks was not necessarily wrong, these banks would have formed 'leverage points' which could have seriously distorted the regression results.

Table 3-1 reports the country of origin and the type of the 1974 banks which were left over after cleaning the data. In the table we have grouped the banks into four categories: Commercial Banks (Commercial), Savings Banks and Cooperative Banks (Savings), Real Estate/Mortgage Banks (Mortgage), and Medium & Long Term Credit Banks and Non Banking Credit Institutions (Long-term and Non-bank). We will follow this classification throughout the rest of the paper. From this table, which fairly well covers the overall European banking industry, and Figure 3-1 it can be seen that the structure of the banking sectors of the EU-15 countries varies considerably. In particular, Austria, Germany, Italy and Spain have relatively many savings banks (more than 40 percent of the total). On the other hand, in Ireland, Greece, Luxembourg, The Netherlands, Sweden and the

UK, less than 10 percent of the credit institutions of our sample are savings banks.[24]

Definition of the Variables

All data for total costs and output comes from Bankscope. The banks' total costs are defined as the sum of 'interest expense', 'total operating expense' and 'commission expense'.

We have defined 5 output variables: customer deposits, loans, equity investments, off-balance sheet items, and other services. *Customer deposits* comprise demand, savings and time deposits.[25] The variable *loans* is created by taking 'total loans' in BankScope and subtracting 'loans to municipalities / government' and 'loans to group companies / associates'. The latter two variables are subtracted as we suspect that relatively few actions need be undertaken when offering loans to these groups of borrowers. Consequently these assets do not significantly lead the bank to incur costs. We share the opinion that mortgages may also imply a different amount of services per unit than other loans and therefore should be treated as an separate output variable. However, unfortunately for most countries BankScope data does not separate mortgages from loans. *Equity investments* are obtained by adding up 'equity investments' and 'other investments'. These items comprise the book value of participations and shares in companies with related business and shares in other non-financial affiliates. Here we have to remark that in many cases this latter output can be substantially under-valued since its book value, as taken from Bankscope, is usually determined on the basis of historic costs instead of its market value. However, this does not necessarily pose as a problem in measuring financial services as long as banks use similar accounting techniques. Evidently, there is a potential danger of mis-measurement of the level of the output variable equity investments for our bank set. We have included this variable as we suspect that equity

[24]While not in our data set, there are a few savings banks in Sweden and Greece.

[25]As we were not sure whether demand and saving deposits on the one hand and time deposits on the other represent the same amount of service production per unit, we attempted to create two output variables rather than one. Unfortunately, however, the data for German banks do not allow for such an analysis.

investments imply activities such as the selection of the shares and active monitoring and risk-management. *Off-balance sheet items* correspond to 'off-balance sheet items' in BankScope which contains contingent liabilities arising from guarantees, irrevocable letters of credit, irrevocable facilities, discounted bills, etc. Derivatives are not included in this item. Like loans, off-balance sheet items force the bank to screen and monitor projects and hence provide services. Finally, the variable *other services* is equal to the variable 'commission revenue'. In order to keep the commission revenues in different years comparable we divide through a price index for banking services.[26] Appendix 1 explains how this price index is created. Contrary to all other output variables, which are stock variables on and off the balance sheet, other services is a flow variable taken from the profit and loss account.

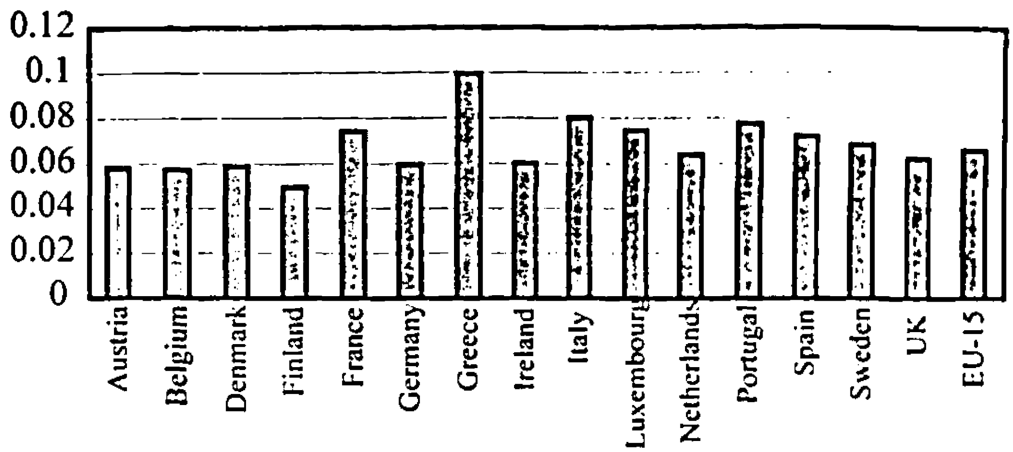
In our opinion the remaining earning and non-earning assets on the banks' balance sheet, such as securities, treasury and other bills, bonds, certificates of deposits, cash balances, and the like, do not require the provision of a significant amount of services and are hence not included in the output vector.

Three input prices have been defined: the price of loanable funds, the price of labour and the price of buildings.[27] The *price of funds* is obtained by taking a weighted average of the average 3-month interbank rate and the deposit rate. This data is obtained from Datastream International and IFS, respectively. The weights are determined by the amount of deposit funding as part of total funding (total assets) of each bank. The *price*

[26]Obviously, changes in prices of the other output variables can be relevant too. Unfortunately no adequate data on these prices are available. As mentioned in section 3, we minimise this problem by scaling all the output variables, including the deflated commission revenue, and total costs by total assets. One could argue that, in the special case of the output variable *other services*, this scaling is redundant since we already divide through a price index. Recall however that there is also an econometric argument for it since scaling reduces the problem that the model errors are not orthogonal to the regressors in a cost model specification and on that score the fundamental orthogonality condition is not fulfilled. Division through price indices cannot solve the latter problem.

[27]The reader could correctly point out that banks face more prices when acquiring their inputs. We think, though, that the three prices we have included are the most important input prices. Also, our assumption here is not so much that the bank faces only these three prices, but that a linear combination of these can sufficiently well approximate the prices that the bank might face. Other bank efficiency studies typically also include up to three prices.

Figure 3-2: Costs over total assets in the European Union, full sample



of labour represents the average wage rate in the banking sector in each country. The data needed to construct an index for the price of labour is taken from BankScope and the OECD. The *price of buildings* is created by taking an appropriate price index for newly delivered buildings and correcting it for the relative price levels in each country. A detailed description on these calculations is given in Appendix 1.

Descriptive Statistics

Figure 3-2 shows that across Europe there are considerable differences in the cost levels. The ratio of costs over total assets ranges for most countries between 5 and 8 percent with Greece being a striking outlier. Average costs are also relatively high in France, Italy, Luxembourg and Portugal when compared with the EU-15 average.[28] Figure 3-3 shows that in all European countries the costs per unit of assets substantially decreased over the period 1993-1997. For the overall European banking industry average costs fell by about 25 percent.

It would be premature to conclude from Figure 3-2 and Figure 3-3 that Greek banks

[28]The EU-15 averages in Figures 1 to 10 are constructed by applying country weights on the basis of the share of each country in total European assets.

are more inefficient than other European banks or that the performance of European banks has improved over time. For testing these kinds of hypotheses we have to take into consideration changes in the input prices and changes in the level and mix of the outputs. For example, the decrease in cost could be fully due to the fact that in the same period the price of funds decreased substantially. Figure 3-3 to Figure 3-6 show the developments of average costs and the fund rate for the EU-15, Italy, Germany, and the UK. Evidently, given the sharp fall in the price of funds in Europe one may expect substantially lower average costs for banks in general.

Substantial differences across the banking industry are also revealed by looking at the banks' output structure in the respective European countries. Figure 3-7 shows the decomposition of the earning assets. Equity investments are relatively small compared to 'other earning assets' (treasury and other bills, etc) and loans. On average, equity investments are less than 2 percent of total assets, whereas 50 percent of the balance total consists of loans and mortgages. Luxembourg and Greece have relatively many other earning assets.

Let us turn to the other three outputs of our cost model: deposits, off- balance sheet activities, and services related to other activities (brokerage services). The ratio of deposits over total assets reveals remarkable differences in country output structures (see Figure 3-8). Both the ratio of off-balance sheet items to total assets, and the ratio of brokerage services to total assets, reveal two outliers with respect to financial product mix. In 1997, Belgium and Swedish banks report relatively high off-balance sheet activities (see Figure 3-9). Furthermore, brokerage services are relatively high in Ireland and the UK (see Figure 3-10). Note that for Greece earned brokerage fees are not reported in BankScope. We should bear all these facts in mind when discussing the cost efficiency of banks in the next section.

