7

# A General Equilibrium Asset-Pricing Approach to the Measurement of Nominal and Real Bank Output

J. Christina Wang, Susanto Basu, and John G. Fernald

In many service industries, measuring real output is a challenge, because it is difficult to measure quality-adjusted prices. In financial services, however, there is not even an agreed upon conceptual basis for measuring *nominal,* let alone *real,* output.[1] This chapter presents a dynamic, stochastic, general equilibrium (DSGE) model in which nominal and real values of bank output—and hence the price deflator—are clearly defined. We use the model to assess the inadequacy of existing national accounting measures, and we derive a theoretically preferred alternative. Our model is a general equilibrium (GE) extension of Wang's (2003a) partial equilibrium framework, and it validates Wang's proposed bank service flow measure.

The biggest challenge for measurement is that banks and other financial service providers often do not charge explicit fees for services. Instead, they charge indirectly by the spread between the interest rates they charge and pay. The *System of National Accounts, 1993* (United Nations et al. 1993; hereafter SNA93) thus recommends measuring these "financial intermediation services indirectly measured" (FISIM) using net interest, defined as "the total property income receivable by financial intermediaries minus

J. Christina Wang is a senior economist at the Federal Reserve Bank of Boston. Susanto Basu is a professor of economics at Boston College and a research associate of the National Bureau of Economic Research. John G. Fernald is vice president of macroeconomic research at the Federal Reserve Bank of San Francisco.

We thank Erwin Diewert, Dennis Fixler, Charles Hulten, Alice Nakamura, Emi Nakamura, Marshall Reinsdorf, Paul Schreyer, Jack Triplett, and Kim Zieschang for helpful discussions, and we thank Felix Momsen for data assistance. The views in this chapter are those of the authors and should not be construed as necessarily reflecting the views of the Board of Governors or anyone else affiliated with the Federal Reserve System.

[1] For a recent sample, see chapter 7 in Triplett and Bosworth (2004a), the comment on that chapter by Fixler (2004), and the authors' rejoinder.

their total interest payable, excluding the value of any property income receivable from the investment of their own funds."[2] The so-called user cost approach to banking is taken to be the theoretical basis for measuring nominal output via interest rate margins and for interpreting interest rate spreads as implicit prices for financial services.[3] As a practical matter, the SNA93 approach more or less equates nominal output from FISIM with the net interest income that flows through banks.

Net interest income from lending is the conceptual difference between interest income received and the opportunity cost of funds lent. As currently implemented (e.g., in SNA93), net interest is imputed using the difference between actual lending rates and a riskless interest rate, such as a short-term Treasury rate, which is meant to capture the opportunity cost of funds. However, Wang (2003a) shows that net interest contains not only nominal compensation for bank services but also the return due to the systematic risk of bank loans. This return is part of the opportunity cost of funds, and according to the essence of the user cost framework, it should be excluded from bank output. In modern finance theories of asset pricing, the required rate of return depends on risk. Hence, the user cost of money needs to be adjusted for risk. Wang's (2003a) key contribution is the extension of the user cost approach to a world with risk. This contrasts with the unrealistic riskless framework in the existing literature. Thus, the net interest portion of Wang's service flow measure of nominal bank output can also be characterized as total net income of the opportunity cost of funds, provided this opportunity cost is correctly adjusted for risk. (All agree that explicit fee income received is also part of nominal bank output.)

The GE model here verifies the partial equilibrium conclusion reached by Wang (2003a). As in Wang (2003a), we use a user cost framework in which banks' optimal choice of interest rates must cover the (risk-adjusted) opportunity cost of funds, as well as the cost of implicitly provided services. The primary contribution of our GE model is to endogenize the cost of funds, which Wang (2003a) takes as exogenously given by financial markets.

Like SNA93, the 2003 U.S. National Income and Product Accounts (NIPAs) benchmark revisions allocate the FISIM between borrowers and depositors using a reference rate. The NIPAs impute the nominal value of services to borrowers as the volume of interest-earning assets multiplied by the difference between the (average) lending rate and that reference rate—the user cost of funds. Likewise, it imputes nominal output of services to depositors as the volume of deposits multiplied by the difference between that reference rate and the (average) deposit rate (Fixler, Reinsdorf, and Smith 2003).

[2] SNA 1993, paragraph 6.125.

