ANTHONY M. SANTOMERO

Modeling the Banking Firm

A Survey

THIS PAPER REPORTS on the status of the literature on micro bank modeling and assesses our understanding of the banking firm's optimal behavior. This is no mean task, for much has been written on banking, broadly defined, over the past couple of decades.

The review is developed in pieces. Each major subproblem is outlined and the analysis used to deal with the issue explicated. This is not the best way to summarize the development of a field, it should be immediately recognized. One would prefer to have a smooth continuum of development, moving the frontier of knowledge evenly through time and across subareas. Yet, this is rarely the way a field develops. More likely, individual questions attract attention and are the subjects of a substantial number of contributions. After a time, the field moves on to the new area of interest. The banking field is no exception.

Before embarking upon the review, however, a couple of lines should be devoted to previous attempts. There have been essentially three. First, Pyle (1972) analyzes the uncertainty portfolio models at a time when little existed in the literature, and hence one finds the review a bit vague and sketchy. Baltensperger's contributions (1978, 1980) are the next serious and rather extensive reviews. The quality of these

This paper was prepared for the Eighth Annual Economic Policy Conference of the Federal Reserve Bank of St. Louis, held on November 4–5, 1983. The author would like to thank Mark J. Flannery, Stephen M. Goldfeld, Michael L. Smirlock, and Richard Startz for their insights and discussions of earlier drafts.

ANTHONY M. SANTOMERO *is professor of finance, the Wharton School, University of Pennsylvania.*

Journal of Money, Credit, and Banking, Vol. 16, No. 4 (November 1984, Part 2)

comes through after reading the literature itself. This review, in fact, owes much to them.

1. WHY DO BANKS EXIST?

Before entering the analysis of the banking firm, it is, perhaps, relevant to ask why the institutional banking structure exists at all. This question has two functions. First, it allows us to discuss some of the more fundamental work in the banking area that centers upon the nature of the financial market structure that is being exploited by the banking firm. Second, it focuses our subsequent discussion of the behavioral models by specifying the nature of the role played by these institutions.

There are basically three approaches to the question of why internal financial institutions exist in the financial market. Each approach centers upon a specific portion of the bank's activity. The first relates to the role played by these institutions as asset transformers. Here, interest centers upon diversification potential and asset evaluation as a reason for these financial firms. The second refers to the nature of the liabilities issued and their central function in a monetary economy. Indeed, the existence of a medium of exchange creates an opportunity for its issuer to gain some form of seigniorage. Finally, some have emphasized the two-sided nature of these financial firms as critical in any explanation of their behavior. Although each of these explanations can find its roots in the seminal literature of financial intermediation, such as Goldsmith (1969) or Gurley and Shaw (1960), all have been specified more formally in the papers cited here.

A. *Asset Transformation Function*

Within this set of explanations for financial intermediation there are two distinct views, namely, asset diversification and asset evaluation functions of the financial firm. The first reason given is presented in the work of Klein (1973), Benston and Smith (1976), and Kane and Buser (1979). The authors argue that a fundamental role of intermediation is transformation of large-denomination financial assets into smaller units. Klein (1973) emphasizes the ability of the firm to exploit the sub-optimal portfolio choice of the depositor faced with unit constraints. The firm offers a risk-return combination in its financial assets that dominates the households' constrained set, even though economic profit is being earned by the bank providing such divisibility services. Benston and Smith (1976) make a similar argument concerning transaction cost minimization. Kane and Buser (1979) argue, at least implicitly, that these divisibility problems favor the use of a financial institution to diversify for both its depositors and equity holders. The latter authors also offer evidence to suggest that such diversification is supported by the portfolios held by the banking sector.

The second explanation on the asset side is currently receiving much attention. It argues that the bank is fundamentally an evaluator of credit risk for the uninitiated depositor. These banks function as a filter to evaluate signals in a financial environment with limited information. Financial agents are either pathologically honest or dishonest, but due to imperfect information, participants find it difficult to evaluate the quality of signals or the honesty of agents. This gives rise to financial intermediaries whose primary role is the evaluation and purchase of financial assets.

Leland and Pyle (1977) were the first to propose this view of financial intermediation. Much of the recent literature springs from their contribution. Starting with a firm's debt-equity decision, these authors argue that because of imperfect information concerning the value of the underlying project investors can glean some information about its quality by observing the willingness of the insider to invest equity capital in the endeavor. Accordingly, the financial structure of the firm adds information to the market. Leland and Pyle (1977) extend this approach to financial intermediation, arguing that this signaling problem could explain financial intermediation. The lack of adequate information on the quality of financial assets requires a set of firms whose primary output is signal evaluation. However, the output from such firms is quite fragile. Once resources are invested in obtaining such information, it becomes available to the market as a public good. The firm, therefore, has difficulty obtaining the return associated with its value. So, it is argued that capturing a return to information can be overcome if the firm that gathers the information becomes an intermediary, holding assets that are found to be of sufficient value.

Campbell and Kracaw (1980), however, counter that this solution to the information problem is conditional upon the assumptions made about the nature of information and its lack of general availability. First, they argue that portfolio choice by knowledgeable firms will be monitored and known in the market, so that honest intermediaries cannot hide portfolio choices. Second, they point out that, in the absence of prohibition, firms with low quality assets would find it in their best interest to produce false information. In this case, even honest firms would have information credibility problems. Therefore, given the uncertainty of the value of the intermediaries' information, the mere observation of the portfolio is not sufficient to resolve the signaling issue. Campbell and Kracaw's solution to the problem is similar in tone to the Leland and Pyle (1977) answer to the nonfinancial firm signaling problem. By dedicating wealth to the firm, the owner-managers of financial intermediaries demonstrate their commitment to the portfolio and signal the value of the underlying assets. They also argue that such firms can emerge which produce information and additional products, such as divisibility or liquidity as noted above. Diamond (1982) subsequently expands on this notion and demonstrates that diversification is an important characteristic in the information uncertainty equilibrium. His results, however, are dependent upon the assumed information and evaluation cost structure. Boyd and Prescott (1983), using a similar framework, note that financial institutions result in optimal consumption decisions with pathologically honest agents. They also argue that the dishonesty issue, which plays a relatively important part in the solutions of these information models, is overemphasized. Criminal charges make the cost of such dishonesty too high, they contend.

This information literature is still in its formative stages, at least as it relates to financial intermediation. Yet, it offers some interesting insights into the existence of the depository firm. The flow of information is obviously a key element in the market for financial assets, but, until now, the reasons for intermediation never clearly defined a role for it in the financial structure. Many questions remain to be answered, however. For instance, how important is honesty in firm information? Why is it so difficult for investors adequately to evaluate the project's specific risk? Are side payments to generate false information a realistic concern, particularly in a multiperiod setting?

B. The Role of the Bank's Liabilities

The second reason given to explain the existence of the banking firm is the central role played by its demand deposit liability as the medium of exchange. It has been many years since Clower's (1967) contribution pointed out the unique function of a monetary unit. The subsequent work of Niehans (1969, 1971, 1978) formalized the choice of monetary unit, and exchange patterns. Throughout, the central feature of a monetary unit is its ability to minimize the cost of transactions that convert income into the optimal consumption bundle. In fact, Barro and Santomero (1976), in their model of a monetary exchange economy, derive an effective price level for consumption goods as a direct result of the costs incurred in converting income into utility-yielding consumption goods.

Brunner and Meltzer (1971) take a different approach to the role of money, but one that is increasingly relevant in light of the recent literature on imperfect information rational macro models. To these authors, money holdings are part of the household's attempt to maximize a utility function in the first and second moments of consumption. Money allows the economic agent to search across the distribution of prices. As Starr (1972) points out, money balances are accepted even if the trade is not excess demand diminishing.

For whatever reason, money is held by the private sector. These balances, although subject to standard cost-minimization analytics, generate profit potential for the issuing institutions. The extent of these gains is dependent upon the characteristics of the money-type liability issued and the explicit pricing structure of the financial institution. The demand for money literature is replete with models that generate positive money balances. The obvious references here are Baumol (1952), Tobin (1956), Barro and Santomero (1972), Feige and Parkin (1971), Miller and Orr (1966), and Santomero (1974). Interested readers should refer to Laidler (1977) and Feige and Pearce (1977) for a thorough review.

