Industrial Organization of Banking: A Review

DORIS NEUBERGER

ABSTRACT Empirical research about structure, conduct and performance in banking markets has developed mostly independently from the microeconomic theory of banking. The present paper reviews the literature by focusing on the links between theoretical and empirical research. It considers basic conditions, variables of market structure, conduct and performance and public policy special to the banking industry. It is shown that the competitive conditions are different in different market segments, and that the trend towards universal banks which are active in different geographic markets gives new challenges to research.

Key words: Banking; Industrial organization; Structure-conduct-performance paradigm; Market imperfections; Public policy.


1. Introduction

In his seminal paper “What’s different about banks?” Eugene Fama (1985) explains the comparative advantage of banks vis-à-vis capital markets by the superior capability of banks to provide debt with inside information. Within the theory of asymmetric information, Diamond (1984) shows that a special role of banks is to minimize the agency costs between borrowers and lenders by monitoring the borrowers at low cost. As shown by Diamond and Dybvig (1983), another special function of banks is to transform illiquid assets into liquid liabilities, providing insurance against liquidity risk with private information to agents. Hence, banks arise because of incomplete and asymmetric information in the financial markets.

Whereas since the 1980s the microeconomic theory of financial intermediation has developed within the theory of incomplete markets and the ‘new institutional economics’ (for surveys see Swank, 1996; Thakor, 1995; Bhattacharya and Thakor, 1993; Neuberger, 1994), many empirical studies on market structure, conduct and performance in the banking sector have been conducted from the beginning of the 1960s (for surveys see Berger, 1995; Gilbert, 1984). However, the
hypotheses tested there are mostly derived from the theory of oligopolistic goods markets, and it is rarely asked whether they can be applied to the banking markets without restrictions. Attempts to base the empirical research on IO topics in banking on the micro-economic theory of the bank are new (e.g. Hannan, 1991b; Calem and Carlino, 1991). They follow the new approach of research in industrial organization where empirical studies on single industries are closely linked to theory (Bresnahan and Schmalensee, 1987; Martin, 1993: 9; Schwalbach, 1994). In the last decade, IO-related research in banking has tremendously increased, being driven by new questions with regard to competition and strategic behavior after important deregulations have taken place in the US banking industry (Rhoaides, 1997).

The present paper reviews the industrial organization research in (commercial) banking within the revised structure–conduct-performance paradigm. It considers basic conditions, variables of market structure, conduct and performance and public policy which are special to the banking industry. We focus on the links between theoretical and empirical research, rather than attempting to review the whole literature. The aim is to ease the interpretation and critique of empirical findings, finding out how robust the results are to alternative assumptions about the type of competition or structural variables affected by deregulations and technological progress. The paper is organized as follows. In Section 2, the revised structure–conduct-performance (SCP) framework is specified for banking markets. Section 3 reviews the basic supply conditions, Section 4, the basic demand conditions, Section 5, market structure (5.1, market delineation, 5.2, cost structures and entry barriers), Section 6, conduct, Section 7, performance and Section 8, public policy. A conclusion is given in Section 9.

2. Revised SCP-Framework for Banking Markets

The theory of incomplete information and principal-agent problems as an important branch of the new industrial economics has special relevance to banking markets, being the basis for the microeconomic theory of banking (Swank, 1996; Thakor, 1995; Neuberger, 1994). Since banks are special players in the market for information, the conditions of this market are of importance to their structure, conduct and performance. Therefore, the usual SCP-framework has to be completed by aspects of incomplete information to be useful for the analysis of banking markets. This is done in Figure 1.

We consider the revised SCP-paradigm, where all variables are endogenous because of interdependencies between variables of market structure, conduct and performance and feed-back effects on basic conditions and public policy (see Scherer and Ross, 1990; Schwalbach, 1994). To use this paradigm for an analysis of banking markets, we integrate market imperfections (uncertainty, asymmetric information and transaction costs) into the basic conditions. Especially, asymmetric information between borrowers and lenders and the costs of gathering information have an impact on the activity of banks, their structure and performance. Important variables of the basic conditions are risks, attitudes towards risks and principle–agent-relationships. They have special effects on market structure (e.g. diversification), conduct (e.g. information gathering, risk taking) and performance (e.g. allocation of risks and information). The special public policy towards banks, i.e. protective and prudential regulations and competition policy, can be explained as reactions to market failures in the banking sector.
Figure 1. Revised SCP-framework for banking markets.
3. Basic Supply Conditions

On the supply side, outputs and inputs of banks differ from those of industrial firms. Whereas the latter produce goods mainly by employing labor and physical capital, banks produce services primarily with labor and financial capital (notably debt capital). A service is defined as a physical or an intellectual activity, which satisfies a need at the moment of provision without the indirect way of a physical good (Kulhavy, 1974: 456). A service cannot be isolated from its producer, cannot be held in stock, is non-material and is “co-produced” by provider and user (De Bandt, 1996: 20). As a consequence, service markets are characterized by the following features:

(1) Image and trustworthiness of the suppliers are important for demand.
(2) The distribution of services is important for supply: since the provision of a service often requires direct contact to the consumer, a service company has to be located where it is desired by its clients.
(3) As the production of services cannot be standardized, services are not patentable.