[3] See, for example, Fixler and Zeischang (1992). Important contributors to the user cost approach also include Diewert (1974), Barnett (1978), and Hancock (1985).

The challenge is to determine the appropriate reference rate or rates. The NIPAs use a basically risk-free rate for both borrower and depositor services. As in Wang (2003a), however, we show that the risk-free rate is not appropriate for borrowers. Because the cost of funds is risk dependent, the reference rate must incorporate the systematic risk of a bank's loan portfolio. Hence, the imputed value of banks' implicit borrower services *excludes* the risk premium. The premium represents capital income to bank shareholders and uninsured debt holders for bearing the systematic risk on funds used in production *outside* the bank.

A simple example shows the intuition for excluding the risk premium from bank output and illustrates the shortcomings of the NIPAs approach. Consider two otherwise identical borrowers who seek to obtain additional financing in a world with no transactions costs or informational asymmetries; one borrows in the bond market, the other borrows from a bank. The bond-financed firm's expected return equals the risk-free rate plus a risk premium. It is clear that the entire return represents the value added of the borrower.

Now consider the bank-financed firm. To keep things simple, suppose that banks hire *no* labor or capital and produce no services whatsoever. Banks are merely an accounting device that records loans (perhaps funded by bank shareholder equity) to borrowers. They are a perfect substitute for the bond market, so they charge the risk-free interest rate plus the same risk premium paid by the bond-financed firm: by construction, the risk is the same, and in equilibrium, there is no arbitrage. (Note that in equilibrium, bank shareholders are indifferent between buying the bonds or holding shares in the bank.) But NIPAs would attribute positive value added to a bank equal to the risk premium multiplied by the face value of the loan—even though, by assumption, the bank does nothing!

Conceptually, the two firms should be treated symmetrically: they are identical, apart from an arbitrary and (to the firm) irrelevant financing choice. But under NIPAs conventions, they appear to have different value added, inputs of financial services, and productivities. In contrast, the approach we recommend would treat the two firms symmetrically by excluding the risk premium from the bank's nominal financial output—a premium the borrowing firm must pay, regardless of whether it is financed by bonds or bank loans.

Thus, the national accounting measure leads to inconsistency, even in the very simplest of possible models, where banks produce no services that use real resources. Our model, as well as Wang's (2003a), shows that the conceptual inconsistency extends to realistic cases, where banks provide actual services. The NIPAs measure corresponds to the empirically irrelevant special case, when either investors are risk neutral or bank loans have no systematic risk.

Quantitatively, the potential mismeasurement under the current system is large. In 2001, commercial banks in the United States had nominal output of \$187 billion,[4] of which half was final consumption and half was intermediate services to businesses. Our theoretical results imply that the measured figures overstate true output; Wang (2003b) suggests that NIPAs banking output measures are about 20 percent too high. This figure reflects both an overestimate of lending services provided to consumers (hence an overstatement of gross domestic product [GDP]) and an overestimate of intermediate services provided to firms (which does not overstate GDP but distorts measures of industry output and productivity).

Similar considerations apply to measuring the output of financial services more generally, so the total NIPAs mismeasurement can be substantial. Furthermore, the distortions affect relative GDP measures across countries. For example, banking services account for 37 percent of Luxembourg's exports, which in turn are 150 percent of GDP. Thus, our work suggests that Luxembourg's GDP could be overstated by about 11 percent—substantial by any measure.

In addition, time variation in risk premia distorts growth rates (Wang 2003b). The distortion is particularly large during transitions such as those taking place now. For example, as banks securitize a growing fraction of their loans, they move the risk premium off their books, even if they continue to provide substantially the same real services (e.g., screening and servicing loans). Several studies find that financial services contributed importantly to the post-1995 U.S. productivity growth revival, so it is important to measure the growth as well as the level of these sectors' outputs correctly.