The common feature of this literature is the determination of positive money holdings that are a function of transaction costs, uncertainty, and relative rates of return. The monetary mechanism, along with bank pricing decisions, offers the financial firm the opportunity to attract deposits, which may be reinvested at a positive spread. The extent of this profit will depend upon the nature of competition, as Black (1975) and Fama (1980b) point out, and the nature of the transactions network itself. Specific to the latter, the ease of transfer between accounts, the development of cash dispensing options nationwide, and the advent of home banking are all central to the evolution of the banking system's monopoly position.

C. *The Two-Sided Nature of the Financial Firm*

The last line of argument explaining the existence of a banking firm centers upon the conditions necessary for banks to exist as internal financial firms. Here, the most quoted reference is Pyle (1971), which develops a model of the maximizing firm in a financial market with uncertain rates of return. Pyle concludes that covariance between the return on loans and deposits fosters intermediation by encouraging the risk-averse maximizer to transform deposits into loans. Sealey (1980) recently has extended this argument to consider the case where rates are set by the intermediary, rather than given by the open market. Here, a correlation between profits and the level of rates is shown to be equally important as an explanation for financial intermediation. In both cases, the covariance reduces the uncertainty around expected profits and, due to concavity, encourages intermediation activity. Therefore, intermediation becomes possible because the firm can engage in risky arbitrage across markets that have different, though uncertain, interest rates. The firm's value is achieved by the expectation of a positive expected spread across markets, and a sufficiently small variance around this value.

An important question that is not asked in the Pyle (1971) framework is why such expected spreads exist in the financial markets. Presumably, it is because of the nature of its asset or liability services, but this is not modeled explicitly. The literature covered in the earlier two sections, however, offers several viable explanations for the existence of this positive spread for financial intermediaries. In each case, the reason involves some form of deviation from perfect market assumptions. Either there exists nontrivial transaction costs, or there are information problems, or bank liabilities have inherent monopoly potential. In the following sections we deal with optimal bank behavior to exploit such imperfections.

2. THE OVERALL VIEW OF A BANKING FIRM'S PROBLEMS

A financial institution is viewed in the literature as a microeconomic firm that attempts to maximize an objective function in terminal wealth. It uses quantity and/or price variables, such as asset quantities or prices, as control variables. The extent of such choice is model specific and dependent upon the assumed environment and the degree of regulation. Indeed, regulatory constraints may constrain the opportunity set of assets or liabilities, restrict the domain of the solution for one or more of the endogenous variables, or increase the monopoly position of the industry. In all of these cases the banking firm is viewed as seeking a mutatis mutandis solution to its optimization problem. The general form of the problem can be specified as

$$\max E[V(\tilde{W}_{t+\tau})] \tag{1}$$

subject to

$$W_{t+\tau} = W_t(1 + \tilde{\Pi}_{t+1})(1 + \tilde{\Pi}_{t+2}) \ldots (1 + \tilde{\Pi}_{t+\tau}) \tag{2}$$

$$\tilde{\Pi}_{t+k} = \frac{\sum_i \tilde{r}_{A_i} A_i - \sum_j \tilde{r}_{D_j} D_j - C(A_i, D_j)}{W_{t+k-1}} = \frac{\tilde{\pi}_{t+k}}{W_{t+k-1}}, \tag{3}$$

where

$V(\bullet)$ $\equiv$ the objective function, where $\partial V/\partial W_{t+\tau} > 0$ and $\partial^2 V/\partial W^2_{t+\tau} \leqslant 0$
$(\tilde{W}_{t+\tau})$ $\equiv$ the value of terminal wealth at the horizon time τ
$\tilde{\Pi}_{t+k}$ $\equiv$ the stochastic profit per unit of capital during period $t + k$, where $0 \leqslant k \leqslant \tau$
$\tilde{r}_{A_i}$ $\equiv$ the stochastic return from asset i
A_i $\equiv$ the asset category i, where $1 \leqslant i \leqslant n$
$\tilde{r}_{D_j}$ $\equiv$ the stochastic cost for deposit j
D_j $\equiv$ the deposit category j, where $1 \leqslant j \leqslant m$
$C(\bullet)$ $\equiv$ the operations cost function, where $\partial C/\partial A_i \geqslant 0 \quad \forall i$ and $\partial C/\partial D_j \geqslant 0 \quad \forall j$.

Equation (1) is the general form of the objective function to be maximized by the bank and as such allows for two distinct types of behavior. From the first derivative, more terminal wealth is preferred to less. However, the degree of marginal utility depends crucially upon the second derivative. The firm may be an expected value-maximizer or a risk-averse investor. Both choices have been made in the literature, depending upon the author's goals. Specifically, as will be outlined below, if the bank is viewed as selecting a mean-variance efficient portfolio, as in Parkin (1970), Pyle (1971), Hart and Jaffee (1974), or Koehn and Santomero (1980), some form of wealth concavity is assumed. On the other hand, if mean-variance efficiency is not a criterion, or, more generally, when the marginal rate of substitution between risk and return is not the focus of attention, expected profit maximization is assumed. This, of course, is consistent with a linear objective function in terminal wealth, that is, $\partial^2 V/\partial W^2_{t+\tau} = 0$. Models by Klein (1971), Porter (1961), and Orr and Mellon (1961) are typical of this approach.

The choice is not irrelevant to our understanding of the motivation of the firm. Some would contend that it is at the heart of how one views the behavior of the financial institution. Specifically, when modeling the optimization problem, one implicitly defines the motivating force behind bank decisions by specifying the agent and his objectives for the firm. Two choices seem readily apparent, namely, equity investors, or bank management. For the former, a cogent argument can be offered in favor of a linear objective function. This is particularly true in a perfect capital market where financial intermediaries need not exist and the investor's opportunity set spans the institution's choice. Accordingly, any efficient bank portfolio can be perfectly duplicated or hedged by the investor.

In fact, if one views the investor's objective function as the relevant one for bank choice, a more appropriate approach would be one in which the bank's portfolio decisions are determined by global utility maximization of the owners. In this case, equation (1) would be replaced by the risk-averse utility function of the equity holder who views the bank's portfolio selection as a subset of his or her overall decisions. If the investor's opportunity set completely spanned that of the bank, nothing would be gained by the artificial separation of the portfolio choice problem. On the other hand, if any of the arguments of section 1 are correct, the bank's choice set differs from that of the investor due to its ability to exploit financial market opportunities that are not available to the nondepository institution. Yet, even in this case, a relevant factor in the choice of a bank portfolio is the covariance of the returns between the bank's equity with the other components of the investor's portfolio. This idea, however, has never been incorporated into bank modeling. Rather, the bank is viewed as determining a submaximization conditional upon an assumption about other personal wealth allocations being fixed. Consistent with this view, therefore, it is argued that a risk-neutral objective function should be selected for the banking firm to assure its investors efficient allocation, without regard to the risk level that may be hedged elsewhere in the investor's portfolio.

Yet, credible arguments have been offered in favor of the assumption of utility function concavity. The traditional corporate finance explanation rests upon the assumption that management is responsible for decision making and its inability to diversify its human capital. Fama (1980a) has developed this line of reasoning most completely. A similar argument has been made concerning the insufficient owner diversification by O'Hara (1981). In an industry in which only about 1 percent of the institutions have widely traded equity, the latter appears at least partially valid. The agency problems reported by Ross (1973) have also been used to motivate the concavity by Santomero (1983b). According to this view, investors with linear utility functions establish reward schedules for management that lead to risk-averse behavior. Draper and Hoag (1978) use a similar approach to intermediation based upon the information arguments above and Ross (1977). Recently, the presence of bankruptcy costs, first modeled rigorously by Stiglitz (1972), has added yet another explanation. Using this approach, the expected value maximizer behaves as if variance had negative value due to the probability of firm default with its attendant bankruptcy cost. Whether any or all of these reasons are sufficient to motivate the assumption of risk aversion is still open to question. Yet, most of the literature seems to have embraced this approach to the banking problem, at least when necessary.

Accepting concavity does not end the problem for the model builder, as the exact form of the objective or utility function must be specified. The common approach here is to rely upon the quadratic or the exponential forms. Alternatively, a simple two-term Taylor expansion is taken as indicative of the objective function, with appropriate caveats from Pratt (1964).