The services of a bank comprise functions of providing liquidity, information and transformation (of risks, terms and size), which is done by simultaneously taking deposits and granting loans.\(^3\) Whereas deposits *per se* are an input (debt capital) to banks, deposits with attributes such as ‘liquid’ or ‘safe’ may be conceived as outputs. This explains why there are different approaches to the classification of inputs and outputs in banking studies. The “asset approach” or “intermediation approach” holds that banks produce loans and other investments from deposits, other funding sources, labor and materials (Wheelock and Wilson, 1995: 41). Alternatively, the “production approach” views deposits and loans as outputs and only the non-interest expense as inputs (e.g. Gilligan *et al.*, 1984). Different empirical approaches try to find out which banking services should be considered as outputs and which as inputs. The value-added approach classifies activities for which banks create high added value (loans, demand deposits, time and savings deposits) as outputs and labor, physical capital and purchased funds as inputs (e.g. Berger and Humphrey, 1992). The user–cost approach, on the other hand, classifies a bank asset as an output if its financial return exceeds its opportunity cost and a bank liability as an output if its financial cost falls short of its opportunity cost. Empirically, both approaches lead to similar results, with the main exception of demand deposits which are classified as an output in most user–cost studies and as both an input and an output in value-added studies [Wheelock and Wilson (1995: 42) for more detail see Berger and Humphrey (1992)].

In the supply of banking services, equity capital plays a special role. Whereas industrial firms need equity primarily to finance investments in assets, the key role of equity in banking is to protect the depositors from the risk of a bank failure. Since safeness of deposits is an important service to the banks’ customers, banks have a special interest in reducing bankruptcy risk. Moreover, the protection of depositors may be in the interest of the economy as a whole because of negative externalities of production in banking. The task of banks to provide insurance against liquidity risks by supplying demand deposits may cause a bank run, where all depositors withdraw their funds in a panic and the bank collapses (see Diamond and Dybvig, 1983; Calomiris and Gorton, 1991). If this run spills over to other
banks, it is likely to cause strong negative effects on the real sector. This possible market failure due to negative externalities is the reason for a special public policy towards banks which aims at improving efficiency and protecting small investors and depositors.

Supply conditions in banking are dependent on principal–agent relationships and the ways to minimize agency costs. The task of banks to gather information makes long-term relationships to customers efficient. The longer the bank–customer relationship, the more information and experience is gained, which reduces agency problems (Haubrich, 1989; Diamond, 1991).

### 4. Basic Demand Conditions

In thinking about the basic demand conditions, it is important to distinguish between retail and wholesale banking industries. Wholesale banking is defined as the provision of financial services to large corporate customers, whereas the customers of retail banking are households and small to medium-sized firms. In the wholesale segment, there is little or no scope of banks to exert market power, because large corporations usually can choose between different suppliers (including foreign banks, investment banks and the commercial paper market).

In the demand for retail banking services, demand substitutability should be lower because of asymmetric information and transaction costs. Often, retail banking customers are badly informed about the services, which are “experience goods” rather than “search goods”. Gathering information may be especially costly and at the same time its expected return is likely to be especially low. The search costs are high, if only few relevant information is publicly available or if the comparison of different offers is very complex. The expected return of search is low, if the acquisition of the banking service does not cost much, obtaining a small weight in the consumer’s budget. Many financial services have characteristics which restrict the search for information therefore, as represented in Table 1 (see Sondhof, 1989: 105). For example, a bank account is typically characterized by product characteristics 3, 4 and 5 which explains why customers mostly do not possess enough information to choose the most favorable account.

This informational problem reduces the market power on the demand side of a retail banking market. It is severest for private customers who invest little in search for information. Their demand is likely to be characterized by “bank loyalty”, i.e. the tendency to maintain a banking relationship, after having chosen a bank. They

<table>
<thead>
<tr>
<th>Product Characteristic</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Availability of product information</td>
<td>Low</td>
</tr>
<tr>
<td>2. Complexity</td>
<td>High</td>
</tr>
<tr>
<td>3. Price structure</td>
<td>Not transparent</td>
</tr>
<tr>
<td>4. Frequency of purchase</td>
<td>Seldom</td>
</tr>
<tr>
<td>5. Budget burden</td>
<td>Low</td>
</tr>
</tbody>
</table>
will only slowly react to an interest differential, once their house bank has gained reputation and has reduced the informational asymmetry in a long-term relationship. The switching costs on the demand side should therefore be especially high in those market segments. Empirical studies show that switching costs are important in deposit markets and negatively affect bank retail deposit interest rates (Sharpe, 1997). However, technological progress and an increasing willingness of households to search for information by using telecommunication technologies has led to the foundation of direct banks, which offer their products by telecommunication channels. The customers of direct banks are less loyal, i.e. their switching costs are lower, because they have to incur less costs to get informed and to buy banking products (Neuberger, 1997b).

Market power on the demand side may also be low because of low substitutability of bank products due to product differentiation. The less customers search for information, the higher is the incentive of banks to differentiate their products by creating a brand or an image. In banking markets, an image is usually not directed to the products, but to their suppliers who seek to create consumer preferences in this way (Sondhof, 1989: 108).