The model provides additional insights. First, to measure real output (and hence the banking price deflator), one wants to count just the productive activities of banks (such as those related to screening loans), not the (real) amounts of associated financial assets (i.e., the loans). Importantly, the two are not generally in fixed proportion to one another, so one cannot use the volume of loans as a proxy or indicator for the value of services. The model provides conceptual guidance on how to weight different real services in the absence of clearly attributable nominal shares in cost or revenue. Second, being dynamic, our model highlights the potential timing mismatch between when a service is performed (e.g., screening when a loan is originated) and when that service is compensated (with higher interest income over the life of the loan). Third, being stochastic, the model points out that the *expected* nominal output of monitoring (services that are performed during the lifetime of a loan, after it is originated) can be measured from ex ante interest rate spreads, but the *actual* monitoring services produced are difficult to measure from ex post revenue flows. Finally, our service-flow perspective suggests major shortcomings of the book-value output mea-

[4] Figures are from Fixler, Reinsdorf, and Smith (2003) and reflect the December 2003 comprehensive revisions.

sures that are universally used in the empirical microeconomic literature on bank efficiencies.

Our use of a DSGE model offers several advantages for studying measurement issues. First, national income accounting imposes a set of adding-up constraints that must hold in the aggregate; GE models impose the same restrictions. By applying actual national income accounting procedures to the variables generated by the model, we can ask whether and under what conditions the objects measured in the national accounts correspond to the economic concepts we want to measure.

Second, and more specific to our current project, the study of banking intrinsically concerns both goods- and asset-market interactions among different agents, which endogenously determine goods prices, quantities, and interest rates. This nexus of economic connections is naturally studied in a GE setting, which ensures the comprehensive consideration of all the key elements of an economy. For example, one needs to specify an environment in which intermediation is necessary: in the model, households cannot or will not lend directly to firms for well-specified informational reasons. We also need to specify how banks then produce real intermediation services and what determines required rates of return on bank assets.

The DSGE model endogenizes the risk premium on loans that fund business capital, as well as the required rate of return on banks' equity. Our model follows Bernanke and Gertler (1989) and Bernanke, Gertler, and Gilchrist (1999), among others, but it explicitly models the screening and monitoring technology of financial intermediaries, which these authors use to resolve asymmetric information problems in investment. The model highlights the proper measurement of bank service output in both nominal and real terms.

We abstract from some activities banks undertake (mainly transactions services to depositors), as well as from realistic complications (e.g., deposit insurance and taxes). These abstractions, which could be incorporated, are unlikely to interact in important ways with the issues we address here. For example, our approach extends naturally to valuing activities by banks other than making loans and taking deposits, such as underwriting derivatives contracts and other exotic financial instruments; we present one such example. Thus, we begin the process of bringing measurement into line with the new roles that banks play in modern economies, as discussed, for example, by Allen and Santomero (1998, 1999).

One might worry that results from our bare-bones model do not apply to the far more complex real world. But our model provides a controlled setting, where we know exactly what interactions take place and what outcomes result. Even in this relatively simple setting, current methods of measuring nominal and real bank output generate inconsistent results that can be economically substantial. It is implausible that these methods will magically succeed in the far more complex world.

The chapter has four main sections. Sections 7.1 and 7.2 present the basic setup of the model with minimal technicality to build intuition for the economic reasoning behind our conclusions. (The rigorous solution of the model is included in the appendix.) Section 7.1 solves the model with symmetric information between borrowers and lenders and uses this simple setup to show by example the flaws in existing proposals for measuring bank output. Section 7.2 introduces asymmetric information and assumes that banks and rating agencies have a technological advantage in resolving such asymmetries. We derive the correct, model-based measure of bank output in this setting, where financial institutions provide real services. Section 7.3 discusses implications of the model for measuring nominal and real financial sector output. Section 7.4 discusses extensions, and section 7.5 concludes and suggests priorities for future research and data collection.

## 7.1 The Model with Symmetric Information

### 7.1.1 Overview

Our model has three groups of agents: households, who supply labor and who ultimately own the economy's capital; entrepreneurs, who hire workers and buy capital to operate projects; and competitive financial institutions (banks and rating agencies), which resolve information problems between the owners and the final users of capital. It also has a bond market in which entrepreneurs can issue corporate debt.

First, households are the only savers in this economy and thus are the ultimate owners of all capital. Their preferences determine the risk premium on all financial assets in the economy, and their accumulated saving determines the amount of capital available for entrepreneurs to rent in a given period.