Figure 3.3: Costs over total assets and the interbank rate in the EU-15

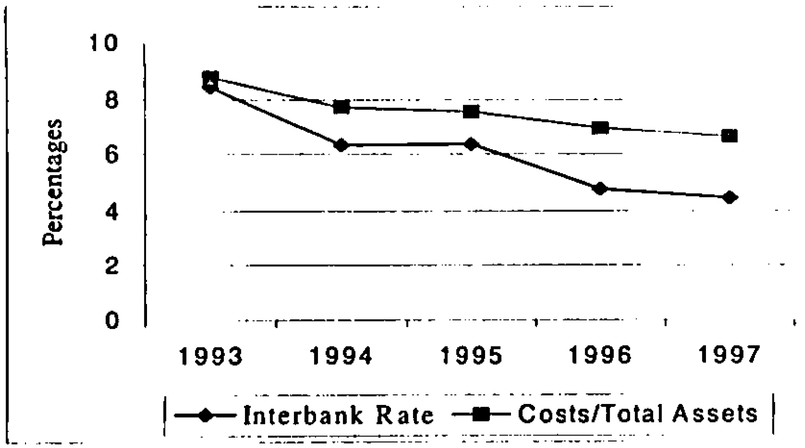


Figure 3.4: Cost over total assets and the interbank rate in Italy

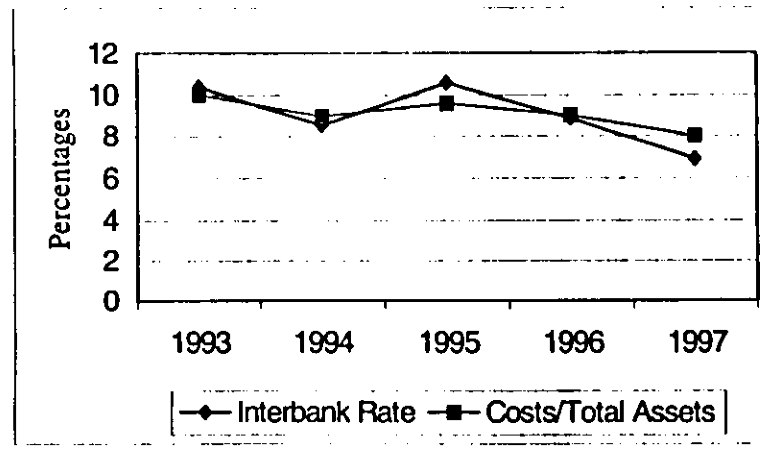


Figure 3.5: Costs over total assets and the interbank rate in Germany

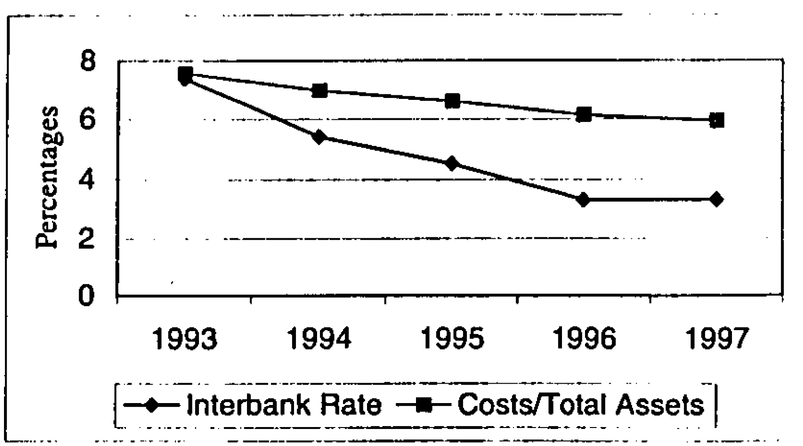


Figure 3.6: Costs over total assets and the interbank rate in the UK

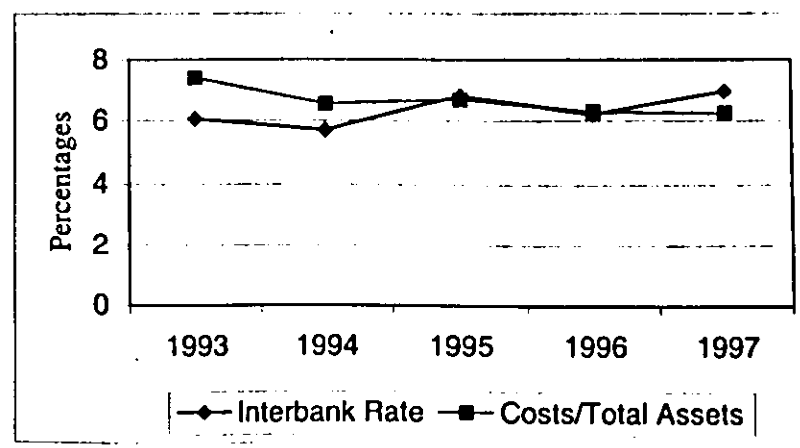


Figure 3.7: Earning assets (outputs over total assets), 1997[29]

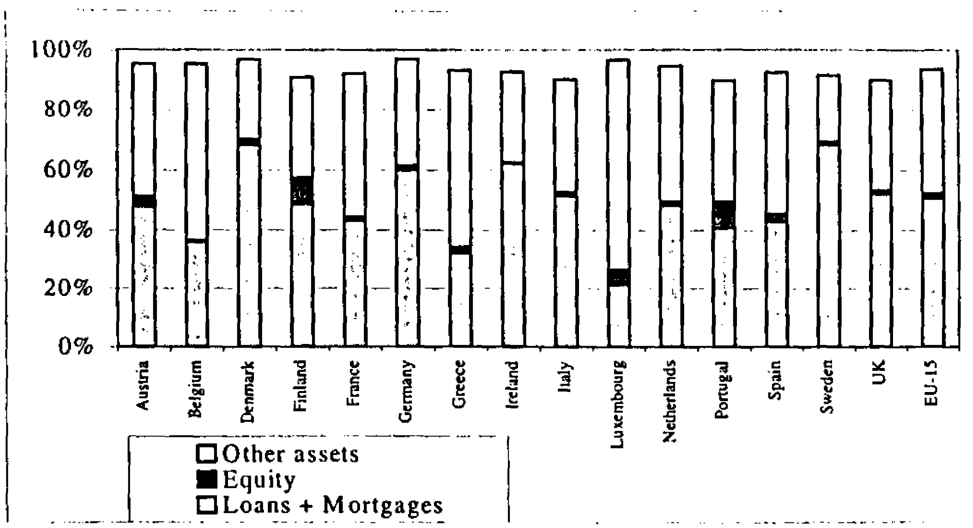


Figure 3.8: deposits over total assets, 1997

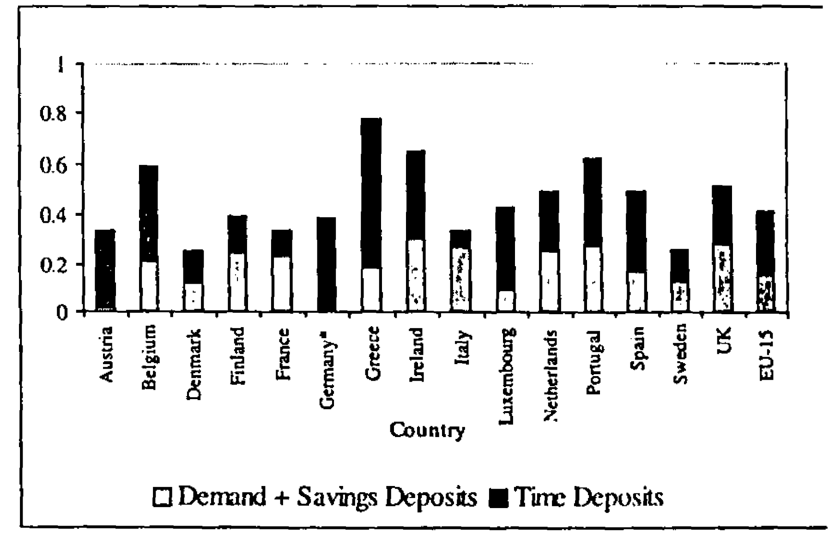


Figure 3.9: Off-balance sheet activities over total assets, 1997m

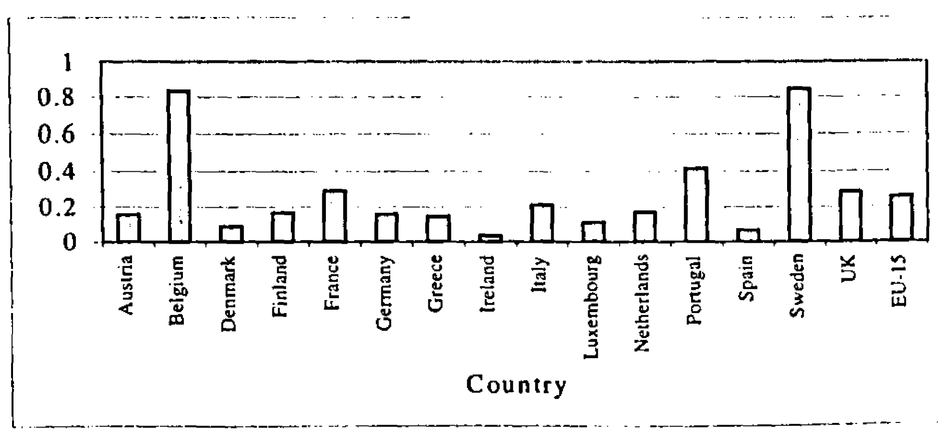
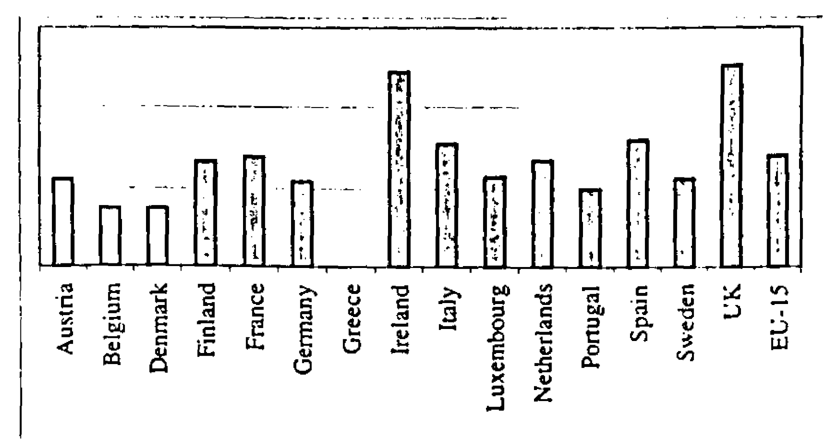


Figure 3.10: Brokerage fees over total assets, 1997



[29] Note: The fixed capital stock, which contains for example buildings, is not explicitly shown in this graph. This explains why the columns do not add up to 100 percent.