In equation (2), the general specification above is defined as a multiperiod valuation problem. Frequently, independence between periods is assumed so as to make the maximization a single-period analysis. This is clearly the case for portfolio models, which, to this author's knowledge, have never been cast in a multiperiod framework for the banking firm. On the other hand, important contributions have developed from a multiperiod view of the customer, loan, or deposit relationships. Goldfeld and Jaffee (1970), Flannery (1982), and Blackwell and Santomero (1982) are symptomatic of intertemporal approaches that show that balance sheet constraints make single-period problems dependent upon multiperiod decisions. This will be true, in general, even though those decisions are state variables at the time of quantity or pricing decision making. To motivate such models, some intertemporal stickiness must be asserted which is often difficult to model carefully. Accordingly, these intertemporal models frequently look like ad hoc specifications of reasonable a priori intuition.

Equation (3) defines profit per unit of capital invested by the owners of the firm or their management representatives. In the second specification, one sees that the optimization procedure involves the dual choice of leverage and portfolio components. Each has its own developed literature. The first involves the models to derive the optimal capital structure of a banking firm, for example, Taggart and Greenbaum (1978) and more recently Orgler and Taggart (1983). The second represents a large array of models of profit maximization, conditional upon leverage. In fact, most of the present review will devote itself to profit maximization or objective function maximization without reference to capitalization issues. Implicitly, the latter is assumed to be solved via a submaximization conditional upon the given portfolio or exogenously determined by regulatory standards.

Only recently have these two issues been brought together. Solving (1), (2), and (3) results in a joint decision of portfolio structure and leverage, along the lines of Pringle (1974), Kahane (1977), and Koehn and Santomero (1980). However, this is accomplished by using only the most general asset choice specification and ignoring operating cost considerations totally.

The issues surrounding operating or real resource costs have attracted increased attention lately as the industry itself views these operational issues of paramount importance. In fact, it plays a fairly central role in Baltensperger's (1978) review. This is because there is a large literature on the trade-off between rate spreads and production cost using, traditionally, a single-period profit-maximization framework. For example, the issue of reserve management, brought to our attention by Orr and Mellon (1961), involves such considerations. Sealey and Lindley (1977) and Towey (1974), have developed these models to consider appropriate product mix and scale. Mitchell (1979), Startz (1983), and others have recently been analyzing the interplay between explicit versus implicit charges for operating costs.

Finally, the pricing and quantity decisions in both the loan and deposit markets depend crucially upon the nature of market structure assumed, even in a single-period case. On the deposit side, there is a long literature on deposit rate setting based upon various assumed deposit market structures, for example, competitive, monopolistic, sticky, and tied-product relationships. Models by Goldfeld and Jaffee (1970), Klein (1972), Stigum (1976), Sealey (1980), and Flannery (1982) are representative of this genre. Likewise, but somewhat belatedly, the loan market decision has been analyzed. As above, we consider competitive, monopolistic, sticky, and tied-product relationships. The portfolio models of Hart and Jaffee (1974) and

Parkin (1970) are typical of the first; Porter (1961) and Klein (1971) are representative of the second. The sticky price paradigm, also known as the non-market-clearing or rationing literature, has its roots in Freimer and Gordon (1965), Jaffee and Modigliani (1969), and, more recently, Jaffee and Russell (1976) and Stiglitz and Weiss (1981). On the loan side, the pricing issue also includes a treatment of the interrelationship between the choice of interest rate and credit risk. This aspect of lending activity has received increased interest, with Ho and Saunders (1981), Deshmukh, Greenbaum, and Kanatas (1981), and Santomero (1983b) as examples of this approach.

To treat the individual areas listed above, it is necessary to concentrate much more heavily on the structure of the problem and its solution technique. For this reason, the subsequent sections are devoted to the individual issues raised in bank modeling. However, it is important to recognize that these individual problems feed directly into the global maximization problem outlined in equations (1), (2), and (3). For consistency's sake, one would like to have a decomposition procedure that results in an optimal global choice. Although this may be desirable, little academic work has attempted such a global reconciliation.

3. ASSET ALLOCATION MODELS

The first set of issues treated here will be those relating to asset choice. Models of asset allocation are primarily of two types, namely, reserve management models and portfolio composition models. The first of this set considers the problem of the optimal quantity of primary or secondary reserves to be held by a bank that is subject to stochastic reserve losses due to uncertain deposit levels. The second is devoted to the allocation across risky assets according to risk and return. As Patinkin (1965) points out, the separability of these two problems is not in general accurate as the opportunity cost associated with holding reserves is determined by the quantity of risk in the asset portfolio. However, this notion does not play a primary role in the modeling below.

A. *Reserve Management Modeling*

This literature has perhaps the oldest lineage in bank modeling, tracing its early roots to Edgeworth (1888). In addition, and perhaps more relevantly, the modeling of reserve management was started and largely completed in the two decades following the Orr and Mellon (1961) contribution. Subsequently, contributions and corrections have been made by Poole (1968), Modigliani, Rasche, and Cooper (1970), Cooper (1971), Frost (1971), Baltensperger (1972a, 1972b, 1974), Brown (1972), Knobel (1977), and Ratti (1979). Baltensperger (1980) has the most exact review of this area from which the present section draws.

The basic model of reserve management can be outlined as follows. Consider a bank that has two assets: noninterest bearing reserves, R, and earning assets, A, yielding r_A. Its deposits, credit lines, and loan repayments are all subject to random shifts without notice, with net bank withdrawals defined as X, with the density function $f(X)$. If net withdrawals exceed reserves, that is, $X > R$, the bank must undergo a proportional cost, c, to obtain the additional funds. The bank, therefore, wishing to maximize expected profit from its deposit balances, must maximize the following function:

$$\tilde{\pi} = r_A(D - R) - \int_R^{\infty} c(\tilde{X} - R) f(X)\, dX \, , \tag{4}$$

where loans are equal to deposits less reserves for expositional simplicity. The first-order conditions of this problem result in a reserve quantity that satisfies the condition

$$r_A = c \int_R^{\infty} f(X)\, dX \, , \tag{5}$$

that is, the opportunity cost of reserves, on the margin, must equal the expected reduction in operating or transactions costs devoted to reserve adjustments. This is the essential condition that must be met in reserve management models.

Poole (1968) notes that for a symmetric distribution with $E(X) = 0$, this condition implies that $c > 2r_A$. Both he and Baltensperger (1980) devote considerable time to explaining why this might be the case. For example, the cost of reserve deficiency, c, may not be linear; access to funds may be viewed as uncertain; and, administrative and regulatory hassles may ensue for firms with excessive dependence on the open market to adjust reserve positions.

Incorporating reserve requirements into the analysis further complicates the matter, but the economics are unaltered. Specifically, the existence of required reserves under a contemporaneous reserve requirement regime reduces the effect of an unexpected deposit drain by the amount of reserves held in its support. On the other hand, lagged reserve requirements, unless carryover provisions exist, appear to leave the analysis unaffected.

The Poole (1968) paper obtains another interesting result. Analyzing the effect of a mean preserving spread of $f(X)$, he concludes that increased uncertainty does *not* necessarily result in higher reserve positions. The economics of the result are that, with increased uncertainty, the bank is more likely to have both excess and deficient reserves. Inasmuch as the former encourages a reduction in reserves, the net effect of uncertainty on reserve assets is ambiguous.

Frost (1971) extends the reserve management issue to an environment where the bank inherits the previous reserve position. At any point in time, the ex post reserve position results from reserve planning and a single draw from the $f(X)$ distribution. The question addressed is whether reserves should be immediately adjusted to the ex ante optimal level for the next period. Here, the author applies a Miller and Orr (1966) approach to adjustment of reserves, where fixed adjustment costs and the previous period's endowment must be evaluated to determine if a movement to the optimal level is profitable. As is the case with all such models, a range of "nonoptimal" reserve positions exists where no adjustment will take place.

Baltensperger (1972b, 1974) and Baltensperger and Hellmuth (1976) extend this analysis of reserve management into its relationship with information and operating costs. They note that information on customer habits may reduce the variance of the underlying subjective distribution of stochastic withdrawals. The latter could reduce reserve costs sufficiently to warrant investment in such information. Accordingly, the profit function (4) should be written as

$$\pi = r_A(D - R) - \int_R^{\infty} c_1(X - R) f(X \mid \phi)\, dX - c_2\phi , \tag{6}$$

where ϕ is an information unit that costs c_2 to obtain. The ex ante expectation of the underlying deposit risk is dependent upon the quantity of information investment. The resultant first-order condition requires that the expected marginal return from information production be equal to its constant marginal cost.