Second, entrepreneurs operate projects that produce the economy's final output. There is only one homogeneous final good, sold in a competitive market, that can be consumed or invested. Entrepreneurs' projects differ from one another because the entrepreneurs differ in their ability levels (or equivalently, in the intrinsic productivity of their projects). The technology for producing final goods in any project has constant returns to scale. Thus, without asymmetric information, the social optimum would be to give all the capital to the most efficient project. But we assume that due to asymmetric information problems, entrepreneurs face a supply curve for funds that is convex in the amount borrowed.[5] We also assume that entrepreneurs are born without wealth—they are the proverbial impoverished geniuses,

[5] Given that all entrepreneurs are borrowing without collateral, this seems quite realistic. Our specific modeling assumption is that the cost of screening is convex in the size of the project, but other assumptions—such as leveraging each entrepreneur's net worth with debt—would also lead to this result. See Bernanke, Gertler, and Gilchrist (1999).

whose heads are full of ideas but whose purses hold only air—so that one way or another, they will need to obtain funds from households.

The focus of this chapter is on how the entrepreneurs obtain the funds for investment from households and on the role of financial intermediaries in the process. A large literature on financial intermediation explains (in partial equilibrium) financial institutions' role as being to resolve informational asymmetries between the ultimate suppliers of funds (i.e., the households in our model) and the users of funds (i.e., the entrepreneurs who borrow to buy capital and produce). We incorporate this result into our general equilibrium model.[6]

We consider both types of information asymmetry—hidden information and hidden actions. Households face adverse selection ex ante as they try to select projects to finance: they know less about the projects (e.g., default probabilities under various economic conditions) than the entrepreneurs, who have an incentive to understate the risk of their projects. Moral hazard arises ex post, as savers cannot perfectly observe borrowers' actions (e.g., diverting project revenue for their own consumption).

Thus, the third group of actors in our model are institutions such as banks that exist (in the model and largely in practice) to mitigate these informational problems.[7] We focus on two specific services they provide: (a) screening to lessen (in our model, to eliminate) entrepreneurs' private information about the viability of their projects and (b) monitoring project outcomes (e.g., auditing after a default) to discover entrepreneurs' hidden actions.[8] To conduct screening and monitoring, intermediaries engage in a production process that uses real resources of labor, capital, and an underlying technology. The production process is qualitatively similar to producing other information services, such as consulting and data processing.[9]

We call the financial intermediaries banks mainly for convenience, although the functions they perform have traditionally been central to the activities of commercial banks. But the analysis is general, as we will show

[6] Most general equilibrium models of growth or business cycles abstract from this issue: implicitly, households own and operate the firms directly, so there are no principal-agent problems.

[7] Financial institutions prevent market breakdown (such as in Akerlof [1970]) but cannot eliminate deadweight loss. Another major function of banks is to provide services to depositors, as discussed in the introduction. But we omit them from the formal model, because their measurement is less controversial and has no bearing on our conclusion about how to treat risk in measuring lending services. Yet, we note practical measurement issues about them in section 7.3.

[8] Many studies, all partial equilibrium, analyze the nature and operation of financial intermediaries. See, for example, Leland and Pyle (1977), who model banks' role as resolving ex ante adverse selection in lending; Diamond (1984) studies delegated monitoring through banks; Ramakrishnan and Thakor (1984) look at nondepository institutions.

[9] Only a handful of studies analyze the effects of financial intermediaries on real activities in a general equilibrium framework. None of them, however, consider explicitly the issue of financial intermediaries' output associated with the process of screening and monitoring, nor the properties of the screening and monitoring technology.

that loans subject to default are equivalent to a risk-free bond plus a put option. So, our analysis also applies to implicit bank services associated with other financial instruments, as well as to other types of intermediaries, such as rating agencies and finance companies. We assume that banks and other financial service providers are owned by households and are not subject to informational asymmetries with respect to households.[10]

As suppliers of funds, households demand an expected rate of return, commensurate with the systematic risk of their assets. This is of course true in any reasonable model with investor risk aversion, regardless of whether there are informational asymmetries. Banks must ensure that the interest rate charged compensates their owners, the households, with the risk-adjusted return in expectation. Banks must also ensure that they charge explicit or implicit fees to cover the costs incurred by screening and monitoring.

The primary focus of this chapter is on how to correctly measure the nominal and real service output provided by these banks, when the services are not charged explicitly, but rather are charged implicitly in the form of higher interest rates. Hence, we need to detail the nature of the contract between entrepreneurs and banks, because that determines the interest rates banks charge. Indeed, most of the complexity in the formal model in the appendix comes from the difficulty of solving for the interest rate charged under the optimal debt contract and from decomposing total interest income into a compensation for bank services—screening and monitoring—and a risk-adjusted return for the capital that households channel to firms through the bank. The payoff from this complexity is that the model provides definite insights on key measurement issues.