The demand for retail banking services may be further characterized by network externalities of branch and ATM systems. The quality of a banking service perceived by a consumer is the higher, the larger the branch network, i.e. the more consumers demand this service from the same bank. Then, demand depends on the anticipated total number of the bank’s customers. This interdependency between utilities may cause coordination problems which imply a suboptimal allocation of resources (Vives, 1991: 20).

5. Market Structure

5.1. Market Delineation

To analyze market structure, it is first necessary to define the relevant market. Banks are typically multiproduct firms, being active on the deposit market, the loan market and the securities market. The products offered on each market are not homogeneous, due to different preferences on the demand side and to different activities and factor inputs on the supply side. As a consequence, the loan market and the deposit market are structurally segmented, with respect to either customer groups or bank products. A segmentation according to customer groups implies the submarkets of wholesale banking (big corporate customers) and retail banking (small and medium-sized firms and households). The consumer preferences differ with respect to the value of search for information, the degree of risk aversion and the need for liquidity. A segmentation according to bank products leads to loan-submarkets (mortgage loans, corporate loans, consumer loans) and to deposit-submarkets (demand, time and savings deposits). Here, those products define a market which satisfy the same needs (e.g. liquidity, safety, profitability) and are strongly related by substitutability. They are largely homogeneous with respect to their physical attributes (e.g. deposits books, credit cards), but differentiated by qualities (e.g. branch network, automatic teller machines, telebanking) or images.

Having defined a product market, we have to determine the geographic extent of the market. For information-intensive retail banking products, the relevant market area is not only limited by transportation costs, but also by costs of
information. Therefore, the traditional approach to market definition in retail banking supposes that financial services are offered mainly by local commercial banks. It assumes that the costs of information and transportation would be prohibitive if customers searched for and used services of distant banks or if banks evaluated a non-local business seeking credit (Elliehausen and Wolken, 1990: 801).

A variety of empirical approaches to delineate geographic banking markets have been suggested in the literature (for surveys see Wolken, 1984; Whitehead, 1980). They can be grouped into three basic categories. The first, 'supply-oriented' approach, is concerned with the geographic distribution of a bank's customers; i.e. the bank's primary service area. The second, demand-oriented approach, attempts to identify the location and characteristic of users of some type of banking services, establishing a geographic market for these services. It emphasizes the reasons why a customer chooses a specific banking office to determine which groups of customers are limited to local banks. The third approach comprises studies which use some demand and/or supply variables to generate on a theoretical base a generalized method for identifying geographic banking markets. Some studies focus on the price equalization characteristic of markets by analyzing price variances or on price changes by analyzing the correlations of prices over time in different regions (e.g. Keeley and Zimmerman, 1985; Jackson III, 1992). Other studies define markets according to overall economic integration, relying on secondary economic and demographic data.

Until the 1980s, most of the US studies indicated that banking markets are local as opposed to regional or national. This local market approach has been adopted for antitrust purposes (Wolken, 1984: 25). Local banking markets are most often approximated by the geographic areas of cities, counties and standard metropolitan statistical areas (SMSA) (Whitehead, 1980: 26). This approach has lost relevance in those banking markets where the geographic extent of the market has broadened due to advances in telecommunication and deregulations. Some studies show that geographic limitations have weakened and that some financial services are provided in regional or national markets (Wolken, 1984: 11; Keeley and Zimmerman, 1985; Dunham, 1986). The local market approach is, however, still valid for the markets of financial services to small and medium-sized customers which demand financial services primarily from local commercial banks (Elliehausen and Wolken, 1990).

In the US, most structure–performance studies focus on retail banking markets, using concentration in local market areas as a measure of market structure (Wolken, 1984: 29). Because comparable sub-market and microdata are typically not available in European countries, structure–performance studies in Europe consider the European banking market as a whole, regarding each individual country as a local market (e.g. Goldberg and Rai, 1996; Molyneux and Forbes, 1995: 158; Molyneux et al., 1994; Ruthenberg, 1994; Short, 1979). The drawback of this approach is that it does not take into account that different banking services are provided by different banks in different geographic markets. The universal banking systems of Europe typically comprise not only large banks with nationwide business, but also small banks of only local importance. As far as we know, this has been taken into account only by Mooslechner and Schnitzer (1994) in a structure–performance study on Austrian banks.
5.2. Cost Structures and Entry Barriers

The number and the size of banks in a market depend on the public policy towards banks, economies of scale and scope in the production of banking services and entry barriers (Molyneux et al., 1994: 447). From the view of public policy, a low number of competing banks may be preferable because it is easier to monitor and, by reducing the intensity of competition, it lowers the risk of bank failures. Increasing returns to scale may be brought about by transaction costs (fixed costs of investments in capital, branches, ATM systems etc.), but also by uncertainty. Models of bank behavior under uncertainty explain economies of scale in reserve holding (Baltensperger, 1972a, 1972b, 1980) and economies of scale from diversification (Diamond, 1984; Neuberger, 1995). Economies of scope are relevant to banks, if the joint production of different financial services by the same fixed factors reduces costs and if the provision of different services to a single customer reduces costs of gathering information (e.g. Vale, 1991).