In all, this literature is fairly self-contained and complete. It solves a problem that appears central to the banking environment and obtains fairly sensible results. However, it could be argued that, with the advent of a well-developed federal funds market, the overall importance of the problem as a micro banking issue is significantly reduced. In fact, casual empiricism suggests that total excess reserves amount to only about 0.1 percent of total bank assets, whereas the proportion of the academic literature associated with its determination is one hundred times that figure. This is, perhaps, unfair as the focus of the reserve management literature transcends just excess reserves and includes the quantity of short-term assets and the availability of credit facilities to meet reserve deficiencies. Yet, this feature of the reserve management problem is not part of the model framework. Indeed, the major shortcoming of this literature is its implicit assumption that all liquidity considerations can be reduced to the reserve management problem. It never squarely addresses the issues surrounding the definition of liquidity, its measurement in the balance sheet, or its optimal quantity. In an environment where liability management has gained increased importance as a liquidity and reserve adjustment tool, a dependence upon reserve items seems increasingly less relevant. In addition, given the interdependent nature of the individual withdrawals from any particular bank, the lack of consideration of contagion or bank-run phenomena seems to avoid a significant aspect of the bank's short-term funds management.

B. *Portfolio Choice Models of Asset Allocation*

Asset choice modeling in the literature generally takes two forms. Either the bank is viewed as possessing some degree of monopoly control over its loan price or the asset market is modeled as a perfectly competitive one where the bank must select appropriate quantities of loans of various characteristics. The first set is typified by Shull (1963), Klein (1971), and Porter (1961), and the second by the contributions of Pyle (1971), Parkin (1970), Hart and Jaffee (1974), and Sealey (1980).

The first group of papers seeks to obtain an optimal loan and/or asset size from the maximization of expected profit of the firm. In virtually all cases, the objective function is linear in profit. Frequently, uncertainty does not even exist, for example, Shull (1963) or Stillson (1974). In all cases, the approach is a rather straightforward microeconomic setup from which the traditional marginal conditions result. A controversy central to this literature and perhaps one which separates it from work on the behavior of nonfinancial firms is the modeling of the most liquid, or highest elasticity, market.

In its simplest form, the typical model structure is a two-sided discriminating monopoly. Marginal revenue equals marginal cost. A variation in any one market feeds through the model to a comparative static shift in all marginal conditions. Shull (1963) exhibits such behavior.

Another approach, apparently initiated by Tobin (1958), has one market as a perfectly competitive one. Tobin used the discount window for this purpose, a position that was later rejected, at least implicitly, by Goldfeld and Kane (1966). Klein (1971) uses the government security market. Others have suggested it is the federal funds markets. In any case, the maximization associated with a model with one infinitely elastic source (use) of funds results in a general convergence of all marginal revenues and costs to this infinitely supplied (demanded) asset. Separation exists between asset allocation and deposit structure. Variations in demand for loans of one type have no effect on the other loan decisions. This result, first pointed out explicitly by Klein (1971), caused considerable debate, with comments by Pringle (1973) and Miller (1975) questioning the validity of the infinite elasticity assumption.

To formalize the discussion, one may view this type of model as the result of the following maximization process:

$$\max_{A_i, D_j} \pi = \sum_i r_{A_i} A_i - \sum_j r_{D_j} D_j \,, \tag{7}$$

where

$$\frac{\partial r_{A_i}}{\partial A_i} < 0 \qquad \forall i$$

$$\frac{\partial r_{D_j}}{\partial D_j} > 0 \qquad \forall j \neq m$$

$$\frac{\partial r_{D_m}}{\partial D_m} = 0 \,.$$

The first-order conditions result in

$$\left(\frac{\partial r_{A_i}}{\partial A_i}\right) A_i + r_{A_i} = r_{D_m} = \left(\frac{\partial r_{D_j}}{\partial D_j}\right) + r_{D_j} \,. \tag{8}$$

Baltensperger (1980) takes issue with this formulation of the bank's behavior on the grounds that no operating costs are included in the analysis and differential assets or deposits are indistinguishable within the formulation. Although both of these criticisms are entirely correct, they do not appear to be particularly relevant or fundamental. Stillson (1974) incorporates servicing costs, as do Towey (1974) and Sealey and Lindley (1977). In none of these cases does one find results that substantially affect the model's basic conclusions.

To obtain a nonobvious generalization through the use of operating costs or a production technology, one must go in the direction of Adar, Agmon, and Orgler (1975). There the authors argue that the banking firm has a joint production problem so that output mix is a critical determinant of operating expense. Within this framework, the equilibrium conditions from profit maximization are interdependent. Separation of loan decisions, as well as asset decisions from liability structure decisions, becomes impossible. This, of course, makes the operating cost function much more complex than it is generally considered to be in the theoretical literature. Perhaps it is also more realistic. In an era in which financial intermediaries are broadening their menu of financial series to include activities in both the deposit and securities areas, this joint production view of operating costs may warrant serious consideration.

Before turning from these simple micromodels of the firm it is worth noting what is not treated in most of these theoretical constructs. Almost without exception, these models represent the bank's problem as a single-period maximization. Intertemporal demand considerations are rarely explicitly modeled, in spite of the fact that the institutional literature spends much of its time discussing such problems. The customer relation that Hodgman (1963) and others have analyzed is rarely incorporated into bank models. Where it is, the treatment tends to be rather sterile, with the entire relationship captured by a geometric lag or by cross-product proportionality. Wood (1974) or Blackwell and Santomero (1982) are typical of such attempts, but represent only the most rudimentary approach to capture the dynamic effects of customer relationships.

Multiperiod firm maximization, too, is only peripherally treated. Where intertemporal concerns exist, the objective function is taken to be a multiperiod discounted valuation function. However, if the literature on agency costs is relevant, and there is a conflict between management and investor goals, it would be interesting to see a better development of this issue in the multiperiod environment. For example, my colleagues have argued that front-end fees in the syndication area have caused bank management to accept risk levels that would have been rejected if a present value decision rule had been used (see Guttentag and Herring 1980). Perhaps this could be due to a conflict between manager and investor discount rates or horizons. Unfortunately, without better multiperiod modeling, theory can say very little about this subject.

Where intertemporal models do exist (e.g., Wood 1974), the solutions have some curious characteristics. For example, consider a simple n period model, such as

$$\max_{r_t, r_{t+1}, \ldots} \pi = \sum_{t=1}^{n} \beta^{t-1} \pi_t , \tag{9}$$

subject to

$$\pi_t = r_t l_t - r_{D_t} D_t ,$$

where $l_t \equiv l(r_t, l_{t-1})$ and $\partial r_{D_t}/\partial D_t = 0$. The first-order conditions for $r_t, r_{t+1}, \ldots$ are

$$\frac{\partial \pi}{\partial r_t} = 0 = l_t + (r_t - r_{D_t})\frac{\partial l_t}{\partial r_t} + \beta(r_{t+1} - r_{D_{t+1}})\frac{\partial l_{t+1}}{\partial r_t} + \cdots . \tag{10}$$

The implication of this optimal behavior may be characterized as follows. The bank will initially accumulate loans from period to period as a result of the build up of intertemporal demand based upon previous decisions. This increased loan demand implies that the interest rate will rise each period in spite of a constant cost of funds. This appears to be an awkward way of looking at the bank's dynamic behavior. Yet, only Blackwell and Santomero (1982) consider the properties of their model at the long-run stationary equilibrium.

The second approach to asset selection is the portfolio choice models using risk and return as criteria. The most elementary of these is the work of Pyle (1971), which set out to determine the necessary conditions for financial intermediation within a mean-variance framework. Using a three-asset case where there is a riskless asset, as well as loans and deposits in the intermediary, he derives the optimal behavior of the financial firm. Necessary and sufficient conditions for the existence of intermediation are also derived. The results imply that covariance of rates across the balance sheet is extremely important, and a positive loan premium and/or negative deposit premium must exist for the firm to make risky loans and to intermediate deposits.