3.5 Empirical Evidence Regarding the Efficiency of European Banking

We performed separate regressions for the full bank sample, for commercial banks, and the sample of savings banks. Due to data limitations we could not perform separate regressions for mortgage banks or long-term and non-bank credit institutions. However, before turning to the regression results several remarks are in order. The first remark relates to how well our Cobb-Douglas specification manages to explain the data. Taking Theil's adjusted R^2 as our measure of fit, we see that the model best explains the data of savings banks. In this case, our model explains 81 percent of the variation of total costs over total assets, compared to 44 percent and 43 percent for the full sample and the commercial banks, respectively. Table 3-2 and Table 3-3 contain the results for the full sample and the commercial banks respectively while Table 3-4 shows the results for the savings banks. The differences in explanatory power might be explained by the fact that savings banks make up a more homogeneous group of banks. Besides that the group of commercial banks is more heterogeneous, within this category, some banks possibly exploit better economies of scope than others, leading to cost differences which the Cobb-Douglas specification is unable to detect.

Based on R^2, it is difficult to judge whether our model explains the data well compared to other studies. In recent efficiency studies of banks it is not unusual to report an adjusted R^2 exceeding 0.95. However, as we have mentioned before, in these studies the regression model is not scaled by total assets. Hence, to a considerable extend the high R^2 of these models is due to the fact that 'big' banks face higher costs. Related to the reliability issue, for the full sample regression we find that 340 of our 1974 credit institutions in Europe are on the cost frontier, 173 of which are German.[30] This implies that the cost frontier is based on 340*5=1700 observations. As far as the number of

[30] Table A2-1 reports exactly how many banks are on the frontier and which are their home countries for the full sample, the commercial banks and the savings banks.

Table 3-2: The estimated Cobb-Douglas cost frontier for the period 1993–1997, full sample

	Parameter Estimate	t-value
Constant	0.0016*	5.92
Deposits over Total Assets	0.0248*	9.75
Loans over Total Assets	0.0693*	20.31
Equity Investments over Total Assets	0.0114*	7.65
Off-balance Sheet over Total Assets	-0.0075*	-4.91
Brokerage over Total Assets	0.0993*	26.57
Price of Funds	0.4256*	46.25
Price of Labour	0.2345*	9.64
Price of Buildings	0.3398*	13.79
Dummy, Total Assets ≤ 100 Million ECU	1.0745*	3.42
Dummy, 100 Million < Total Assets ≤ 300 Million	1.0601*	3.17
Dummy, 300 Million < Total Assets ≤ 600 Million	1.0147	0.76
Dummy, 600 Million < Total Assets ≤ 1 Billion	1.0236	1.18
Dummy, 1 Billion < Total Assets ≤ 5 Billion	1.0248	1.35
Dummy, 10 Billion < Total Assets ≤ 50 Billion	1.0614*	2.70
Dummy, 50 Billion < Total Assets	1.0436	1.71
Dummy Commercial Banks	0.9620*	-4.63
Dummy Mortgage Banks	0.8052*	-6.25
Dummy M-LT & NB Credit Institutions	0.7958*	-8.19
Adjusted Coefficient of Determination	0.44	
Binomial Test	3.73	
Number of banks on the cost frontier	321	

Notes:

(1) The regressand is 'total costs over assets'.

(2) Brokerage is scaled by the annual average index of the price of banking services in the respective countries.

(3) Type dummies are defined with respect to savings and cooperative banks.

(4) Time dummies are defined with respect to the year 1993.

(5) Size dummies are defined with respect to the class: 5 Billion < Total Assets ≤ 10 Billion ECU.

(6) We have adopted the RTFA method. The "Binomial Test" statistic asymptotically converges to a Chi-squared distribution with one degree of freedom.

(7) The price of funds is computed as the weighted average (according to the relative amount of deposits in total assets) of the (yearly average of the) 3-months LIBOR and the deposit rate.

(8) $\chi^2_{0.01}(1) = 6.63$ and $t(\infty)_{0.025} = 1.96$. Parameters significant at the 95 percent confidence level are marked with an asterisk.

Table 3-3: The estimated Cobb-Douglas cost frontier for the period 1993–1997, commercial banks

	Parameter Estimate	t-value
Constant	0.0076*	3.42
Deposits over Total Assets	0.0045	1.61
Loans over Total Assets	0.0525*	7.92
Equity Investments over Total Assets	0.0156*	6.46
Off-balance Sheet over Total Assets	-0.0154*	-6.18
Brokerage over Total Assets	0.1618*	22.02
Price of Funds	0.5602*	32.87
Price of Labour	0.1413*	3.57
Price of Buildings	0.2985*	7.99
Dummy, Total Assets ≤ 108 (ECU)	0.7748*	-7.28
Dummy, 108 < Total Assets ≤ 3*108	0.9126*	-3.87
Dummy, 3*108 < Total Assets ≤ 6*108	0.9412*	-2.58
Dummy, 6*108 < Total Assets ≤ 109	1.0119	0.46
Dummy, 0.5*1010 < Total Assets ≤ 1010	0.9013*	-3.10
Dummy, 1010 < Total Assets ≤ 0.5*1011	1.1015*	4.21
Dummy, 0.5*1011 < Total Assets	0.9798	-0.76
Adjusted Coefficient of Determination	0.43	
Binomial Test	5.10	
Number of banks on the cost frontier	143	

Notes:

(1) The regressand is 'total costs over assets'.

(2) Brokerage is scaled by the annual average index of the price of banking services in the respective countries.

(3) Type dummies are defined with respect to savings and cooperative banks.

(4) Time dummies are defined with respect to the year 1993.

(5) Size dummies are defined with respect to the class: 1 billion < Total Assets ≤ 5 billion.

(6) We have adopted the RTFA method. The "Binomial Test" statistic asymptotically converges to a Chi-squared distribution with one degree of freedom.

(7) The price of funds is computed as the weighted average (according to the relative amount of deposits in total assets) of the (yearly average of the) 3-months LIBOR and the deposit rate.

(8) $\chi^2_{0.01}(1) = 6.63$ and $t(\infty)_{0.025} = 1.96$. Parameters significant at the 95 percent confidence level are marked with an asterisk.

Table 3-4: The estimated Cobb-Douglas cost frontier for the period 1993-1997, savings and cooperative banks

	Parameter Estimate	t-value
Constant	7.21E-05*	4.35
Deposits over Total Assets	0.0299*	4.31
Loans over Total Assets	0.2119*	21.87
Equity Investments over Total Assets	0.0090*	4.53
Off-balance Sheet over Total Assets	-0.0012	-0.87
Brokerage over Total Assets	0.0369*	9.58
Price of Funds	0.3718*	27.73
Price of Labour	0.6355*	20.88
Price of Buildings	-0.0073	-0.37
Dummy, Total Assets ≤ 108 (ECU)	1.1599*	10.40
Dummy, 108 < Total Assets ≤ 3*108	1.0314*	2.73
Dummy, 3*108 < Total Assets ≤ 6*108	1.0471*	4.23
Dummy, 109 < Total Assets ≤ 0.5*1010	1.0168	1.63
Dummy, 0.5*1010 < Total Assets ≤ 1010	1.0015	0.11
Dummy, 1010 < Total Assets ≤ 0.5*1011	1.0130	0.96
Dummy, 0.5*1011 < Total Assets	1.1169*	5.97
Dummy 1997	0.9102*	-2.61
Dummy 1996	0.9483	-1.45
Dummy 1995	0.9538	-1.33
Dummy 1994	0.9819	-0.51
Adjusted Coefficient of Determination	0.81	
Binomial Test	6.53	
Number of banks on the cost frontier	147	

Notes:

(1) The regressand is 'total costs over assets'.

(2) Brokerage is scaled by the annual average index of the price of banking services in the respective countries.

(3) Type dummies are defined with respect to savings and cooperative banks.

(4) Time dummies are defined with respect to the year 1993.

(5) Size dummies are defined with respect to the class: 6*108 < Total Assets ≤ 109.

(6) We have adopted the RTFA method. The "Binomial Test" statistic asymptotically converges to a Chi-squared distribution with one degree of freedom.

(7) The price of funds is computed as the weighted average (according to the relative amount of deposits in total assets) of the (yearly average of the) 3-months LIBOR and the deposit rate.

(8) $\chi^2_{0.01}(1) = 6.63$ and $t(\infty)_{0.025} = 1.96$. Parameters significant at the 95 percent confidence level are marked with an asterisk.

degrees of freedom can tell, the regression results are thus reliable.

A second remark is that one should be careful when interpreting our parameter estimates. Especially the estimated coefficients of the input prices and the outputs are not necessarily the impact on the bank's average costs of changing the variable in which one is interested (i.e. the partial effect). The reason is that some input prices and outputs may be significantly cross-correlated.[31] To give an example, it is likely that both the price of labour and the price of buildings are correlated to inflation. The same can be true for some outputs. It may well be that the outstanding amount of loans of (commercial) banks is strongly related to their off-balance sheet activities.