Many studies on US banks suggest that the average cost curve has a flat U-shape with meaningful economies of scale typically applying only to relatively small banks (for surveys see Berger et al., 1993; Tichy, 1990; Humphrey, 1990, 1985; Clark, 1988; Kolari and Zardkoohi, 1987). Diseconomies of scale have been mostly found for middle-sized banks. For large banks there are too few studies to provide clear-cut evidence (Tichy, 1990), but economies of scale due to diversification are relevant (McAllister and McManus, 1993). Some studies have found average cost dispersion to be higher in the same size class of firms than across different sizes (e.g. Gilligan et al., 1984). The studies disagree about the minimum efficient scale of production, which appears to be sample specific (Clark, 1996: 343). It has been shown that the degree of scale economies increased in the 1980s, which may be attributable to the regulatory and technological changes in the US banking industry (Gropper, 1991).

On the existence or extent of scope economies, there is little consensus. There is some evidence for economies from joint production among specific product pairs (especially deposits and loans), but little support for global economies of scope. The inconsistent results are attributed to various conceptual and empirical problems with measuring scope economies (Berger et al., 1993; Clark and Speaker, 1994; Tichy, 1990: 365). For example, scope economies are computed by comparing the predicted costs of producing a given bundle of outputs by two or more specializing firms to the joint production by a single firm. In banking, however, virtually all firms produce the whole set of products specified in the cost function which makes data about specializing firms too rare. More clear-cut results may be found, if scale and scope economies are measured from the profit function rather than from the cost function (Berger et al., 1993: 225).

For some European countries and international samples, similar results about economies of scale and scope have been found (for a survey see Lang and Welzel, 1995b, 1996; Tichy, 1990: 370). However, the European studies suffer from severe limitations in the data employed, implying often implausible or non-representative results. For Germany, plausible support for economies of scale and scope has been found for the relatively small, cooperative banks (Lang and Welzel, 1996).

Barriers to entry specific to banking result from special regulations (e.g. minimum capital requirements), but also from economies of scale, product differentiation advantages and absolute cost advantages. The latter two may arise from
different investments to build up a clientele and a reputation for solvency (e.g. investments in branches, ATM systems, advertising). If banking services have to be offered close to the customers, an extensive branch network is a crucial factor of competitiveness. Location models show that, by crowding space, an extensive branch network is an entry barrier in the retail banking market. Incumbent banks have an incentive for branch proliferation to such an extent that entry with an additional network would become unprofitable (Vives, 1991: 20; Schmalensee, 1978). Branch proliferation is increased by network externalities of branch and ATM systems. The quality of banking services perceived by a consumer is higher, the higher the expected total number of the bank’s customers. Therefore, the bank has an incentive to create growth expectations by attracting a larger number of customers. At the same time, it may deter entry of competitors by making its ATM system incompatible, denying access to a well-established network (Vives, 1991: 20). The incentive for entry-deterrence by branch proliferation is reduced by direct banking, where the geographical proximity to the customer plays no role any more. But also direct banks need a network for cash withdrawals and deposits. The costs of investing in ATM networks are sunk.

A further barrier to entry results from another quality variable, a bank’s reputation for solvency or its probability of failure. Here again, the expected customer base plays an important role. The larger it is, the higher is the bank’s stability, because of switching costs and diversification. Since risk-averse depositors prefer safe banks to riskier ones ceteris paribus, low-risk banks should enjoy larger margins and market shares. This kind of vertical differentiation may explain a natural oligopoly structure of banking markets. Riskier banks are not viable, if the initial advantage of safer banks is large or if the depositors do not differ much in terms of risk-aversion (Vives, 1991: 21; see Shaked and Sutton, 1983).

These aspects of entry barriers apply only to retail banking markets based on a branch network. Wholesale banking markets, where switching costs are low and customers have access to alternative financial institutions or securities markets, have properties of contestability. In the lending business, this applies to large corporations with access to the capital markets, whereas the markets for small and medium-sized firms in need of bank loans are not likely to be contestable. Here, an entry may be impeded by sunk costs of information gathering: the accumulated knowledge is borrower-specific and private information. An incumbent bank obtains an informational advantage by knowing the characteristics of applicants first (Sharpe, 1990). In the case of price competition, uninformed entrants face the winner’s curse problem of attracting the loan applicants, which have been evaluated as bad risks by the incumbents (Broecker, 1990). The sunk costs and the informational advantage of incumbents preclude that the commercial loan market is contestable (Neuberger, 1994: 48).

Non-contestability of banking markets may not only be caused by natural, but also by regulatory barriers to entry and/or exit. Market entry of banks is controlled by the imposition of licensing conditions, and the exit of banks may be impeded by acquisition or bailout. Unlike the bankruptcy of an industrial firm, the bankruptcy of a bank has negative social effects due to externalities, which have to be prevented. However, this need not be done by bailouts, if prudential regulations succeed in preventing a crisis (see Section 8).

The evidence on contestability of banking markets is mixed. An analysis of entry into US banking markets found evidence for non-contestability in rural markets, but not in urban markets that are considered attractive for entry (Amel
and Liang, 1997). More indirectly, contestability has been tested by applying the Ross–Panzar test and the mark-up test to banking markets (for a survey see Shaffer, 1994). Support for contestability has been found for US and Canadian banks (Shaffer, 1993; Shaffer, 1982; Nathan and Neave, 1989), but not for European banks (Molyneux et al., 1994; Lang, 1997).