These results appear fairly obvious, and Baltensperger (1980) argues that the major failing of the work is its inability to motivate this simple intuitive result. Yet, the paper has attracted quite a lot of attention, primarily because it clearly articulates the portfolio choice modeling of the financial firm.

The portfolio models, of which both Parkin (1970), and Pyle (1971, 1972) are the prototypes, have a rather standard set up. The investor or manager is viewed as maximizing a concave function in end of period profit, with a quadratic or exponential function often used to represent the firm's preference ordering.

$$E(\pi) - \left(\frac{b}{2}\right)\sigma_\pi^2 , \tag{11}$$

where profit and its variance are defined as

$$\pi = \sum_i r_{A_i} A_i - \sum_i r_{D_j} D_j$$

$$\sigma_\pi^2 = E[(\pi - E(\pi))^2] .$$

The solution methodology of these papers is to define the characteristics of the feasible set of assets and/or liabilities of the firm. Then, asset choice is restricted to the efficient frontier, where additional return is only achievable at the expense of added variance. The bank, then, selects a point on the efficient frontier where the objective function's marginal rate of substitution between risk and return is equated with the market's opportunity set. Rather obviously, the determinants of the optimal solution are the specifications of the returns to each asset and liability, and the choice of the objective function of the bank. On the first of these, the return characteristics of the assets are assumed exogenously given and independent of bank decision making. This view significantly minimizes the effect of the bank on the asset return patterns through the manipulation of lending terms and conditions. However, if the entire set of assets from which the bank constructs its portfolio includes multiple pricing options for each loan category, this framework is theoretically correct. As far as the second factor determining the solution, section 2 notes the difficulty in determining the all-important objective function to be maximized.

The above reservations notwithstanding, these models lend new insight into the bank's portfolio selection. Hart and Jaffee (1974) is perhaps the best of the papers using this approach to examine the asset structure of the depository intermediary. Their results show that under some rather restrictive conditions, one can segment the scale of bank operations from its risk-return choice. Such a separation theorem can be developed for the bank in an environment in which no risk-free asset exists and a fully liability funded portfolio structure is assumed.

In total, the asset choice models get mixed reviews. One branch of the literature applies Economics 1 to the bank to obtain results that border on the trivial in retrospect, that is, marginal revenue equals marginal cost. It depends, to an excessive extent, on demand and supply curve slopes that are not well motivated or understood. Alternatively, the Tobin-Markowitz asset portfolio models are used for quantity choice in a perfect market. The results follow directly from the finance literature and add little more. Their insights, and they are many, come from the realization that the bank asset problem is a special case of the standard portfolio choice model. As such, that perception of asset management must become a part of the banking literature.

4. LIABILITY CHOICE MODELS

The modeling of the liability side of the balance sheet has taken two separate paths. For deposits, the modeling has been akin to the simple monopolistic modeling structure outlined for the asset side. For capital and leverage issues, the literature has used more sophisticated techniques, including bankruptcy considerations and models from corporate financial theory. We consider each in turn.

A. *Deposit Modeling*

A useful starting point for the analysis of liability modeling techniques is the simple deposit market form used by Klein (1971). Here, as indicated above, there is an infinitely elastic market from which an unlimited quantity of funds may be obtained. As noted in equation (7), the bank is a deposit rate setter with some monopolistic control over the deposit market. First-order conditions from such a setup indicate that the marginal cost of funds from each deposit source must equal the marginal cost (use) of funds from the competitive market, as in (7).

This model may be complicated by allowing production costs related either to balances held on deposits or to the number of accounts. Sealey and Lindley (1977), Klein and Murphy (1971), and Baltensperger (1972b) are typical of this genre. The solution of this type of model is a fairly obvious extension of the earlier work; that is, with production costs, the total marginal costs must be equilibrated. The analysis only becomes interesting when different types of production functions are integrated into the analysis. For example, Flannery (1982) considers the case of a quasi-fixed production process whereby the cost function relates to changes in deposit balances rather than (or in addition to) the level itself. Assuming that such customer-specific costs are shared by both the customer and bank, the model demonstrates that deposit-rate variation will be reduced relative to open-market rate movements. Flannery argues that such a production process can explain both the long-run customer relation and the tendency for deposit rates sometimes to lag behind open-market rates, both on the up and downside of the market. To Flannery, such quasi-fixed production technology explains the concept of "core" deposits.

Mitchell (1979) extends the analysis of the production or operating costs to a case where, due to regulatory constraints, explicit interest is incapable of rewarding the depositor sufficiently. Implicit payments then result from the willingness of the bank to absorb some of the real production costs associated with transactions in order to attract larger balances. Barro and Santomero (1972) and Santomero (1979), who treated the customer behavior within such a regime, demonstrate both the relevance and importance of such implicit payments. Mitchell (1979) and Startz (1983) go the next step to consider the effect of this implicit payment option on the behavior of the banking firm. Central to their analysis is the ability of the bank to subsidize banking activity, such as check cashing, funds transfers, and account maintenance costs, to encourage deposit balances. However, their treatment depends critically upon the relationship between such account activities and average balances. The basic model may be developed as follows. Denote the number of transactions using the deposit account as n and the cost and service charge per transaction as c and k, respectively. The profit function of (7) may be written more generally as

$$\pi = \sum_{i} r_{A_i} A_i - \sum_{j} r_{D_j} D_j - \sum_{j} [(c_j - k_j) n_j(D_j)] . \tag{12}$$

The bank must now consider the return to deposit balances, the cost of deposit services, and the interrelationship between the two. Mitchell (1979) notes that increasing the explicit payments allowed on deposit balances within such a framework may result in either increases or decreases in implicit payments according to the relationship $n_j(D_j)$. Although this work is still in its formative stages, it does have some interesting implications for bank decision making and pricing. In fact, the whole area surrounding the relationship between various forms of payment, deposit reaction, and, ultimately, profitability needs further work in a real-world environment that is shifting to a greater extent to rate deregulation.

The movement of deposit rates relative to open-market interest rates is another area of continual attention in the literature. A simple neoclassical model of the firm argues that rates on deposits should move along with market rates. Yet, Weber (1966) and others have suggested that rates will move sluggishly for institutions with long-lived assets. The general argument is that due to the inability of the firm to achieve a satisfactory increase in the asset yields from its portfolio, the bank will be forced to adjust its deposit rates more slowly to overall market forces. The Weber hypothesis elicited a significant number of subsequent responses, of which the major contributors were Goldfeld and Jaffee (1970) and Stigum (1976).

The former authors demonstrate that on a theoretical plane the phenomenon can be explained by boundary conditions forced upon the firm by its balance sheet constraint. Specifically, consider a firm that maximizes the profit function contained in equation (7) and attracts assets and liabilities in period one. Optimum profit is obtained for that period and over the expected horizon, defined by expected rates and demand functions. Subsequently, however, assume that there is an unexpected decline in asset yields. The lower rates imply lower optimal deposit rates given a fixed liability schedule. However, if the asset portfolio is nonmarketable and relatively fixed in quantity, the bank may find itself in a position that requires it to pay whatever rate is necessary to finance its asset portfolio. Later upward movements in asset yields will, accordingly, not elicit an appropriate rate increase, as the initial deposit rate was a nonequilibrium choice by a constrained firm. As Goldfeld and Jaffee note, such sluggishness is introduced only because of the possibility that a corner solution will lead to perverse behavior. In the long run, such behavior is ruled out.

Stigum returns to this issue to present a more general treatment of the problem. Here, two innovations are offered. First, she considers the effect of sticky deposits on the analysis. Assuming that deposit balances can be written as a function of both current rates of return and the previous deposit balance, she demonstrates that deposit inertia will result in relatively sluggish movement in deposit rates even if no boundary condition is reached. This is an interesting insight into the pricing dynamics but the author offers little to motivate, in a rigorous economic sense, this deposit inertia. Flannery's (1982) joint cost-sharing framework is perhaps one such explanation from the industrial organization literature, as is Adar, et al. (1975). However, without adequate understanding of the cause of such sluggish deposit movement, it is difficult to be convinced that it is a satisfactory answer to the relative rate movement observed in the deposit markets.