For the most part, we try to specify the incentives and preferences of the three groups of agents in a simple way in order to focus on the complex interactions among the agents. We now summarize the key elements of the incentives and preferences of each agent to give the reader a working knowledge of the economic environment. We then derive the key first-order conditions for the optimal pricing of risky assets, which must hold in any equilibrium, to draw implications from the model that are crucial for measurement purposes. At the end of this section, the reader may proceed to the detailed discussion of the model found in the appendix or may proceed to section 7.3 to study the implications for measurement.

### 7.1.2 Households

We assume households are infinitely lived and risk averse. For most of the chapter, we assume that households can invest their wealth only through a financial intermediary, because they lack the ability to resolve information asymmetries with entrepreneurs directly. In contrast, households own

[10] We could extend our model to allow for this two-tier information asymmetry at the cost of considerable added complexity. We conjecture, however, that our qualitative results would be unaffected by this change.

and have no informational problems with respect to the intermediaries. All households are identical, and they maximize the expected present value of lifetime utility—expressed here in terms of a representative household:

$$E_t\left[\sum_{s=0}^{\infty}\rho^s V(C_{t+s}^H, 1-N)\right], \tag{1}$$

subject to the budget constraint:

$$C_t^H = W_t N_t + \prod_t + \tilde{R}_{t+1}^H X_t - X_{t+1}. \tag{2}$$

The variable $C_t^H$ is the household's consumption, $N_t$ is its labor supply, and ρ is the discount factor. The variable $E_t(.)$ is the expectation, given the information set at time $t$. We assume that the utility function $V(.)$ is concave and that $V'(0) = \infty$. The variable $W_t$ is the wage rate, $X_t$ is the household's total assets (equal to the capital stock in equilibrium), and $\Pi_t$ is pure economic profit received from ownership of financial intermediaries (equal to zero in equilibrium because we assume that this sector is competitive). The variable $\tilde{R}_{t+1}^H$ is the ex post gross return on the household's asset portfolio (real capital, lent to various agents to enable production in the economy). Corresponding to the ex post return is an expected return—the required rate of return on risky assets, which we denote $R_{t+1}^H$. This is a key interest rate in the following sections, so we discuss it further.

We define the intertemporal pricing kernel (also called the stochastic discount factor), $m_{t+1}$, as

$$m_{t+1} \equiv \frac{\rho V_c(C_{t+1}^H, 1 - N_{t+1})}{V_c(C_t^H, 1 - N_t)}, \tag{3}$$

where $V_C$ is the partial derivative of utility with respect to consumption. In this notation, the Euler equation for consumption (which is also a basic asset-pricing equation in the consumption capital asset-pricing model [CCAPM]) is:

$$E_t(m_{t+1}\tilde{R}_{t+1}^H) = 1. \tag{4}$$

Now suppose a one-period asset, whose return is risk free because it is known in advance. Clearly, for this asset, the rate of return $R_{t+1}^f$ satisfies $E_t(m_{t+1}R_{t+1}^f) = R_{t+1}^f E_t(m_{t+1}) = 1$. So,

$$R_{t+1}^f = \frac{1}{E_t(m_{t+1})}. \tag{5}$$

As is standard in a CCAPM, the Euler equation (3) allows us to derive the risk-free rate, even if no such asset exists—which is the case in our economy, where the only asset is risky capital.[11]

[11] For more discussion, see chapter 2 in Cochrane (2001).