6. Conduct

In the banking literature, there is no consensus about the instrument of competition. Some models assume Bertrand competition (Yannelle, 1988, 1989), others Cournot competition (Calem and Carlino, 1991; Berg and Kim, 1994; Neven and Röller, 1997). Both should be appropriate for homogeneous wholesale banking products, but not for retail banking services. A more realistic formalization of competition in retail banking should allow for a demand for banking services that is not infinitely price-elastic, because product differentiation (quality competition) is important. This is done in models of product differentiation with price competition (Hannan, 1991b; Chiappori et al., 1995; Bouckaert and Degryse, 1995).

Bertrand competition may not apply for several reasons. A banking market with a high asymmetry of information between suppliers and demanders is non-transparent, which implies that the price cannot be an “objective” means for buyers to take their decisions. Competition refers to the quality of services or, since the quality cannot easily be observed or evaluated, to the image created by the supplier (De Bandt, 1996: 21). According to the theory of incomplete credit markets, prices do not adjust to eliminate situations of excess demand (Stiglitz and Weiss, 1981) and should be highly dispersed because price competition hardly works (for empirical evidence see Gischer, 1995; Winker, 1994). On the other hand, it may be the case that pricing of interest bearing products are so inextricably linked to money market rates, and these are strongly correlated with cyclical factors that margins on interest bearing business are very similar for banks, thus limiting competition. This will encourage banks to compete on quality related to interest business and on price for fee-based services.7

Quality variables of bank competition are the density of the branch network, the density of automated teller machines, the reputation for solvency, the quality of the staff or of the premises (Neven, 1990: 164).8 Quality, however, is also subject to asymmetric information. Because banking services are primarily experience goods with a low productivity of search for information by consumers, advertising becomes an important instrument for banks to create a brand or an image.9

In the lending business, gathering information by evaluating credits and monitoring borrowers is an important variable of conduct. Credit-worthiness tests have special implications for banking competition. If the production of information by a bank can be costlessly observed by rival banks, a free-rider-problem with underinvestment in information production is to be expected (Hellwig, 1991: 46). If, on the other hand, the results of information gathering remain private information, a winner’s curse-problem arises. Each bank is afraid of attracting those consumers which have been evaluated as bad risks and have been rejected by rival banks (see Broecker, 1990).

Switching costs and the multiproduct nature of most banks are sources of price discrimination. Since switching costs are likely to decrease with wealth, banks may discriminate across customers, offering higher returns to wealthy customers.
Moreover, banks discriminate across products, i.e. they charge prices below costs for some services and prices in excess of costs for other services. This cross-subsidization is, however, difficult to estimate, because product-specific costs cannot be easily calculated due to the joint production of many banking services (Neven, 1990: 169).

It has been suggested that banking collusion may be fostered by multimarket contact. Since a non-cooperative behavior in one market can be punished in other markets, competition in several markets leads to “mutual forbearance”. However, the evidence for this hypothesis is ambiguous (Gilbert, 1984: 629). Theoretical and empirical work by Mester (1987, 1992) suggests that to the contrary, multiple market contact between banks leads to more competition than indicated by the structure of the individual markets.

Banks are claimed to be especially prone to expense preference behavior, which means that managers increase “staff expenditures, managerial emoluments, and discretionary profit”, for which they have a positive preference, beyond the profit-maximizing point (Edwards, 1977: 147). This may be due to regulations which shield banks from competition or to agency problems which are especially severe because of a high dispersion of ownership and a coordination failure of small, uninformed depositors (Dewatripont and Tirole, 1993). Empirical studies provide support for the expense preference hypothesis in banking (Edwards, 1977; Smirlock and Marshall, 1983; Mester, 1991a, 1993; for a survey see Mester, 1991b). Banks may also translate their market power into risk avoidance by way of a low risk loan portfolio [for a theoretical foundation see Sealey (1980)]. A weak relationship between concentration and profitability in banking has been explained by an increase in risk-avoidance with market power (Edwards and Heggestad, 1973; Heggestad and Mingo, 1976; Heggestad, 1977; Rhoades and Rutz, 1982; Bourke, 1989).

7. Performance

The performance of firms can be measured in terms of their productive (cost and profit) efficiency\(^{10}\) and allocative efficiency (market power). To measure efficiency, inputs and outputs have to be related to each other. Here again, researchers of banking markets face the problems of how to define and measure inputs and outputs. This explains why no technique for measuring efficiency has been generally accepted and why there are considerable differences in measured efficiency (Wheelock and Wilson, 1995; Colwell and Davis, 1992). As stated above (Section 3), it is not clear which financial services constitute inputs and which ones outputs to a banking firm. Moreover, there are different approaches to measuring inputs and outputs. In the 1970s and early 1980s it was common to measure output by the number of transactions or accounts serviced, whereas today output is mainly measured in terms of the dollar volume of loans or deposits, which includes interest (Wheelock and Wilson, 1995: 41). The measurement problems arise not only from the multiproduct nature of most banks. A problem fundamental to any service industry is that output consists to a large degree in quality, which is not taken into account if productivity is measured by volume or price alone (De Bandt, 1996: 29). An additional quality dimension especially relevant to banking is risk.