The second generalization contained in Stigum (1976) is the addition of deposit uncertainty to the analysis of Goldfeld and Jaffee (1970). Considering both the case of risk neutrality and concavity, the author investigates the relative movement of rates. Not surprisingly, the degree of absolute risk aversion is a crucial determinant of bank behavior. This same factor comes into play in Sealey's (1980) slightly more general treatment of the deposit rate setting problem. In addition the latter author considers the correlation of loan and deposit rates as well. This critical feature of bank modeling will be discussed more fully below when issues related to the management of the total bank portfolio are discussed in section 5.

B. *The Capital Decision*

The capital decision of the financial firm is more complicated than it may first appear. This is true because the optimal choice of scale and leverage is determined by the assumed financial environment and the raison d'être of the firm. One must keep in mind the substantive contribution of Modigliani and Miller (1958), illustrating that in the absence of frictions and taxes there exists *no* optimal capital structure. Accordingly, to derive an optimal capital structure, one must determine, first, the role played by the financial institution and, second, the extent to which one wishes to deviate from the perfect market paradigm in explaining its operation. Pringle (1974) does so by asserting exogenous and fixed deviations from perfect market behavior, while at the same time assuming a CAPM valuation model. Not surprisingly, his results suggest that an optimal capital is obtained in this model whereby such imperfections are appropriately exploited. In his framework, there is a divergence between lending and borrowing rates, excess expected returns exist on loans, and capital cost diverges from the risk-adjusted open-market rate. The author shows that optimal capital is attained when the excess marginal revenue on loans equals the excess marginal cost of capital.

Taggart and Greenbaum (1978) develop their analysis of capital under the assumption that excess loan revenues and transaction service profits can be exploited by the value-maximizing firm. An extension of this approach by Orgler and Taggart (1983) argues that the optimal capital structure for the bank depends upon its efficiency of producing such intermediary services as check processing and record keeping, along with the interplay of private and corporate tax rates on profits. An interesting result obtained by the framework is its ability to explain variations in bank capital across firms by differential scale and efficiency of bank operations.

Kahane (1977) and Koehn and Santomero (1980) use the portfolio model approach, outlined above, on the bank capital issue. Starting from the general specification of (1), they optimize the bank's rate of return on capital by selecting a portfolio of assets and leverage position that optimizes shareholders' returns. The Koehn and Santomero paper is interesting because it analyzes the effect on bank portfolio behavior of a regulatory shift in capital adequacy regulations. The authors demonstrate that although capital increases as a fraction of assets, the resultant portfolio is unambiguously more risky than before the capital constraint. Essentially, the constrained firm attempts to offset some of the effect of the leverage limit by absorbing greater risk in its portfolio than before the regulation. Koehn (1979) finds a similar effect of asset restriction regulation. These results lend credence to the idea that regulation, if it is to be effective, must be combined with adequate understanding of the behavioral response of the banking firm.

Talmor (1980) uses a different approach, employing the gambler's ruin technology to the problem. According to this approach, appropriate capital involves the determination of an acceptable probability of bankruptcy and derives an optimal structure to achieve this ex ante acceptable level. Building on the work by Wilcox (1971) and Santomero and Vinso (1977), the paper derives the appropriate level of capital from a continuous time diffusion process whereby the management determines the appropriate ex ante probability of failure by portfolio choice. An interesting characteristic of the model is that the determinants of the process are derived endogenously from the underlying balance sheet of the bank.

For all its benefits, however, this approach has a fundamental problem. It is not clear how one defines the acceptable probability of failure. Presumably, the firm should be managed to optimize firm value for its stockholders. How one links this objective to an ex ante criterion for probability of failure is not developed in such models. This could be done by some form of general objective function in which bankruptcy costs affect shareholders' value, as was discussed in section 2. However, some would contend that the presence of regulation and deposit insurance precludes this market response mechanism from functioning adequately.

The mantle of regulation has in and of itself a built-in incentive to increase risk and leverage. The deposit insurance structure guarantees all depositors up to a statutory limit. For these depositors, the liability of the depository institution is de jure a riskless asset. Accordingly, there is no incentive for these depositors to respond to bank riskiness per se. The noninsured depositors would seem to require some assurance of the solvency of the institution before deposits are made. However, regulation has made their concerns less relevant in terms of the market disciplining the depository institution. Over the past several decades, failed institutions have been dealt with in a manner that has protected all depositors, rather than just the insured category. Assumption of the bank location and balance sheets has been the dominant form of regulatory action. Advances from the discount window to facilitate this manner of resolution lead Kareken and Wallace (1978) to conclude that such advances are essentially loans to the FDIC.

If one accepts this view that bank liabilities are essentially 100 percent insured, then the entire issue of bank capital and risk taking should be recast in terms of a discussion of insurance pricing. The role of the market to price the riskiness of the depository institution has been usurped by the regulator in its blanket insurance of the industry.

What remains, therefore, is to discuss the form and substance of the insurance and the impact it has and can have on the quantity of capital and risk held by the banking firm. Merton (1977) and Sharpe (1978) have made significant contributions in this direction. These papers apply the option pricing literature to the insurance of bank liabilities to obtain considerable insight into the deposit insurance problem. The basic approach used is to show that the payoff pattern of the insurance scheme is like a put option on the underlying assets of the institution. Then, using the work of Black and Scholes (1972), one can derive the optimal price for such insurance and its functional dependence.

Following Merton (1977), one may consider essentially three parties to the insurance of a liability of the bank: (a) the bank itself, (b) the depositor, and (c) the insurance company (Federal Deposit Insurance Corporation—FDIC). The bank offers the depositor a return, B, that compensates the latter for the time value of the funds left at the bank. These funds are commingled with the equity supplied by the bank's owners in the asset portfolio. After one period, the asset value is given by V, which is stochastic ex ante. If the ex post value of the asset portfolio is greater than B, that is, if $V \geq B$, the bank redeems the depositor's liability and the insurer plays no role. The equity holder receives the difference between V and B as a return on equity, presumably less the insurance fee charged by the FDIC. In the event that V is less than the maturity value of the liability, that is, if $V < B$, the depositor still receives the insured value, B, but, in this case, the return is paid partially by the residual value of the assets and partially from the insurance fund. The equity holder receives nothing, whereas the insurer receives $V - B$, which is unambiguously negative. One may rewrite the payoff pattern of the insurance fund equivalently as

$$\min[O, V - B] = \max[O, B - V] . \tag{13}$$

This is identical to the payoff structure of a put option issued by the FDIC against the value of the assets at the bank. Using the Black-Scholes option pricing model to derive the optimal price per dollar, Merton (1977) derives the value of deposit insurance.

Sharpe (1978) develops this same perspective of the deposit insurance pricing problem using a state preference approach. His main insight is obtained by reversing the logic of the optimal deposit insurance pricing scheme. Specifically, the institutional realities in the United States are that FDIC insurance is granted under a fixed net insurance fee. Therefore, there is an inherent tendency for financial firms to accept higher risk levels than they would in the absence of the insurance subsidy. However, note that, given other factors, the value of a fair insurance fee declines as the capital-asset ratio increases. Sharpe points out that adequate capital is that quantity which would make the current fixed rate insurance fee the correct price for the underlying put option implicitly issued by the FDIC.

Buser, Chen, and Kane (1981) argue that appropriate pricing of insurance benefits may not be what the regulator wishes. They argue that the FDIC uses explicit and implicit costs to offset the inherent benefits accruing to the insured liability issuers. The latter includes capital regulation, safety regulations, community development accountability, and the like. Once accepting the benefits of insurance without paying the full cost explicitly, the institution can be manipulated by the regulator. However, the resultant profit of the firm must at least equal the uninsured case to maintain control.

In all, this literature on optimal bank capital is a bit vague and very model specific. This has led to a significant divergence between practice and theory as has been suggested in Santomero (1983a). Although more work needs to be done, the field seems to be awaiting a breakthrough in either of two areas. The corporate finance literature needs to develop further in order to aid in the search for a private determination of optimal capital. In addition, the interest expressed in variable pricing of deposit insurance must become much more of an institutional reality. Without both of these, the capital area remains murky.