A related remark is that in such cases insignificant parameter estimates need not imply that the impact of the corresponding variable is insignificant. For instance, for commercial banks we find that the parameter estimate of the price of labour is not significant in the regression which includes the time-dummies. However, this does not necessarily indicate that the price of labour is irrelevant to the costs of a bank. A significant correlation between the price of labour and the price of buildings may drive this result.

A final remark is made about model reduction. In the cases of the full sample and the group of commercial banks, none of the time dummies are significantly different from 1 on the 95 percent confidence level.[32] We therefore repeated the regressions without including time dummies in model 3.5. The results of which are reported in Tables 2 and 3 respectively. The parameter estimates of the complete model for the savings banks are shown in Table 3-4.

Input Prices and Costs

It becomes immediately clear from Table 3-2, which gives the estimated cost frontier

[31] If, by contrast to the situation sketched above, a particular variable is little correlated to the remaining variables in model 3.5, the corresponding parameter estimate can be interpreted as an elasticity. This is a standard feature of log-linear models.

[32] These regression results are not reported but are available on request.

for the full sample of banks, that the most important input price of banks is the price of loanable funds. This confirms our impression of a close relationship between costs per asset and the price of funds (shown in Figures 3 to 6). For the full sample we get a coefficient of 0.43 and we believe that this figure can well be interpreted as the fund price elasticity of total costs over assets.[33]

We notice from Tables 3 and 4 that the costs of commercial banks are more sensitive to changes in the fund rate than the costs of savings banks. The price elasticities of loanable funds are 0.56 and 0.37 respectively. An explanation for the considerable difference in these estimates may be that savings banks rely more on long-term finance than commercial banks while our price of funds is based on two short-term rates.

We conclude with a key point. Given the big impact of the price of funds it is of crucial importance to take the correct price for this variable. If a price is taken which is inappropriate for a specific country, we will have incorrect estimates for the X-efficiency measures for the banks in that country.

In the full sample the coefficients of the price of labour and the price of buildings are 0.23 and 0.34 respectively. When splitting up the sample these estimates change substantially. We believe this is due to the high correlation between the two prices (they may both be driven by inflation). We therefore refrain from any interpretation of these results.

Bank Outputs and Costs

We will first discuss the relationships that we observe between the production of different outputs and average costs for the full sample of banks (Table 3-2). Then we pinpoint some differences in output elasticities between commercial and savings banks (Table 3-3 and Table 3-4).

For the full sample, brokerage and loans are the most important outputs. These

[33] We did a regression of the price of funds on the other two prices and found that in the sample period there was very little correlation between the price of funds and the other two prices.

explanatory variables enter the cost frontier with significant parameter estimates equal to 0.099 and 0.069 respectively. Although the estimate of deposits over total assets is smaller (0.025), deposits do also contribute significantly to explaining the bank's cost. This confirms our view that deposits have output characteristics. Equity investments appear with a positive significant coefficient of small magnitude (0.011).

It is striking that off-balance sheet items enter the regression equation with a significant negative coefficient (-0.008). This result is due to a strong positive correlation between loans and off-balance sheet items. Namely, when loans were left out from the model we found that the sign of the coefficient for off-balance sheet items reversed. A possible economic explanation for the reported negative coefficient of the off-balance sheet variable could be that a bank with many off-balance sheet items incurs less production costs per unit of loans relative to other banks. This could be for two reasons. First, the production costs per unit of a large loan is probably lower than the production costs of one unit of credit supplied to small lenders. Banks with a high level of credit commitments and guarantees are usually larger and may also have larger clients. Therefore, they provide loans of larger volume compared to banks with few off-balance sheet items. Second, off-balance sheet items contain many credit commitments. The loans provided by banks with many off-balance sheet items often stem from these credit commitments and these are usually only made to customers with an unviolated payment record. This means that the screening and monitoring expenses incurred are lower.

The results regarding the effects of output on cost become interesting once we compare the parameter estimates of the separate regressions for commercial and savings banks. The positive relationship between brokerage and costs is much more pronounced for commercial banks than for savings banks. For commercial banks the parameter estimate for brokerage services is 0.162 whereas for savings banks we find a value of 0.04. On the other hand, loans are more important for the cost function of a savings bank. For savings banks and commercial banks we find significant parameter estimates of 0.212 and 0.053, respectively. Finally, the coefficient for deposits is substantially lower for the commercial banks than for savings banks and not even significant. All these differences in output elasticities can be explained by the fact that commercial banks offer more credit lines and market related services such as security trading, risk management, underwriting of assets, etc.

Cost Differences Between Different types of Credit Institutions

The full sample regression results reveal that mortgage banks and long-term and non-bank credit institutions operate at significant lower costs than savings banks. In both cases the ratio of costs to total assets is about 20 percent lower than for savings banks. Structural differences between different credit institutions may underlie this result. For example, the nature of the outputs or the institutional environment of mortgage banks and long-term and non-bank credit institutions may fundamentally differ from savings banks. For this reason the cost differences mentioned above need not reflect differences in X-inefficiencies.

Our analysis also suggests that on average commercial banks operate at 4 percent lower costs than savings banks (see Table 3-2). Again this can be due to differences in structure or X-efficiency. For example, a difference in X-efficiency could occur when managers of savings banks have more discretion over the use of the bank's cash-flow. If this were the case, demutualisation of savings and cooperative banks would lead to lower costs.

Technological Progress

Has the cost frontier shifted over time in the sample period? For the full sample, none of the t-values of the time dummies are significant. We have, therefore, no evidence that the cost level of a typical efficient bank changes over time in the period 1993 - 1997. We find the same result for the commercial banks. By contrast, we see that costs over assets of an average efficient savings bank decreases over time. In particular, for X-efficient banks we find a steady reduction in the costs over total assets of about 2 percent each year. To be more precise, in the period 1993 - 1997 efficient savings banks reduced their costs by 9 percent.[34]

With our limited study we are not in the position to judge what are the driving forces behind the drop in costs for savings banks, and why this effect did not occur for commercial banks. One can think however of several explanations. As was mentioned above, savings banks are on average less efficient than commercial banks. The reduction in the cost per unit of assets of the group of managerial efficient savings banks, could simply reflect that these banks have reduced their distance to the even more efficient commercial banks. The possible reasons for observing such a rise in X-efficiency are numerous. For instance, small savings banks may reduce costs by centrally organising the acquisition of funds on the money markets or the portfolio management of securities. Within this view, German "Sparkassen" provide an illustrative example. Cost reductions can possibly also be ascribed to the implementation of new (computer) technology that facilitates data processing, data communication with other institutions, credit risk evaluation and decision-making. It is not unlikely that savings banks were slower in adopting the latest technology in comparison with commercial banks since the latter group of banks are usually more market orientated. Commercial banks may have started earlier with exploiting new technology in comparison with savings banks, but the returns have faded away or were offset by other structural changes. That does not mean that technological innovation such as Internet banking will have no impact on commercial banks in the future. However, for our sample period, technological progress was statistically irrelevant for commercial banks.

Size Inefficiencies

From the parameter estimates of the size dummies in Table 3-2 we find initially increasing returns to scale and afterwards constant returns to scale. The estimates for the size dummies initially decrease in the size class. For very large banks the dummy

---

[34] See the value of the 1997 time dummy in Table 3-4.

increases again and becomes significantly different from unity only for the size class with total assets between 10 and 50 billion but is insignificant for the largest banks in the sample with total assets above 50 billion. It therefore seems that only very small banks face higher costs than the reference class of banks. In particular, banks with less assets than EUR 100 million have approximately 7.5 percent higher costs per asset and banks with assets between EUR 100 and 300 million have approximately 6.0 percent higher costs per asset.

Turning to Table 3-4 for savings banks we clearly find a U-shaped average cost curve. This indicates that small savings banks face increasing returns to scale while very large banks have decreasing returns to scale. Savings banks with less assets than EUR 100 million have approximately 16 percent higher costs per asset than the savings banks falling in the reference class. Also the next two smaller size groups have significantly higher costs per asset of roughly 3 percent and 5 percent, respectively. After that there are constant average costs until we arrived at the ten very large savings banks with total assets exceeding EUR 50 billion. These banks have roughly 10 percent higher costs over assets than the medium-sized reference class. Summarising, small and very large savings banks can improve efficiency by choosing their total assets between EUR 600 million and EUR 50 billion.

For the group of commercial banks the size picture is much less transparent, as costs seem to jump up and down with increasing size class. In our view these rather strange results are due to the fact that commercial banks form a very diverse group of banks. Some small investment banks that offer a range of products which is substantially different from the average product mix, could belong to this group. This could also be taken as evidence that there is scope for niche players to play an important role in the banking industry.

Using the results above we can determine to which extent the banking sector may improve its performance by exploiting the increasing returns of scale. The European banking sector as a whole hardly would improve efficiency by choosing the right scale of

Table 3-5: Weighted average of the estimated size inefficiencies in the European Union (percentages, number of banks in each country given in parentheses)

Country	1997	1996	1995	1994	1993
Austria (50)	0.02	0.03	0.04	0.05	0.04
Belgium (69)	0.02	0.02	0.02	0.03	0.03
Denmark (82)	0.14	0.15	0.16	0.17	0.15
Finland (7)	0.00	0.00	0.00	0.00	0.00
France (295)	0.01	0.02	0.02	0.02	0.02
Germany (886)	0.06	0.07	0.08	0.09	0.10
Greece (17)	0.03	0.01	0.03	0.12	0.12
Ireland (7)	0.00	0.00	0.00	0.00	0.00
Italy (194)	0.02	0.02	0.03	0.03	0.03
Luxembourg (97)	0.05	0.07	0.06	0.08	0.08
Netherlands (35)	0.01	0.01	0.01	0.01	0.02
Portugal (24)	0.00	0.00	0.00	0.01	0.00
Spain (125)	0.01	0.02	0.02	0.02	0.03
Sweden (12)	0.00	0.00	0.00	0.00	0.01
United Kingdom (74)	0.01	0.01	0.01	0.01	0.01
EU-15	0.03	0.04	0.04	0.05	0.05

Note: This table is derived using the results in Table 3-2. The weight of each bank is obtained from its total asset amount.