Nevertheless, in the US, the banking industry has been considered as a “convenient laboratory” for testing structure–performance hypotheses (Rhoades, 1982a: 477). In contrast to the manufacturing sector, “... banking presents us with
what is essentially a local market industry with relatively well defined geographic markets and with firms producing roughly the same services. In banking, most problems with diversification can be substantially reduced by confining the analysis to unit banks, i.e., banks that operate at a single location" (Rhoades, 1982b: 376). This explains the large number of structure-performance studies on US banking markets, which are approximated with the rural, county, or SMSA definitions (Wolken, 1984: 29).

Unlike inter-industry studies, these banking studies have not consistently found a positive relationship between concentration and profitability (see the reviews in Gilbert, 1984; Shaffer, 1994; Smirlock, 1985; Goldberg and Rai, 1996). When it was found, it has been interpreted in different ways. The relative market power hypothesis (RMPH) asserts that in concentrated markets, banks are more able to collude and extract monopolistic rents by offering lower deposit rates and charging higher loan rates. In contrast, the efficient-structure hypothesis (ESH) explains a positive relationship between profits and concentration by lower costs achieved through superior management or production technologies. According to this view, more efficient firms gain larger market shares, resulting in higher levels of concentration. Since larger market shares may reflect higher efficiency and/or higher market power, the two hypotheses cannot be distinguished by including market share as an independent variable (Shepherd, 1986). For US banks, a positive relationship between market share and profitability generally exists, but has been interpreted in support of both hypotheses (Smirlock, 1985; Rhoades, 1985; Evanoff and Fortier, 1988; Frame and Kamerschen, 1997).

To distinguish between the RMPH and ESH, some studies control directly for efficiency variables (Berger, 1995; Frame and Kamerschen, 1997), while others use price data as dependent variable (Berger and Hannan, 1989; Calem and Carlino, 1991; Hannan, 1991a). They indicate that the empirical support for the RMPH hinges on the existence of regulatory entry barriers and/or natural entry barriers linked to certain product lines. Banks which are shielded from competition by severe intrastate branching restrictions have market power in accordance with the RMPH, but not those in unit banking states or with statewide branching (Berger, 1995; Frame and Kamerschen, 1997). Price-concentration studies for local retail deposit markets show that high market concentration is correlated with low deposit rates according to the RMPH (Berger and Hannan, 1989; Calem and Carlino, 1991), but that even in unconcentrated markets, deposit rates reflect monopoly power (Calem and Carlino, 1991). This may be due to product specific transaction or search costs. Likewise, commercial loan rates have been found to depend significantly on market concentration and local market effects (Hannan, 1991a).

Unlike US banking, universal banking in Europe is no convenient laboratory for testing structure-performance hypotheses: geographical market delineation is less clear, and the availability of the necessary bank-by-bank database is poor. Therefore, the structure-performance studies on European banking markets are based on approximations which contrast sharply to those of their US counterparts. Mostly, they neglect local markets and individual product lines, approximating geographic markets by countries and product markets by all banking assets or total deposits. Using international databases, they conduct cross-section and pooled time-series estimates for different pools of European countries. In contrast to the US studies, they mostly provide support for the RMPH, finding a positive effect of concentration, but not of market share on profitability (Molyneux and Forbes,
1995; Ruthenberg, 1994; Molyneux and Teppet, 1993; Schuster, 1984; see Goldberg and Rai, 1996). However, those studies which incorporate efficiency directly in the model obtain conflicting results. Whereas Vander Vennet (1993) finds evidence for collusion in some European countries, Goldberg and Rai (1996) do not find support for the RMPH, but for the ESH for the largest banks in 11 European countries.

For individual European countries, there is little evidence. Lloyd-Williams et al. (1994) find support for the RMPH for Spanish banks with pooled data for the national market. The first European analysis based on micro-databases with banks of different size and different geographical expansion has been conducted by Mooslechner and Schnitzer (1994) for Austrian banks. They find that the results depend crucially on the set of control variables included that control for differences between banks in balance sheet composition and costs and differences in the “real” market structure of different geographical markets (demand factors). They doubt that the US results of studies which do not incorporate such control variables are stable and that the standard US approach of SCP-studies can be applied to European banking. Obviously, much more research needs to be done to find out the relevance of the RMPH and ESH for different geographical markets, different product segments and different banking sizes.

More insight into the effects of different structural variables is gained by looking at the theoretical foundation of the equations estimated. In a banking model of price competition under product differentiation, Hannan (1991b) shows that the return on assets depends on market structure variables of a potentially large number of markets according to the shares of the relevant businesses in the bank portfolio, on the capital–asset ratio and on the ratio of fixed costs to assets. Deposit market concentration should have a greater impact on the return on assets, since it, unlike loan market concentration, negatively affects total assets. For a multiproduct banking firm, profitability or price effects of concentration in single markets may be blurred by cross-subsidization. This explains why empirical SCP-studies, which differ with respect to the variables included, lead to ambiguous results and may be mis-specified. Mostly, as a measure of market structure, only a single concentration ratio of a local deposit market is used, whereas performance is measured by variables of overall profitability or interest rates on single markets (see Gilbert, 1984). This approach may have been appropriate for the US unit banks in the periods of strict regulations, but is not adequate for banks which operate on different product and/or geographic markets.