5. TWO-SIDED MODELING

Thus far, the analysis has been focused upon the modeling of each side of the balance sheet. Frequently, the models discussed considered the entire portfolio choice problem, but for expositional convenience a review of the analysis was placed in either of the two sections above. For example, Klein (1971) considers the entire asset-liability management problem. However, the treatment is separable due to the structure of the infinitely elastic market. So, too, Parkin (1970) includes positive and negative assets within the portfolio choice under uncertainty, but the main insights gained from the model pertain to its asset choice structure.

Several models have recently appeared, however, which are intimately related to the two-sided nature of the banking problem. The first of these is the recent work by Deshmukh, et al. (1983), in which they look at the alteration in the financial intermediation process associated with increased interest rate variability. These authors argue that the typical bank serves as both an asset-transformer and broker. The former borrows funds at a fixed rate before interest rate uncertainty on the asset side is resolved, and the latter borrows only after interest rates are known. Increased rate volatility shifts the bank's activity more toward a brokerage function and away from the more traditional asset transformation function. The limited upside gain from higher rates is insufficient to compensate the bank for the reduced profit from asset transformation when rates decline. Accordingly, variability results in the shift to brokerage activity even for an expected profit-maximizing firm.

Ho and Saunders (1981) analyze the bank's brokerage function in more detail, using the finance literature on broker bid-and-ask spreads to explain bank margins. Using a diffusion process for rate movement and a concave objective function, they are able to define the optimum equilibrium spread for the banking firm. As one might anticipate, this margin is dependent upon the risk aversion of the firm and interest rate volatility.

One way of shedding interest rate risk is to construct a portfolio so as to immunize it from the effects of rate shifts. There is a whole literature on this problem and the method of application to the banking firm. Bierwag and Kaufman have been the mainstays in this duration and immunization literature, with their most recent work, Bierwag, Kaufman, and Toevs (1982). These contributions develop an appropriate measure of the interest rate exposure of both sides of the balance sheet and a portfolio structure that will not be adversely affected by interest rate movement. This is achieved by the use of Macauley's duration measure, among others, and the two-sided immunization of the balance sheet through duration matching.

This literature is interesting and critical to any treatment of actual bank interest rate exposure estimation. Its development, use, and limitations require more space than can be given here but interested readers can look elsewhere for a critique, for example, Haley (1982) and Bierwag and Toevs (1982).

Some caveats are in order. First, this technique involves the matching of present value changes associated with an arbitrarily small but specific variation in the yield curve. Rightly or wrongly, bankers worry much more about cash flow than the academic researcher. Inasmuch as regulation is frequently cast in current book value terms, this may be rational. By implication the banking firm may not wish to hedge its present value position. Second, one should keep in mind that there is a distinction between balance sheet hedging and equity-value immunization. These two concepts are in direct conflict inasmuch as a duration hedged balance sheet will cause the present value of the earnings stream, that is, the bank's equity value, to vary with rate movement. Third, duration matching requires active portfolio management, which may not be practical in the short run or for small institutions. In fact, Flannery (1981, 1983b) demonstrates that net current operating income is not affected by rate movements for large institutions, but is affected for smaller ones.

An allied area which must be discussed in the context of the banking firm is modeling interest rate risk taking. This area has only recently developed, though there was an early work by von Furstenberg (1973). Santomero (1983b) models the bank's choice between fixed and variable rate loans to analyze the payoff functions and to consider the optimal quantity of each in the overall portfolio. Dealing simultaneously with both credit and interest rate risk, it is demonstrated that both types of risk are likely to be present in the optimal portfolio choice. The driving force behind this analysis is the asymmetry of the variable rate payoff function, as noted by Deshmukh, et al. (1983) in a much different context. The paper demonstrates that credit or default risk is generally related to interest rate fluctuations. Using a concave utility function, marginal conditions require that the bank accept some nonzero interest rate exposure in order to minimize the sum of credit and interest rate risk.

In all, these models begin to investigate the financial firm as a true intermediary that deals simultaneously in both the asset and liability markets and bears both interest and credit risk. Yet much needs to be done before they can purport to explain bank behavior adequately. First, the credit risk decision must be more adequately integrated into the analysis than it has been heretofore. Rather than treating credit risk as a simple mean-variance-covariance decision process, serious consideration must be given to the interdependence both between interest rates and default probabilities and between credit risk at one firm precipitating failure at others. Second, the entire area of interest rate risk management and its optimal level needs further work. With the advent of a developed futures market, the average institution has the capability of shedding interest rate risk rather easily. Yet, little has been done to model this process or characterize the optimal solution. This, in particular, is an area that requires serious modeling and empirical investigation to bring the academic literature up to institutional realities.

6. CREDIT LINES AS AN EXAMPLE OF OFF-BALANCE SHEET RISK MANAGEMENT

While the profession has been building models of the banking firm from the perspective of optimal pricing and allocation of balance sheet items, the banking industry has been substantially expanding into off-balance sheet activity. Standby letters of credit, bankers acceptances, and credit lines have grown sufficiently to cause regulatory concern, as the board staff study illustrates (Board of Governors of the Federal Reserve System 1982). In addition, with the erosion of the traditional lines of commerce restrictions, banks have entered such standard securities market areas as futures trading, underwriting, private placement, and, most recently, discount brokerage. Yet, little has been done in the academic literature to integrate these activities into banking theory.

The only exception is the work on bank credit lines. Campbell (1978), Thakor (1982), and Thakor, Greenbaum and Hons (1981) represent significant advances. The basic approach taken by Campbell (1978) is the modeling of pricing with uncertainty in demand and cost. This has been the subject of a number of papers in the micro theory literature, most notably Leland (1972). The Thakor, et al. (1981) paper moves away from the construction of a general model of the market for credit lines in favor of an in-depth analysis of the optimal pricing of the instrument. The fundamental insight is that the bank is issuing a put option to the firm such that it stands ready to purchase a risky claim, the loan contract, over a given horizon. Then, using the intertemporal CAPM model of Merton (1973) to solve the value of the option, an optimal price is obtained.

There are weaknesses in the application of option pricing to this problem, however. Firstly, partial take-downs of the credit line make the pricing issue much more complex. Thakor, et al. argue that historical analysis and the decomposition of the total credit lines into a series of likely draw-downs can eliminate this difficult case. Yet, this is a somewhat heroic assertion. The option on the time of usage is at least as critical as the value associated with the resolution of price uncertainty. Yet, these models neglect the former in favor of the latter.

Secondly, even the price uncertainty resolution appears somewhat artificial. Credit lines have two pricing conventions, fixed and pegged. The former is a minor part of the market in which the bank specifies a fixed rate loan commitment to the firm for a reasonably short space of time. For these lines, the option pricing model is appropriate. For the large majority of commitments, however, the price of the loan commitment is tied to open-market rates. The option in this case relates to a hedge against the variability of the markup over prime or funding cost. The firm is betting either that credit tightness will increase bank margins or that the fortunes of the firm will deteriorate to such an extent that the upfront cost of the credit line is warranted. As the above authors point out, if the borrowers knew that the terms would be identical at draw-down and at commitment dates, there would be no incentive to purchase loan commitments. The uncertain availability, which cannot be modeled very neatly in these sorts of models, appears crucial to our understanding of the credit line market.

Others have taken another approach to the credit line literature, which is more along the lines of availability premiums. Wood (1975), for example, tries to deal with the firm reaction to a nonzero probability of lack of available funds. So, too, Blackwell and Santomero (1982) model the credit lines as a way for firms to jump positions on the queue of acceptable borrowers and therefore assure themselves access to credit. Thakor (1982) also recognizes this interrelationship between availability and credit lines. James (1981) goes further and argues that the decision of the firm as to how the cost of the commitment will be paid, that is, fee or balances, is a self-selection process. Firms that view themselves as more likely to require future funding will opt for fees over the freezing of needed balances.

7. CREDIT RATIONING MODELS

In traditional neoclassical theory of the firm, the lack of available credit is not a concern. According to the standard paradigm, demand equals supply at the equilibrium price. Yet, it has long been argued that nonprice rationing of credit is an integral part of the market. Although data on this issue is sketchy at best, many have attempted to develop models of the credit market and the banking firm that result in some nonprice rationing of credit demands. In fact, this has been somewhat of a growth area of the literature over the last decade. Accordingly, before ending a review of the banking field, an overview of the past and recent literature on credit rationing is appropriate.