From equations (4) and (5), the gross required (expected) rate of return on the risky asset, $R^H_{t+1}$, is:

$$(6) \qquad R^H_{t+1} \equiv E_t(\tilde{R}^H_{t+1}) = R^f_{t+1}[1 - \text{cov}_t(m_{t+1}, \tilde{R}^H_{t+1})],$$

where $\text{cov}_t$ is the covariance, conditional on the information set at time $t$. The risk premium then equals

$$R^H_{t+1} - R^f_{t+1} = -R^f_{t+1}\,\text{cov}_t(m_{t+1}, \tilde{R}^H_{t+1}).$$

Note that when $R^H_{t+1}$ is the *required* rate of return on debt (e.g., loans) subject to the risk of borrower default, there is a subtle but important conceptual difference between $R^H_{t+1}$ and the interest rate (the so-called yield for corporate bonds) that is *charged* on loans—the rate that a borrower must pay if not in default. To illustrate in a simple example, suppose there is probability $p$ that a borrower will pay the interest rate charged (call it $R_{t+1}$) and probability $(1 - p)$ otherwise, in which case lenders get nothing. Then, $R_{t+1}$ must satisfy

$$p \cdot R_{t+1} + (1 - p) \cdot 0 = R^H_{t+1} \quad \Rightarrow \quad R_{t+1} = \frac{R^H_{t+1}}{p}.$$

So, $R_{t+1}$ exceeds the required return $R^H_{t+1}$; the margin $R_{t+1} - R^H_{t+1}$ is the so-called default premium. Thus, $R_{t+1}$ differs from the risk-free rate for two reasons. First, there is the default premium. The borrower repays less to nothing in bad states of the world, so he must pay more in good states to ensure an adequate average return. Second, there is a risk premium, as previously shown. The risk premium exists if the probability of default is correlated with consumption (or more precisely, with the marginal utility of consumption). If defaults occur when consumption is already low, then they are particularly costly in utility terms. Thus, the consumer requires an extra return, on average, to compensate for bearing this systematic, nondiversifiable risk.

In addition to the intertemporal Euler equation, consumer optimization requires a static trade-off between consumption and leisure within a period:

$$(7) \qquad W_t V_C(C^H_t, 1 - N_t) = -V_N(C^H_t, 1 - N_t).$$

In equilibrium, households' assets equal the total capital stock of the economy: $X_t = K_t$. The capital stock evolves in the usual way:

$$K_{t+1} = (1 - \delta)K_t + I_t.$$

Capital is used by intermediaries to produce real financial services or is rented by firms for production.[12]

[12] Because we have assumed identical households, we abstract from lending among households (e.g., home mortgages).

### 7.1.3 Entrepreneurs

Each entrepreneur owns and manages a nonfinancial firm that invests in one project, producing the single homogeneous final good and selling it in a perfectly competitive market. So entrepreneur, firm, and project are all equivalent and interchangeable in this model. Entrepreneurs are a set of agents, distinct from households in that each lives for only two periods, which coincides with the duration of a project. Thus, there are two overlapping generations of entrepreneurs in each period. The number of entrepreneurs who are born and die each period is constant, so the fraction of entrepreneurs is constant in the total population of agents. The reason for having short-lived entrepreneurs in the economy is to create a need for external financing and thus the screening and monitoring by financial intermediaries. Long-lived entrepreneurs could accumulate enough assets to self-finance all investment, without borrowing from households. In addition, by having each borrower interact with lenders only once, we avoid complex supergame Nash equilibria, where entrepreneurs try to develop a reputation for being good risks in order to obtain better terms from lenders.

We assume that entrepreneurs, like households, are risk averse.[13] But we abstract from the issue of risk sharing and assume that the sole income an entrepreneur receives is the residual project return, if any, net of debt repayment.[14] That also means entrepreneurs have no initial endowment.[15] In choosing project size in the first period, entrepreneurs seek to maximize their expected utility from consumption in the second period, which is the only period in which they consume. Thus, the utility of entrepreneur $i$ born at time $t$ is

$$U(C_{t+1}^{E,i}), \tag{8}$$

where $U' > 0$, $U'' < 0$, and $U(0) = 0$. We denote entrepreneurs' aggregate consumption by $C_t^E$, which is the sum over $i$ of $C_t^{E,i}$.

Firms differ only in their exogenous technology parameters. We denote

[13] If entrepreneurs were risk neutral, they would insure the households against all aggregate shocks, leading to a degenerate—and counterfactual—outcome, where lenders of funds would face no aggregate risk.

[14] In fact, this model implicitly allows for the sharing of project-specific risk (i.e., $z^i$) across entrepreneurs (e.g., through a mutual insurance contract covering all entrepreneurs), as all the results would remain qualitatively the same. The model assumes that there is no risk sharing between entrepreneurs and households, because the only contract that lenders offer borrowers is a standard debt contract. Given our desire to study banks, this assumption is realistic.