The weak relationship between market structure and performance in banking may also be explained by a size-dependent capital–asset ratio. If larger banks have a lower capital–asset ratio than small banks, their return on assets is lower ceteris paribus (Peltzman, 1984: 652). Another explanation is that banks may use their market power to reduce risk or to engage in expense preference behavior, rather than report high profit rates. Both hypotheses are supported by empirical evidence (see Section 6 above).

Indeed, X-inefficiencies, brought about by lacking managerial ability to control costs or maximize revenues, have been found to be especially important in banking. They account for at least 15–25% of costs, whereas at most 5% of costs are attributable to scale and scope inefficiencies (Akhaven et al., 1997: 98; Allen and Rai, 1996; Bauer et al., 1993; Berger et al., 1993: 222; Lang and Welzel, 1995a). The largest X-inefficiencies are observed for the largest banks in separated banking countries (Allen and Rai, 1996). Empirical studies on bank mergers in the
US confirm that there is a potential for cost efficiency improvement from mergers, but show that it is not realized (for a survey see Akhavein et al., 1997: 99). Also event studies which examine the short-term reaction of the merging banks’ share prices to the announcement of the merger find only small positive or even negative effects. Compared to the results from event and outcome studies for industrial mergers, this points to significant negative long-run effects of bank mergers (Tichy, 1990: 379).

Profit efficiency effects, which may occur by revenue-enhancing changes in output after a merger have been studied only by Akhavein et al. (1997) for the US banking megamerger of the 1980s. They show that profit efficiency increased by a higher loan/asset ratio due to an improved diversification of risks. However, the potential for profit efficiency improvements by mergers should be very different for the group of small banks which specialize on different functions than large banks. Risk reduction by diversification is a core strategy of large banks, which specialize on large risky credits and large customers in national markets and financial centers. Small banks, in turn, tend to specialize on information gathering and the provision of transaction cost-intensive services to small customers in local markets (Neuberger, 1994: 61; Tichy, 1990: 373).

Following Tichy (1990: 376), we conclude that there is no answer to the question about the optimal size of a bank, because it is ill-posed. The effect of mergers and acquisitions on cost efficiency, profit efficiency and market power is likely to be different for large and small banks. For the largest banks in Europe as in the world, there is no significant relationship between size and profitability, which indicates either absence of market power and efficiency effects or, more plausible, a compensation of market power gains by decreasing returns to scale (Tichy 1990: 370). Also the merging to financial conglomerates does not seem to be efficiency enhancing, because there is no evidence for global economies of scope.

8. Public Policy

The banking industry differs substantially from non-financial industries in the extent and kind of public policy which affects structure, conduct and performance. Banking markets are regulated in order to protect small depositors from wealth losses due to an underrepresentation of their interests in the event of a bank’s default ("representative hypothesis"; Dewatripont and Tirole, 1994: 35). Moreover, the regulations should prevent negative real effects of instabilities, liquidity shortages and bank runs, and negative incentives caused by asymmetric information. They can be grouped into protective and prudential regulations and competition policy (see Figure 1).

Protective regulations aim to protect a bank from an impeding default or its depositors from a loss of their deposits in the event of default (Borghof and Rudolph, 1996: 39). This may be done by a public deposit insurance or by a lender of last resort. Whereas the former is a contractual arrangement to ensure banks against illiquidity, the latter is a discretionary policy to provide liquidity ex post. These policies have, however, undesirable effects on banking conduct, since they are likely to induce moral hazard by the owners or managers. They create incentives to take too many risks or to compete too aggressively by raising deposit rates, while at the same time the depositors have less incentives to monitor the banks. As a consequence, expense–preference behavior by the managers and
X-inefficiency may increase. These adverse incentive effects are less if the scope of insurance is reduced or if the insurance premiums are related to the riskiness of the banks' assets (Baglioni and Marotta, 1993).

Prudential regulations shall reduce the probability of a bank’s bankruptcy and a crisis. They consist in regulations of the bank portfolio (capital rules, liquidity rules, diversification rules, prohibition of some operations), entry requirements, information disclosure rules and accounting requirements (Burghof and Rudolph, 1996: 40). Because bankruptcy risk depends on the ratio between equity, debt and asset risk, capital structure regulations are of key importance (Dewatripont and Tirole, 1994: 5).

Capital adequacy rules may be an effective instrument in constraining asset-substitution moral hazard, or, in its extreme form, gambling for resurrection. A higher capital ratio increases the co-insurance of the owners, reducing their incentive to shift risks to uninformed investors (Baglioni and Marotta, 1993; Dewatripont and Tirole, 1993). Capital requirements may also prevent bank runs if they credibly provide safety to depositors (Dowd, 1992b). On the other hand, they may prevent banks from signaling quality to the market or to reduce agency conflicts by capital decisions (Berger et al., 1995). They may even provoke adverse incentive effects, reducing the incentives of effort-averse inside equity holders to monitor loans, which should result in a higher bank failure risk (Gehrig, 1996; Boot and Greenbaum, 1993).

Some operations may be prohibited, because they are considered as too risky, such as investments in industrial shares. This prohibition may, however, restrain the ability of banks to acquire information and provide efficient long-term relationships to firms (Neuberger, 1997a; Neuberger and Neumann, 1991). The separation of commercial from investment banking (in the US by the Glass–Steagall Act of 1933) may be justified for prudential reasons, or because it prevents conflicts of interest in the commercial banking and investment business. On the other hand, it reduces the scope efficiency of banks, if there are complementarities in the production of both services, or induces inefficient investments, if loopholes are made out (N.N., 1996). Empirical support for higher efficiency, lower loss performance and hence superior profitability of universal banks compared to separated banks has been found by Steinherr and Huveneers (1994).