Table 3-6: Weighted average of the estimated size inefficiencies in the European Union, savings and cooperative banks (percentages, number of banks given in parentheses)

Country	1997	1996	1995	1994	1993
Austria (21)	1.2	1.2	1.2	1.5	1.5
Belgium (19)	5.4	4.8	1.3	4.7	5.1
Denmark (28)	3.5	3.5	3.8	4.3	4.1
France (86)	8.0	8.1	7.7	7.7	7.8
Germany (673)	6.4	6.7	6.6	6.5	6.5
Italy (129)	1.1	1.2	1.2	1.3	1.3
Luxembourg (5)	1.3	1.3	1.3	1.3	1.3
Spain (55)	3.2	3.2	3.2	3.2	1.1
Finland (1)	1.7	1.7	1.7	1.7	1.7
Greece (0)	–	–	–	–	–
Ireland (0)	–	–	–	–	–
Netherlands (2)	11.5	11.4	11.4	11.4	11.5
Portugal (3)	1.2	1.3	1.3	1.3	1.3
Sweden (0)	–	–	–	–	–
United Kingdom (3)	1.1	1.2	1.3	1.3	1.3
EU-15	6.1	6.2	5.9	5.9	5.8

Note: This table is derived using results in Table 3-4. The weight of each bank is obtained from its total asset amount.

operations as shown in Table 3-5. This is because small banks, although there are more than 800 credit institutions in Europe which are smaller than EUR 600 million measured in balance total, account for a small fraction of the European banking sector's assets (see Table A2-2). By contrast Table 3-6 shows that savings banks do have scope for improvement. By choosing the right scale, savings banks can reduce costs per asset by approximately 6 percent. This empirical finding is driven by France and Germany where cost reductions of approximately 8 percent and 6 percent are attainable.[35] Indeed, most of the European savings banks are based in these two countries and many of them are either small or very large.

Table 3-7: Weighted average of the estimated X-inefficiencies in the European Union, full sample (percentages, number of banks in each country given in parentheses)

Country	1997	1996	1995	1994	1993
Austria (50)	11	16	18	14	7
Belgium (69)	13	23	18	16	20
Denmark (82)	20	25	27	37	32
Finland (7)	10	17	11	28	32
France (295)	22	21	21	22	22
Germany (886)	16	19	14	14	10
Greece (17)	59	63	64	67	67
Ireland (7)	21	35	33	35	31
Italy (194)	14	18	26	22	24
Luxembourg (97)	22	20	19	11	20
Netherlands (35)	13	24	21	21	28
Portugal (24)	30	33	36	36	41
Spain (125)	22	24	25	23	29
Sweden (12)	28	30	23	35	39
United Kingdom (74)	-4	8	10	13	20
EU-15	16	20	19	19	20

Note: This table is derived using the results in Table 3-2. The weight of each bank is obtained from its total asset amount.

Other studies using European data (see, for instance, Altunbas and Molyneux, 1996) tend to find positive economies of scale also for larger size classes (in some cases up to a level of total assets of EUR 10 billion). Our results are more in line with evidence on

[35] Lang and Welzel (1996) also find moderate size economies for all size classes of German cooperatives using 1989-1992 data.

banking in the US in the 1980s and early 1990s which do not find large scale economies. McAllister and McManus (1993) argue that the traditional way of choosing input prices may bring about this 'economies of scale puzzle' since larger firms have better risk diversification opportunities and thus lower cost of funding than small firms. However, by the special nature of our dataset these so-called *financial scale economies* would have been revealed by our particular approach. If larger banks pay less than our constructed average price of funds, and thus have lower interest costs, then these banks will be judged as more efficient. Possibly the reason for our result that economies to scale are hardly present is the one brought up by Hughes and Mester (1998). They argue that large banks take more risk due to the financial scale economies mentioned above. As a consequence, the quality of the output mix of larger banks is of a different nature than the quality of the financial products of small credit institutions. Therefore, large banks may incur higher costs per unit of offered financial services.

X-efficiency

In Tables 3-7, 3-8 and 3-9 we have computed country averages of X-inefficiency in each year for the full sample, commercial and savings banks respectively. In constructing these averages we weight the X-inefficiencies of a particular bank by its total assets as a percentage of the total assets of the banks in the respective country. In the same way we also created averages for the European Union. In order to reduce the influence of severe outlying observations we ignore those banks with X-inefficiencies that are tremendously large or small. This can be revealed by means of a (two-sided) trimmed least squares regression of X-ineff$_{ti}$ on a constant and country dummies. We evaluate whether the absolute value of the standardised residuals from this regression exceed the cut-off value 5.[36]

The inefficiency that stems from the sources discussed above is modest. We shall see

[36] We scale the regression residuals by a robust estimate of the standard deviation. For this purpose the Medium Absolute Deviation (MAD) estimator is employed (see, among others, Rousseeuw and Leroy (1987), p.45).

Table 3-8: Weighted average of the estimated X-inefficiencies in the European Union, commercial banks (percentages, number of banks given in parentheses)

Country	1997	1996	1995	1994	1993
Austria (20)	19	25	16	3	-10
Belgium (33)	23	33	26	20	20
Denmark (47)	-2	2	3	8	4
Finland (5)	5	12	-2	20	20
France (171)	24	23	15	19	16
Germany (156)	17	21	14	10	1
Greece (17)	66	69	69	69	70
Ireland (5)	15	29	23	26	13
Italy (57)	14	15	20	18	18
Luxembourg (86)	26	25	19	8	14
Netherlands (28)	17	25	23	22	27
Portugal (18)	23	25	26	23	27
Spain (66)	20	15	10	8	10
Sweden (5)	1	0	-7	2	9
United Kingdom (59)	-16	4	3	8	15
EU-15	13	18	14	14	13

Note: This table is derived using results in Table 3-3. The weight of each bank is obtained from its total asset amount.

Table 3-9: Weighted average of the estimated X-inefficiencies in the European Union, savings and cooperative banks (percentages, number of banks given in parentheses)

Country	1997	1996	1995	1994	1993
Austria (21)	8	4	14	11	9
Belgium (19)	13	11	15	1	1
Denmark (28)	7	13	19	20	27
France (86)	13	4	10	11	11
Germany (673)	4	4	0	3	-3
Italy (129)	6	7	16	8	11
Luxembourg (5)	2	6	3	-4	13
Spain (55)	21	23	22	24	30
Finland (1)	23	23	38	42	53
Greece (0)	–	–	–	–	–
Ireland (0)	–	–	–	–	–
Netherlands (2)	-1	-1	-1	0	18
Portugal (3)	29	29	30	30	36
Sweden (0)	–	–	–	–	–
United Kingdom (3)	41	36	34	24	33
EU-15	9	6	7	7	7

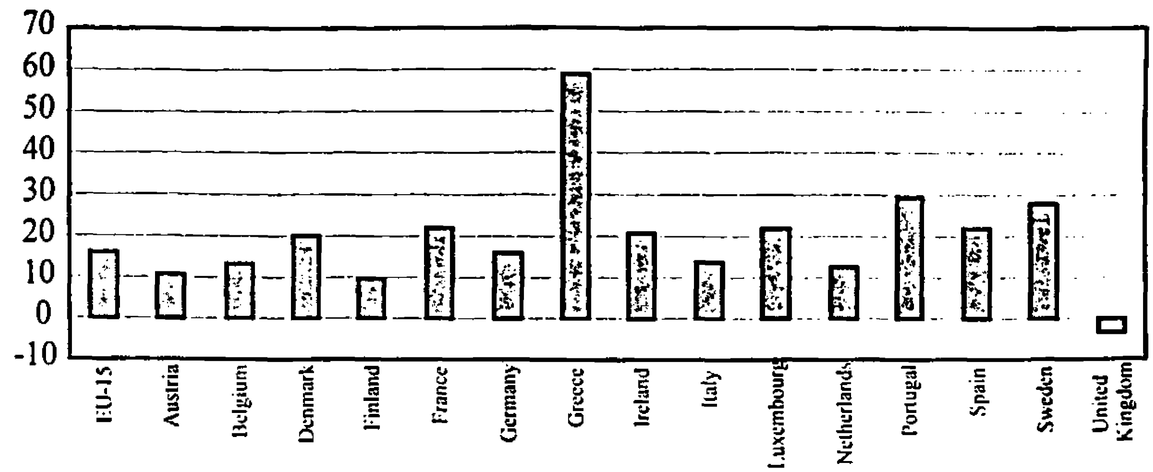
Note: This table is derived using results in Table 3-4. The weight of each bank is obtained from its total asset amount.

now that the largest cost reductions in the European banking industry can be achieved by improving management, i.e. by improving X-efficiency. In Table 3-7 we find that for the full sample of banks the average X-inefficiency in the sector is of the order 15-20 percent throughout the sample period. This figure is similar to what has been found for the US. Average X-inefficiencies within the European Union considerably fell from about 20 percent in 1996 to 16 percent in 1997. There remains, however, plenty of scope for improving the banking sector.