[15] The assumption of zero endowment is mainly to simplify the analysis. Introducing partial internal funds (e.g., with entrepreneurs' own labor income) affects none of the model's conclusions. One potential problem with zero internal funds is that it gives entrepreneurs incentive to take excessive risk (i.e., adopting projects with a high payoff when successful but possibly a negative net present value), but we rule out such cases by assumption. The usual principal-agent problem between shareholders and managers does not arise here because entrepreneurs are the owners-operators.

the parameter $A^i_{t+1}$ for a firm $i$ created in period $t$, because the owner produces in the second period—$t + 1$. We assume that $A^i_{t+1} = z^i A_{t+1}$, where $A_{t+1}$ is the stochastic aggregate technology level in period $t + 1$, and $z^i$ is the idiosyncratic productivity level of $i$, drawn at time $t$ when the owner is born. The variable $z^i$ is assumed to be independently and identically distributed across firms and time, with bounded support, and independent of $A_{t+1}$, with $E(z^i) = 1$. Conditional on $z^i$, the firm borrows to buy capital from the households at the end of period $t$. In keeping with our desire to study banking operation in detail, we assume that lenders offer borrowers a standard debt contract. (We discuss the borrowing process, first under symmetric information and then under asymmetric information, in the next several subsections.)

The aggregate technology level ($A_{t+1}$) is revealed at the start of period $t + 1$, and it determines $A^i_{t+1}(= z^i A_{t+1})$. But because $A_{t+1}$ is unknown when the capital purchase decision is made, there is a risk involved for both the borrower and the lender. Conditional on $A^i_{t+1}$ and the *precommitted* level of capital input, the firm hires the optimal amount of labor at the going wage of time ($t + 1$) and produces the final good. Entrepreneurs first pay their workers, then pay the agreed upon interest to households (as well as return the loan principal, the value of the stock of capital rented for production), and then consume all the output leftover.

If a bad realization of $A_{t+1}$ leaves an entrepreneur unable to cover the gross interest on his borrowed funds, he declares bankruptcy. The lenders (households) seize all of the assets and output of the firm leftover after paying the workers, which will be shown to be less than what the lenders are owed and expect to consume. Entrepreneurs are left with zero consumption—less than what they expected as well. The risk to both borrowers and lenders is driven by the aggregate uncertainty of the stochastic technology, $A$.

### 7.1.4 Equilibrium with Symmetric Information

In order to make an important point about the SNA93 method for measuring nominal bank output, we first consider a case where households can costlessly observe all firms' idiosyncratic productivity, $z^i$.

We assume that the production function of each potential project has constant returns to scale (CRS):

$$(9) \qquad Y^i_t = A_t z^i (K^i_t)^{\alpha} (N^i_t)^{1-\alpha}.$$

Given CRS production, households will want to lend all their capital only to the entrepreneur with the highest level of $z$—or to paraphrase in market terms, the entrepreneur with the highest productivity will be willing and able to outbid all the others and to hire all the capital in the economy. (We assume that he or she will act competitively, taking prices as given, rather than act as a monopolist or monopsonist.)

We define $\bar{z} = \max_i\{z_i\}$.[16] Then, the economy's aggregate production function will be (that of $\bar{z}$):

$$Y_t = A_t \bar{z} K_t^{\alpha} N_t^{1-\alpha}.$$

The entrepreneur with the $\bar{z}$ level of productivity will hire capital at time $t$ to maximize

$$E_t U\{\max[A_{t+1}\bar{z}K_{t+1}^{\alpha}N_{t+1}^{1-\alpha} - (R_{t+1} - 1 + \delta)K_{t+1} - W_{t+1}N_{t+1}, 0]\}. \tag{10}$$

The expression in equation (10) indicates that the entrepreneur gets either the residual profits from his project if he is not bankrupt or gets nothing if he has to declare bankruptcy.

The labor choice will be based on the realization of $A_{t+1}$ and the market wage and will be

$$N_{t+1} = \left[\frac{(1-\alpha)\bar{z}A_{t+1}}{W_{t+1}}\right]^{1/\alpha} K_{t+1}. \tag{11}$$

Production, capital and labor payments, and consumption will take place as outlined in the previous subsection. Note that producing at the highest available level of $z$ does not mean that bankruptcy will never take place or even that it will necessarily be less likely. Ceteris paribus, a higher expected productivity of capital raises the expected return $R_{t+1}^H$ but does not eliminate the possibility of bankruptcy conditional on that higher-required return.[17] Thus, debt will continue to carry a risk premium relative to the risk-free rate.