If entry requirements are especially strict for prudential purposes, the structure–performance relationship should be stronger because pricing is less affected by the threat of entry by banks not already in the market. A special type of entry regulations are legal obstacles to the establishment of branches in different states or regions. They do not only reduce competition, but also the scope for risk reduction by diversification. Cross-country analyses show that in regulated markets, financial margins are larger, profitability is higher and wages are well above average (Gual and Neven, 1993; Lewis and Pescetto, 1996:28).

Competition policy towards banks may consist in price regulations and antitrust laws. Price regulations such as ceiling interest rates on deposits which have been common in the US (Regulation Q) as well as in Europe should weaken the relationship between concentration and interest rates on deposit markets (Gilbert, 1984:626). Moreover, they induce distortions by suppressing price competition and inducing banks to compete through quality and to cross-subsidize products. As shown by location theory, rate regulation induces overbranching (Vives, 1991: 14, Chiappori et al., 1995). Antitrust policy which aims at preventing monopoly power is likely to be less strict in banking markets than in goods markets, because
of the fear that tough competition induces too much risk-taking, causing a banking crisis.\textsuperscript{13} Moreover, credit market competition may be inimical to the formation of mutual beneficial relationships between banks and firms. Banks are more likely to finance credit-constrained firms when credit markets are concentrated because then it is easier for them to internalize the benefits of assisting the firms (Petersen and Rajan, 1995). Thus, competition in banking does not always enhance allocative efficiency.

9. Conclusion

Technological developments and important deregulations in the banking industry have increased the interest in IO-related research in the last decade. Our review of the empirical literature shows that most of the structure–performance studies in the US have been conducted in the years of strict branching regulations. Because most banks were restricted to small, local markets, banking was a convenient laboratory for IO research. This does not apply to the universal banks of Europe and the US banks after the break-down of geographical barriers. These banks compete with banks of different size in different markets.

The theoretical banking literature explains why competitive conditions are different in different segments of commercial banking. Wholesale banking products are relatively homogeneous, so that Bertrand or Cournot competition may work. They are mainly provided by large banks in national or international markets. Retail banking products offered to households and small and medium-sized firms are heterogeneous and affected by market imperfections (transaction costs and asymmetric information). Here, quality competition or price competition under product differentiation are relevant. Because of sunk costs of information gathering and investments in branch or ATM networks these markets are unlikely to be contestable. However, the innovation of direct banking due to technological progress and changing customer preferences has increased competition in the retail banking market.

Theoretical and empirical research suggests that there is no optimal size of a banking firm, because banks of different size specialize on different tasks. Scale and scope inefficiencies are small relative to X-inefficiencies, which indicates that bank managers have large discretionary power to pursue their own goals. This may be due to regulations and/or to severe agency problems between managers, owners and depositors. Special regulations which may increase X-inefficiency or market power of banks are prudential and protective regulations. Since competition may conflict with the stability of the banking sector or with beneficial lending relationships, there are incentives for a lax competition policy. To conclude, more research needs to be done which differentiates between different product and geographic markets and accounts for banking size, portfolio structure and regulatory settings.

Notes

1. However, the Diamond/Dybvig model has been subject to severe criticism which doubts that it explains phenomena of real-world banks (Dowd, 1992b).
2. Asymmetric information and agency problems are not as fundamental for industrial firms, whose existence can also be explained by other market imperfections, especially transaction costs (Coase, 1937).

4. The interest rate differential must reach 1–3% in order to make private customers switch to another bank (Neven, 1990: 170).

5. For a review on arguments for increasing returns to scale which should lead to a natural monopoly in banking see Dowd (1992a).

6. In Germany, many direct banks have been founded as independent affiliates of branch banks, which provide them their network for these purposes. Others have established their own ATM networks (Neuberger, 1997b).

7. I am indebted to an anonymous referee for this argument.

8. To give an example: the German retail banking market was characterized by presence competition with an enlargement of branch networks until 1980 and by quality competition with an improvement of customer services in the period 1980–1990. In the 1990s, the possibilities to gain market share by improved quality had been exhausted, inducing price and cost competition. This led to the foundation of discount brokers and direct banks in the market segment of standardizable services (Schütt, 1995).


10. Cost efficiency is improved, if costs per unit of output are reduced for a given set of output quantities and input prices. This may be done by increasing scale efficiency, scope efficiency, or X-efficiency (managerial efficiency). Profit efficiency is improved, if superior combinations of inputs and outputs are realized (Akhavein et al., 1997:96).

11. However, the relationship between concentration and price varies for different concentration levels and time periods (Jackson III, 1992; Berger and Hannan, 1992).

12. Such a regulation was imposed by the McFadden Act of 1927 in the USA. It has, however, been reformed by the Interstate Banking Efficiency Act of 1994, which allows interstate branching.

13. See the exemption of German banks from the per se rule against cartels in the “law against restraints of competition” (Section 102 GWB). In other countries, too, regulators have allowed banks to collude (Vives, 1991:15).

References