Who are Europe's efficient bankers? There are some striking differences in X-efficiency in Europe that are worth mentioning. These are also illustrated in Figure 3-11. In the UK, bankers were able to reduce their managerial inefficiency from approximately 20 percent in 1993 to full X-efficiency in 1997. On the other hand Greek banks appear to be the most inefficiently managed in Europe. Although Greek bankers improved, average X-inefficiency still exceeded 59 percent in 1997. Like the UK, the Netherlands and Finland show considerable gain in X-efficiency in the sample period. Conversely, Austria, France, Germany and Luxembourg did not improve over time or even worsened. The other differences we observe are less pronounced and sometimes do not match with the prior views that one may have. For example, Sweden is found to have a relatively inefficient banking sector with X-inefficiency ranging between 39 percent (1993) and 28 percent (1997). In Italy on the other hand, which many think is still at an early stage in restructuring, the banking sector is found to be relatively efficient (X-inefficiency fell from 24 percent in 1993 to 14 percent in 1997).

Although differences in X-efficiency across countries are substantial in many cases, we have to be somewhat careful in distinguishing between the performance of banks steering a middle course. The variance of the computed X-inefficiencies corresponding to the companies on the frontier is quite large. Therefore it may happen that average X-inefficiency even becomes negative for a particular country. Managerially efficient banks incur between 10 percent higher costs and 14 percent lower costs than the predicted

Figure 3-11: X-inefficiency of European banks in 1997 (percentages)



optimal costs at the 95 percent confidence interval.[37] The "thickness" of the cost frontier, that is the band around the cost function wherein the average cost of X-efficient firms fluctuate, is relatively small in comparison with the dispersion of the inefficient banks. These latter banks are highly inefficient with an average X-efficiency of 77 percent. In this case, the corresponding 95 percent confidence interval spans from 57 percent to 97 percent. The overlapping part of these two 95 percent confidence intervals indicate a "twilight zone" where banks are close to optimal performance but not fully cost efficient. The conclusion that can be drawn from these findings is that sometimes X-inefficiencies are extremely high but for other cases there are not enormous differences, all things considered.

Splitting up the sample into commercial banks and savings banks reveals some additional interesting results. Looking at the EU averages in Tables 8 and 9 it is clear that commercial banks have higher average X-inefficiencies (around 13 percent) than savings banks (around 7 percent), when each type is compared to its respective cost frontier.

[37]Our estimation method, RTFA, guarantees that X-efficient banks are not systematically located above or below the frontier.

Table 3-10: Weighted average X-inefficiency of small and large banks (percentages)

Year	Large	Small
1997	14	20
1996	18	25
1995	18	24
1994	18	24
1993	19	24

Note: A bank is defined to be 'large' when its total assets in 1997 exceeded ECU 10 billion. Small banks had a balance sheet total which was smaller than ECU 10 billion. In our sample there are 200 big banks and 1774 small ones.

Recall from section 6.3 however that X-efficient savings banks have on average roughly 4 per cent higher costs than X-efficient commercial banks. In other words, it is unlikely that savings institutions on average are much more efficient than their commercial peers when looking at the cost frontier associated with the whole European banking industry.

We also investigated whether there are differences in X-efficiency between small and large banks. Here we defined a bank to be 'large' when its 1997 total assets amount exceeded EUR 10 billion. The other banks were defined as 'small'. In our data set there are 200 big banks and 1774 small ones. Table 3-10 shows that, on average, large banks have around 6 percent lower X-inefficiency than small banks. To us this result seems somewhat counterintuitive as we would expect that smaller banks are easier to manage. Possibly large banks operate in a more competitive environment which forces them to be more efficient. Another reason can be that managers of large commercial banks are better monitored by shareholders. It is interesting to note that both small and large banks reduced their X-inefficiency over time.

3.6 Conclusion

In this paper the efficiency of the European banking sector has been assessed by estimating a cost frontier. Our augmented Cobb-Douglas cost model discriminates between banks of different legal status and size and also allows for shifts in average costs over time. We introduced a method to disentangle the effects of input prices on average costs from other time-related effects such as technological progress. Furthermore, an innovative regression technique, the Recursive Thick Frontier Approach, was used.

We find that the cost structure and performance of European banks over the period 1993-1997 can be characterised by the following key observations:

- Large cost reductions are possible when bank managers organise their businesses better. In 1997 more than 80 percent of the European banks were not located on the cost frontier and these banks can reduce costs by more than 16 percent. The slimming course of the European banking sector has already led to some cost cutting across Europe. X-inefficiency decreased on average about four percent between 1993 and 1997. Within Europe there are considerable differences in cost efficiency. Some countries showed rapid improvement in bank performance (UK, Netherlands, Finland) but others (Austria, France, Germany, and Luxembourg) have yet to step on the scales. A remarkable result is that bankers in the UK were able to reduce X-inefficiencies from over 20 percent to essentially zero in this fairly short time span.

- Different types of credit institutions operate with different average cost levels. We find that X-efficient commercial banks incur, on average, four per cent lower costs per unit of assets than X-efficient savings banks. Mortgage banks have considerably lower costs than savings and commercial banks. It is thus very likely that certain types of niche players may flourish while at the same time the bulk of the European financial institutions could go in the direction of commercial banking.

- The gains from positive economies of scale are "scanty" for the overall European banking industry. Increasing returns to scale do only exist for banks with total assets of less than EUR 600 million. The picture becomes slightly different when looking only at savings banks. By choosing a larger scale many savings banks can achieve cost reductions. Also the ten largest banks, i.e. those with total assets exceeding EUR 50 billion, have significantly higher costs than savings banks in the optimal size class. The total gain that the savings bank sector can make by exploiting scale economies is about six percent.

- For the EU banking sector as a whole time related developments such as technological progress seem not to have played an important role. Yet, relatively efficient savings institutions reduced average costs by about nine percent in 1993-1997.

From a cost reduction point of view one may expect that competitive forces will trigger further restructuring of the European banking sector. Although the European banking industry is at the beginning of a new era with the introduction of the Euro, one can only guess about the rapidity at which this restructuring will take place. Restrictive labour laws, state ownership, other forms of state interference (promoting national champions etc.) and cultural and language barriers to cross-border banking may set the pace of changes in European banking.

3.7 References

Aigner, D., Lovell, C.A.K., and Schmidt, P., 1977, 'Formulation and estimation of stochastic frontier production function models,' *Journal of Econometrics*, **6**, pp. 21-37.

Altunbas, Y., 1997, 'Estimating efficiencies in European banking: an analysis of alternative functional forms,' Research Papers in Banking and Finance, 97/20, Institute of European Finance, University of Wales, Bangor.

Altunbas, Y., Gardener, E.P.M., Molyneux, P. and Moore, B., 1999, 'Efficiency in European banking,' mimeo, University of Wales, Bangor.

Altunbas, Y., Goddard, J., Molyneux, P., 1999, 'Technical change in banking,' mimeo, University of Wales, Bangor.

Altunbas, Y. and Molyneux, P., 1996, 'Economies of scale and scope in European banking,' *Applied Financial Economics*, **6**, pp. 367-375.

Altunbas, Y. and Molyneux, P., 1997, 'Technical change in European banking,' Research Papers in Banking and Finance, 97/17, Institute of European Finance, University of Wales, Bangor.

Berger, A.N., 1993, ' "Distribution-free" estimates of efficiency in the U.S. banking industry and tests of the standard distributional assumptions,' *The Journal of Productivity Analysis*, **4**, pp. 261-292.

Berger, A.N., 1995, 'The profit-structure relationship in banking-Tests of market-power and efficient-structure hypotheses,' *Journal of Money, Credit, and Banking*, **27-2**, pp. 404-431.

Berger, A.N., Demsetz, R.S., and Strahan, P.E., 1999, 'The consolidation of the financial services industry: causes, consequences, and implications for the future,' *Journal of Banking and Finance*, **23**, pp. 135-194.

Berger, A.N. and Hannan, T.H., 1998, 'The efficiency cost of market power in the banking industry: a test of the "quiet life" and related hypotheses,' *The Review of Economics and Statistics*, **80**, pp. 454-465.

Berger, A.N., Hanweck, G.A., and Humphrey, D.B., 1987, 'Competitive viability in banking. Scale, scope and product mix economies,' *Journal of Monetary Economics,* **20**, pp. 501-520.

Berger, A.N. and Humphrey, D.B., 1992, 'Measurement and efficiency issues in commercial banking,' in *Measurement issues in the service sectors,* Z. Griliches (ed.), NBER, Chicago.

Berger, A.N. and Humphrey, D.B., 1997, 'Efficiency of financial institutions: International survey and directions for future research,' *European Journal of Operational Research,* **98**, pp. 175-212.

Berger, A.N., Hunter, W.C., and Timme, S.G., 1993, 'The efficiency of financial institutions: A review and preview of research past, present, and future,' *Journal of Banking and Finance,* **17**, pp. 221-249.

Berger, A.N., and Mester, L.J., 1997, 'Inside the black box: What explains differences in the efficiencies of financial institutions?,' *Journal of Banking and Finance,* **21**, pp. 895-947.

Charnes, A., Cooper, W.W., Lewin, A.Y. and Seiford, L.M., 1994, *Data Envelopment Analysis,* Kluwer, Dordrecht, The Netherlands.

Cooper, J.C.B., 1980, 'Economies of scale in the UK building society industry,' *Investment Analysis,* **55**, pp. 31-36.

DeYoung, R. and Hasan, I., 1998, 'The performance of de novo commercial banks: a profit efficiency approach,' *Journal of Banking and Finance,* **22**, pp. 565-587.