Before entering a discussion of the models devoted to explaining the existence of credit rationing, a more exact specification of the phenomenon to be explained is in order. The traditional approach to the problem is to define a situation in which there is excess demand for credit at the "going interest rate" as a situation of credit rationing. Subsequent work by Freimer and Gordon (1965) and Stiglitz and Weiss (1981) suggests a more exact definition. Credit rationing occurs when a subset of firms seeking credit at the going rate are not granted such loans in spite of the fact that their objective characteristics are identical, or nearly so, to those firms receiving credit. This definition recognizes that some borrowers are not worthy of credit because of loan or project characteristics and are, therefore, formally rejected by the lending institution. Credit rationing situations have been further divided into cases of equilibrium rationing and dynamic rationing by Jaffee and Modigliani (1969). The definition of the former is a case where credit rationing occurs at the long-run equilibrium interest rate, whereas the latter exists in the transitory periods between such equilibriums.

A. Rationing and the Customer Relation

Models devoted to explanations of credit rationing have taken three approaches. The first, and oldest, is the literature that attributes credit rationing to the tied relationship between banks and their customers. Hodgman (1961) was the first to espouse this view, which argues that because of the intertemporal and cross-product relationship between a customer and the bank, preferential treatment is given to prime customers when credit tightening occurs. Accordingly, nonprime, small customers are rationed during periods of interest rate movement. Defining all credit rationing as dynamic and short-run, he argued that the bank was merely behaving as a multiperiod profit maximizer by favoring its best customers. In fact, standard textbooks in the field (e.g., Mason 1979) have made this central to the bank/customer relationship.

This line of analysis has some very serious problems. To make the analysis consistent, it is necessary to argue that the bank's best customers are ones that yield the bank super-normal returns in a present value sense. Therefore, when a period of credit rationing develops, they are given preferential treatment. However, it is unlikely that the bank's best and largest customers can be charged a rate that is sufficiently high to warrant such preferred treatment. In fact, Blackwell and Santomero (1982) have demonstrated that if large firms, with intertemporal demands for credit and multiperiod commitments to the banking institution, are priced correctly, they will, on the margin, be no more profitable than the smaller, less-sophisticated firm. Accordingly, if rationing becomes necessary, there is no reason to believe that they will be given additional consideration. In fact, the authors demonstrate that because of the higher elasticity of demand possessed by large customers with alternative funding options, these customers are less likely to receive credit preference during such periods of constraint in contrast to the traditional result. One is left with the feeling that this literature is of questionable worth.

B. Rationing and Partial Price Discrimination

The second approach to credit rationing argues that there is something about the transactions cost structure that leads banks to refuse credit at the going interest rate. The most well known of these papers is the work of Jaffee and Modigliani (1969). According to this approach, the bank's offer curve for credit to any specific customer or customer group is defined by the standard profit-maximizing criteria. If the bank could operate as a discriminating monopolist, credit would never be rationed and each customer would be granted a profit-maximizing rate. However, if the bank is forced to charge a nonhomogeneous set of customers the same interest rate, then credit rationing will result. Such rationing results because, within the set of nonhomogeneous borrowers, the bank can be shown to charge a rate that lies between the lowest and highest rates applicable to members of the group. For the subset of loans that should have been charged a higher rate than the optimal rate for the class, quantity rationing results. The least risky in the group and the riskless customers, on the other hand, are never rationed.

The major problem with the Jaffee-Modigliani approach to rationing is that the ultimate causes of the rationing are unsubstantiated assumptions about the loan market. Specifically, the model asserts that there is only a small set of loan prices that can be charged to the entire array of borrowers. It is, then, asserted that the bank knows the exact customer risk characteristics, even though it does not use such information to set loan prices. It can be argued that the first assumption is probably a reasonable view of the bank trying to deal with a large set of heterogeneous customers. Transaction costs presumably restrict the extent to which differential pricing is possible. It is not clear that this reasoning is consistent with the second assumption. If the bank, for economic reasons, only roughly categorizes its customers into broad groups, it is hard to accept the notion that it would find it profitable to maintain more explicit knowledge within each group. Yet such additional information is necessary to obtain the results of positive equilibrium and dynamic credit rationing.

Another problem with this approach to loan sorting is its stability. Any two banks within such a system will have, presumably, different groupings and group-specific interest rates. This would seem to imply that customers who are relatively less risky than the "average" within a group may find it desirable to shift to another bank. Their relative ranking would improve and borrowing costs decline as a result of such an action. In the limit, this type of behavior should result in adverse or unique selection and in the general absence of equilibrium credit rationing from the model. Although this may not in general be the case, there does not appear to be anything in the model setup to prohibit such a degenerate case.

C. *Rationing and Information Problems*

The third approach to credit rationing, and the one that appears most promising at the moment, is the information asymmetry or adverse selection approach. This was first suggested by Jaffee and Russell (1976), in which it was argued that expected value pricing of loan rates hides two different types of borrowers. When all loans are priced at a single rate, the bank attracts both honest and dishonest customers. The former fully anticipate repayment, whereas the latter will renege in all states where the implied cost is lower than repayment. An equilibrium competitive interest rate will be set so that the market rate incorporates the probability of default by dishonest or unlucky borrowers. They then demonstrate that variations in the loan rate from this equilibrium level could shift the relative proportion of honest borrowers so as to improve the expected profit for the financial institution. Essentially, by restricting the percentage of an investment project that is financed by the bank, the lender attracts more honest customers and a smaller loan loss experience. Barro (1976) considers a similar problem by treating the quantity of collateral as a determinant of the bank offer curve.

In a recent contribution, Stiglitz and Weiss (1981) extend this approach. Dealing again with customer groupings with low and high-risk characteristics, they demonstrate that as the interest rate charged to borrowers increases, the percentage of low quality loans may increase. They argue that without a priori knowledge of the quality of each loan applicant, the willingness to pay higher interest rates is a screening device in identifying high-risk borrowers. Therefore, the bank would prefer to charge a lower rate than to clear the loan market by discouraging the preferred borrowing group.

Taken as a whole, this credit rationing literature is diverse. On one side of the spectrum, a series of ad hoc assertions about extramarginal pricing is used to explain rationing. Next, we have an approach that dominated the literature for a long period of time, in which custom or costs mandate a small set of loan categories. These, in turn, bring about both equilibrium and dynamic rationing, but somehow do not cause the bank to subdivide its lending classes further. Finally, the recent literature on asymmetric information has been used to explain rationing. Here, equilibrium rationing exists by customers self-sorting in response to interest rate signals but not necessarily dynamic rationing. Although the last approach is clearly the most sophisticated, it also raises many questions. Yet, virtually no empirical work has been done over the past decade to answer them — or to assess the importance of this entire set of literature.

8. WHERE DO WE GO FROM HERE?

There has been much written over the past decade on the subject of bank behavior. This paper lists and discusses some two hundred contributions. To conclude my review of this literature, I offer some thoughts on the future directions of work on bank modeling.

The first point that springs from this review is that there is still much to be learned about financial institutions and their place in the economy. We seem to have converged upon a global view of the maximization process that the firm attempts. Nevertheless there is little solid work on the nature of the financial firm's product. The existence question remains vital to our understanding of these institutions and why they are capable of achieving their objectives in a relatively efficient capital market. On the asset side, more work clearly needs to be done along the imperfect information line of analysis in our explanation of the financial intermediation process. Whether it will be able to explain the evolution of banking, broadly defined, through new product innovation is an open question. On the liability side, understanding the evolution of the role and definition of money in a complex payments system appears central to the explanation of intermediation itself. Liability and transaction product innovation requires considerably more attention than it is currently receiving in the banking literature.

Within our existing view of bank behavior, the modeling of asset and liability management is fairly well formulated. But it is important to proceed further along the line of integrating the two sides of the balance sheet. Issues surrounding credit and interest rate risk management must be more adequately modeled. The recent work on the interaction of these risk factors moves us in the right direction, but much more needs to be done on risk taking, hedging, and diversification in the financial firm. Finally, economists must address the degree of competitiveness in the financial markets that surround the banking firm. In finance theory virtually no market imperfections are considered. Economists, on the other hand, have generally treated the bank as an imperfect competitor in markets that exhibit nonequilibrium pricing, price setting capability, and price discrimination. It would be worthwhile if we could reconcile these opposing views by incorporating modern finance theory in the theory of the banking firm.