European Economic Community, 1977, *First Banking Directive,* First Council Directive, 77/780/EEC.

European Economic Community, 1988, *Second Banking Directive,* Second Council Directive.

Fanjul, O. and Maravall, F., 1985, *La eficiencia del sistema bancario espanol,* Alianza University, Madrid.

Gibson, H.D., and Tsakalotos, E., 1993, 'European Integration and the Banking Sector in Southern Europe: Competition, Efficiency and Structure,' *BNL Quarterly Review*, **186**, pp. 299-325.

Goldberg, L.G. and Rai, A., 1996, 'The structure-performance relationship for European banking,' *Journal of Banking and Finance,* **20**, pp. 745-771.

Hughes, J.P. and Mester, L.J., 1998, 'Bank capitalization and cost: evidence of scale economies in risk management and signaling,' *The Review of Economics and Statistics*, **80,** pp. 314-325.

Lang, G. and Welzel, 1996, 'Efficiency and technical progress in banking. Empirical results for a panel of German cooperative banks,' *Journal of Banking and Finance,* **20**, pp. 1003-1023.

Maudos, J., 1998, 'Market structure and performance in Spanish banking using a direct measure of efficiency,' *Applied Financial Economics,* **8**, pp. 191-200.

McAllister, P.H. and McManus, D.A., 1993, 'Resolving the scale efficiency puzzle in banking,' *Journal of Banking and Finance,* **17**, pp. 389-405.

Mester, 1996, 'A study of bank efficiency taking into account risk-preferences,' *Journal of Banking and Finance,* **20**, pp. 1025-1045.

Molyneux, P., Altunbas, Y., and Gardener, E. (1996). *Efficiency in European Banking*, Wiley, West Sussex, England.

Rogers, K.E., 1998, 'Nontraditional activities and the efficiency of US commercial banks,' *Journal of Banking and Finance,* **22**, pp. 467-482.

Rousseeuw, P.J. and Leroy, A.M., 1987, *Robust Regression and Outlier Detection*, Wiley, New York.

Sealey, C. and Lindley, J.T., 1977, 'Inputs, outputs and a theory of production and cost at depository financial institutions,' *Journal of Finance,* **32**, pp. 1251-1266.

Vander Vennet, R., 1996, 'The effect of mergers and acquisitions on the efficiency and profitability of EC credit institutions,' *Journal of Banking and Finance,* **20**, pp.

1531-1558.

Varian, H.R., 1992, *Microeconomic Analysis*, W.W. Norton & Company, Inc.

Wagenvoort, J.L.M. and Schure, P.H., 1999, 'The Recursive Thick Frontier Approach to estimating efficiency,' Report 99/02, European Investment Bank.

3.8 Appendix 1 Price Data

In the main text we have mentioned three input prices and a price index for banking services. In this appendix we give a detailed description of how we created these data.

The *price of funds* is obtained by taking a weighted average of the 3-month interbank offered rate and the deposit rate. The weight of the deposit rate is taken to be the value of deposits over the total amount of liabilities. The weight of the interbank rate is taken to be one minus the deposit rate weight. The deposit rates are extracted for each country from the IFS data set [Line 601 in the IFS data set of the IMF]. For some countries, we missed data on the last quarter in 1997 in which case we constructed the 1997 value to be the average of the rates in the first three quarters of 1997. The interbank rate was retrieved from Datastream International. We downloaded monthly data on the 3-month interbank offered rates in the EU-15 countries and from these we created year averages. The interbank rates which were thus created are given in Table A1-1.

The *price of labour* was constructed using BankScope data and OECD data. As wages in each country differ, we created 15 different wage rates. The 1996 and 1997 observations in each country are constructed using BankScope. We added up all labour expenses of all banks in the sample in a given country and divided the sum by the number of workers employed by these banks. A considerable number of banks report the necessary data. Only in case of Ireland we had few banks [in 1996, three banks (out of seven), and in 1997, four]. As employment data prior to 1996 is normally not found in BankScope we could not generate wage rates for 1993 - 1995 in this way. However, in the 1998 issue of 'Bank Profitability' of the OECD there is data for 1993 -1996 both on labour expenses in the banking sector, and on the total number of employees.[38] With this we had data for the entire period 1993 - 1997, with double data for the year 1996. Unfortunately, for a few countries the 1996 values of the BankScope calculations differed considerably from

[38]That is, for some countries the data is not available for the banking sector as a whole. In that case the OECD gives data for the commercial banks (Greece, Luxembourg, Portugal, Sweden and the UK) or for commercial plus savings banks (Denmark).

the OECD data, and we decided to base the price of labour in 1996 and 1997 on the BankScope data. Data for 1993, 1994 and 1995 was obtained by extrapolating the 1996 observation from BankScope using the OECD data to compute changes in the wage rate in the period 1993 - 1996. The results are given in Table A1-2.

The *price of buildings* is created by taking a price index for newly delivered buildings and correcting it for the relative price levels in each country. The data on newly delivered buildings is obtained from the CRONOS data set of Eurostat (/theme4/ construc/ isti08a/ i8aa ind, see Table A1-3), and relative price levels are constructed from data of CRONOS and the IFS of the IMF. We used Purchasing Power Parities (PPPs) with respect to Germany. In WEFA we found exchange rates to the German Mark for the EU-15 currencies.[39] In CRONOS we found monthly purchasing power parities from Jan 1993 until Nov 1995 (/theme2/ price/ ppa/ ppam). Dividing the exchange rates by the PPPs we obtained monthly data on the relative price level in Jan 1993 - Nov 1995. The observations from December 1995 onwards could be generated using IFS. In particular, we downloaded monthly data on the 'real effective exchange rate' (i.e. series *reu*) for Nov 1995 - Dec 1997 and used this index to extrapolate the previous data. From the monthly relative price levels in the EU-15 we created yearly averages [see Table A1-4]. The relative price levels in Table A1-4 were used to correct the price indices of buildings of Table A1-3 for price differences between the EU countries. The price index which resulted from this is the price of buildings as used in the study [see Table A1-5].

The price index for *banking services* is obtained from the CRONOS data set of Eurostat. We followed the link /theme2 /price /hicp /haind and selected time series hicp_idx, 125a ('banking services n.e.c.'). Unfortunately CRONOS only has data from 1995 onwards, if at all. To solve this problem, we took the 1993 and 1994 levels to be equal to the 1995 level. The results are given in Table A1-6.

[39] These are given by the series L00RF.M; e.g. for the UK (country code is M112) the series was 'M112L00RF.M'.

Table A1-1: Interbank rates in the EU-15 countries in 1993 - 1997, percentages

Country	Datastream Code	1993	1994	1995	1996	1997
Austria	ASVIB3M(IO)	7.05	5.14	4.57	3.38	3.50
Belgium	BIBOR3M(IO)	8.30	5.78	4.85	3.30	3.50
Denmark	CIBOR3M(IO)	10.83	6.32	6.13	4.02	3.73
Finland	FNIBC3M(IO)	7.81	5.37	5.70	3.63	3.24
France	PIBOR3M(IO)	8.76	5.88	6.52	3.95	3.48
Germany	FIBOR3M(IO)	7.36	5.40	4.53	3.32	3.33
Greece[a]	GDIBK3M(IO)	21.50	30.49	16.46	13.85	14.20
Ireland[b]	EIRED3M(IR)	9.56	5.94	6.25	5.41	6.06
Italy[c]	ITIBK3M(IO)	10.39	8.57	10.57	8.87	6.89
Luxembourg[d]	BIBOR3M(IO)	8.30	5.78	4.85	3.30	3.50
Netherlands	HOLIB3M(IO)	6.88	5.20	4.37	3.01	3.32
Portugal[a]	LISBO3M(IO)	13.42	11.27	9.90	7.39	5.71
Spain	ESMIB3M(IO)	11.88	8.09	9.33	7.58	5.46
Sweden	SIBOR3M(IO)	8.88	7.63	8.80	6.04	4.44
United Kingdom	LDNIB3M(IO)	6.04	5.67	6.80	6.18	6.95

Notes: [a] We missed some observations for Portugal and Greece. For Portugal we missed two observations: 31/12/92 and 31/1/93, so we averaged the other 1993 observations to come to the 1993 rate. Our first Greek observation is the 29/04/1994 one, so we missed little more than a year. For the last of these missing observations we found a good substitute, the 3 month deposit offered rate in Greece. We established our 1994 average as the average of these three observations and the 9 interbank rates we had for 1994. The 1993 average we took to be 21.5 percent, a rough guess based on extrapolating the three 3 month deposit rates we had for Greece.

[b] Datastream has no interbank rates for Ireland, so we took the 3-Month money middle rate [EIRED3M(IR)].

[c] Italy's first 4 monthly observations were taken from the series 'Italy Atic Interbank 3-month (history) - offered rate'.

[d] The Luxembourg data series is the same as the Belgium one [BIBOR3M(IO)].

Table A1-2: Annual wage per employee in the banking sector (ECU) countries, 1993 - 1997

Country	1993	1994	1995	1996	1997
Austria	46062	48002	52913	53827	52115
Belgium	54898	56704	58765	60161	60880
Denmark	43418	45000	47797	49986	51744
Finland	29898	32664	33069	37554	35054
France	51956	51591	52848	54946	57509
Germany	43606	44837	48627	49813	52923
Greece	19439	20788	22758	25555	27602
Ireland[a]	31696	31696	31696	34384	35350
Italy	55323	54966	50784	59409	58476
Luxembourg	54579	60460	64095	64284	63916
Netherlands	36635	40006	44601	47943	55438
Portugal	27731	27605	30322	32380	33905
Spain	39615	36296	37686	40203	37389
Sweden	39919	40048	46305	53385	55883
United Kingdom	41193	40452	40571	39950	42023

Note: [a] For the years 1993 and 1994 no data was available. We assumed the 1993 and 1994 wages to equal the 1995 wage.

Table A1-3: Cost of buildings, price indices for the EU-15 (Source: CRONOS, Eurostat)

Country	CRONOS Code	1993	1994	1995	1996	1997
Austria	Prin_out, b100[a]	95.2	97.7	100.0	101.7	103.3
Belgium	CPI[f]	96.3	98.6	100.0	102.0	103.7
Denmark	Prin_inp, b110[d]	94.5	96.4	100.0	103.1	106.0
Finland	Prin_inp, b100[c]	97.3	98.8	100.0	98.9	101.3
France	Prin_out, b110[b]	99.9	100.1	100.0	101.8	104.2
Germany	Prin_out, b110[b]	95.4	97.6	100.0	99.9	99.3
Greece	Prin_out, b110[b]	87.9	94.4	100.0	106.0	111.5
Ireland	Prin_inp, b900[e]	94.4	96.6	100.0	100.7	104.7
Italy	Prin_inp, b110[d]	94.7	98.1	100.0	101.8	104.3
Luxembourg	Prin_out, b110[b]	97.0	98.2	100.0	100.9	102.4
Netherlands	Prin_out, b110[b]	94.1	96.4	100.0	102.3	106.1
Portugal	CPI[f]	91.2	96.0	100.0	103.1	104.9
Spain	Prin_inp, b100[c]	92.3	95.4	100.0	102.8	104.7
Sweden	CPI[f]	95.4	97.5	100.0	100.5	101.0
United Kingdom	Prin_out, b110[b]	92.1	95.4	100.0	102.0	107.1

Notes:

[a] Output price index of the building sector (national currency)

[b] Output price index for residential buildings (national currency)

[c] Construction cost index of the building sector (national currency)

[d] Construction cost index of residential building (national currency)

[e] Construction cost index of building and civil engineering sector (national currency)

[f] Consumer price index

Table A1-4: Price levels in EU-15 relative to German prices, year averages

Country	1993	1994	1995	1996	1997
Austria	0.99	1.01	1.02	1.02	1.06
Belgium	0.90	0.91	0.91	0.93	0.96
Denmark	1.16	1.16	1.16	1.20	1.27
Finland[a]	1.00	1.00	1.00	0.94	0.96
France	0.92	0.91	0.90	0.91	0.94
Germany	1.00	1.00	1.00	1.00	1.00
Greece	0.69	0.69	0.69	0.75	0.83
Ireland	0.78	0.77	0.74	0.74	0.76
Italy	0.75	0.73	0.66	0.76	0.84
Luxembourg	0.84	0.85	0.86	0.87	0.90
Netherlands	0.91	0.92	0.92	0.91	0.92
Portugal	0.63	0.62	0.61	0.64	0.67
Spain	0.75	0.72	0.70	0.74	0.78
Sweden[a]	1.00	1.00	1.00	1.06	1.07
United Kingdom	0.74	0.74	0.68	0.73	0.95

Note: [a] For Sweden and Finland we missed data for Jan 1993 - Nov 1995. In these months the relative price levels have been taken to be 1.

Table A1-5: Price index buildings (Germany 1995 = 100)

Country	1993	1994	1995	1996	1997
Austria	94.5	99.1	101.5	103.5	109.7
Belgium	86.6	89.7	91.0	94.9	99.6
Denmark	109.7	111.7	116.3	123.3	135.2
Finland	97.3	98.8	99.9	93.2	97.1
France	91.4	91.4	89.7	92.5	97.5
Germany	95.4	97.6	100.0	99.9	99.3
Greece	60.7	65.2	68.5	79.1	92.7
Ireland	73.9	74.8	73.8	74.8	79.5
Italy	71.4	71.7	66.0	77.6	87.5
Luxembourg	81.8	83.7	85.5	88.2	92.4
Netherlands	86.0	88.4	91.8	93.5	97.8
Portugal	57.8	59.1	61.1	65.9	70.6
Spain	69.7	68.2	69.9	76.2	81.2
Sweden	95.4	97.5	100.1	106.2	108.3
United Kingdom	67.9	70.4	68.2	74.4	101.3

Table A1-6: Price index for banking services (Source: CRONOS, Eurostat)

Country	1993	1994	1995	1996	1997
Austria[b]	100	100	100	100	100.1
Belgium[a]	95.5	95.5	95.5	100	100.5
Denmark[a]	98.5	98.5	98.5	100	102.8
Finland[a]	102.5	102.5	102.5	100	101.1
France[b]	100	100	100	100	100.3
Germany[a]	97.5	97.5	97.5	100	102.4
Greece[c]	100	100	100	100	100
Ireland[a]	89.2	89.2	89.2	100	103.8
Italy[a]	93.6	93.6	93.6	100	109.9
Luxembourg[a]	94	94	94	100	112
Netherlands[a]	100.9	100.9	100.9	100	99.7
Portugal[a]	94.5	94.5	94.5	100	105.8
Spain[a]	97	97	97	100	118.7
Sweden[a]	98.1	98.1	98.1	100	114.4
United Kingdom[b]	100	100	100	100	101.2

Notes:

[a] Missing values in CRONOS for 1993 and 1994.

[b] Missing values in CRONOS for 1993, 1994 and 1995.

[c] Missing values in CRONOS for all years.

3.9 Appendix 2 Regression Results

Table A2-1: The number of EU banks on the cost frontier

Country	Full Sample (Ad. Table A2-3)	Commercial Banks (Ad. Table A2-6)	Savings Banks (Ad.Table 3-4)
Austria	8	0	8
Belgium	5	1	5
Denmark	13	3	5
Finland	0	0	0
France	59	23	36
Germany	173	26	51
Greece	0	0	–
Ireland	1	1	–
Italy	32	13	38
Luxembourg	25	17	3
Netherlands	3	5	1
Portugal	1	2	0
Spain	5	9	0
Sweden	2	2	–
United Kingdom	13	13	0

Table A2-2: Number of observations in each size class for the three regressions

Size Class	Full Sample	Commercial Banks	Savings Banks
Total Assets $\leq$ 100 million (ECU)	569	316	204
100 million $<$ Total Assets $\leq$ 300 million	2013	680	1259
300 million $<$ Total Assets $\leq$ 600 million	1551	557	919
600 million $<$ Total Assets $\leq$ 1 billion	1197	457	658
1 billion $<$ Total Assets $\leq$ 5 billion	2901	980	1649
5 billion $<$ Total Assets $\leq$ 10 billion	638	304	233
10 billion $<$ Total Assets $\leq$ 50 billion	703	348	158
50 billion $<$ Total Assets	298	223	45

3.10 Appendix 3 Variances of Regression Coefficients

In this Appendix we show how to obtain the variances of the parameters of interest given by vector w in the main text, after having estimated model 3.5.

For the parameter estimates of the type dummies, γ_j^*, exact standard errors can be computed since $\exp \pi_l$ is log-normally distributed if π_l is normally distributed.[40] For the estimates of the constant, γ_0^*, and the time dummies, δ_t^*, standard errors can only be approximated. A way to do this is given in Kmenta (1986, p.487). Let in the next formula α^* be a function of K other estimators $\beta_1^*, ..., \beta_K^*$, i.e. $\alpha^* = f(\beta_1^*, ..., \beta_K^*)$. Then, for large samples, the variance of α^* can be approximated using a Taylor expansion:

$$var(\alpha^*) \approx \sum_k \left(\frac{\partial f}{\partial \beta_k}\right)^2 var(\beta_K^*) + 2\sum_{j<k} \left(\frac{\partial f}{\partial \beta_j}\right)\left(\frac{\partial f}{\partial \beta_k}\right) cov(\beta_j^*, \beta_k^*), \; j,k = 1,...,K \quad (3.17)$$

We used formula 3.17 to approximate the standard errors of $\gamma_0^*, \delta_1^*, \delta_2^*, \delta_3^*$ and δ_4^*, where we assume that $cov(\eta_{0i}^*, \eta_{0j}^*) = 0$ if $i \neq j$ and for and $cov(\alpha_i^*, \eta_{0j}^*) = 0$ for $i = 1,2,3$ and $j = 1, ..., 4$.[41] Furthermore, we approximated $cov(\eta_{ki}^*, \eta_{kj}^*)$, $i \neq j$ and $k = 1, ..., 4$ by $cov(\eta_i^*, \eta_j^*)$, where $\eta_i^* = (\eta_{i1}^*, \eta_{i2}^*, \eta_{i3}^*, \eta_{i4}^*)$.

[40] Let $x \sim N(\mu, \sigma^2)$. The variance of a log-normally distributed random variable $y = \exp(x)$ equals $var(y) = \exp(2\mu + \sigma^2)(\exp \sigma^2 - 1)$.

[41] Note that:

$$cov(\alpha_1^*, \alpha_3^*) = -var(\alpha_1^*) - cov(\alpha_1^*, \alpha_2^*)$$

$$cov(\alpha_2^*, \alpha_3^*) = -var(\alpha_2^*) - cov(\alpha_1^*, \alpha_2^*)$$

$$var(\alpha_3^*) = var(\alpha_1^*) + var(\alpha_2^*) + 2cov(\alpha_1^*, \alpha_2^*)$$

$$cov(c, \alpha_3^*) = -cov(c, \alpha_1^*) - cov(c, \alpha_2^*), \text{ etc}$$