The national income accounts identity in this economy is

$$Y_t = C_t^H + C_t^E + I_t.$$

### 7.1.5 The Bank That Does Nothing

There is no bank in the economy summarized in the previous subsection, nor is there any need for one. Households lend directly to firms at a required rate of return $R_{t+1}^H$. Suppose, however, a bank is formed simply as an accounting device. Households transfer their capital stock to banks, and in return, they own bank equity. The bank rents the capital to the single most productive firm at the competitive market price.

Because households see through the veil of the bank to the underlying

[16] The maximum is finite because we have assumed the $z$ has a bounded support.

[17] Let us assume, as in section 7.2, that a continuum of entrepreneurs is born every period, so we are guaranteed that $\bar{z}$ is always the upper end of the support of $z_i$. Then, all that happens by choosing the most productive firm every period is that the mean level of technology is higher than if we chose any other firm (e.g., the average firm). But nothing in our derivations turns on the mean of $A$; it is simply a scaling factor for the overall size of the economy, which is irrelevant for considering the probability of bankruptcy.

assets the bank holds—risky debt issued by the entrepreneur—they will demand the same return (i.e., $R^H_{t+1}$) on bank equity as they did on the debt in the economy without a bank. Because the bank acts competitively (and thus makes zero profit), it will lend the funds at marginal cost (expected return of $R^H_{t+1}$, hence a contractual interest rate of $R_{t+1}$) to the firm, which will then face the same cost of capital as before.

However, applying the SNA93 calculation for FISIM to this model economy, the value added of bank, effected via book entries of the capital transfer (the only sign of the bank's existence here), would be

$$(R^H_t - R^f_t)K_t.$$

The variable $K_t$ is the value of bank assets, as well as the economy-wide capital stock. Thus, by using the risk-free rate as the opportunity cost of funds instead of the correct risk-adjusted interest rate, the current procedure attributes positive value added to the bank that in fact produces nothing.[18]

At the same time, from the expenditure side, the value of national income will be unchanged—still equal to $Y_t$—because the bank output (if any) is used as an intermediate input of service by firms producing the final good.[19] But industry values added are mismeasured: for a given aggregate output, the productive sector has to have lower value added in order to offset the value added incorrectly attributed to the banking industry. Clearly, the production sector's true value added is all of $Y_t$, but it will be measured, incorrectly, as:

$$Y_t - (R^H_t - R^f_t)K_t.$$

Thus, the general lesson from this example is that whenever banks make loans that incur aggregate risk (i.e., risk that cannot be diversified away), the current national accounting approach attributes too much of aggregate value added to the banking industry and too little to the firms that borrow from banks. This basic insight carries over to the more realistic cases next, where banks do in fact produce real services.

We shall also argue later that our simplifying assumption of a fully equity-funded bank is completely unessential to the result. The reason is that in our setting, the theorem of Modigliani and Miller (1958; hereafter MM) applies to banks. The MM theorem proves that a firm's cost of capital is independent of its capital structure. Thus, the bank that does nothing can finance itself by issuing debt (taking deposits) as well as equity, without

[18] Financial intermediation services indirectly measured (FISIM) also impute a second piece of bank output—depositor services. But because bank deposits are zero in our model, FISIM would correctly calculate this component of output to be zero.

[19] Mismeasuring banking output would distort GDP if banks' output is used as a final good (e.g., lending and depository services to consumers, or perhaps more importantly, net exports).

changing the previous result in the slightest, either qualitatively or quantitatively.[20]

Even in more realistic settings, the lesson in this subsection is directly relevant for one issue in the measurement of bank output. Banks buy and passively hold risky market assets, as in the example here. Even though banks typically hold assets with relatively low risk, such assets (e.g., high-grade corporate bonds) still offer rates of return that exceed the risk-free rate, sometimes by a nontrivial margin. Whenever a bank holds market securities that offer an average return higher than the current reference rate, it creates a cash flow—the difference between the securities' return and the safe return, multiplied by the market value of the securities held—that the current procedure improperly classifies as bank output.

## 7.2 Asymmetric Information and a Financial Sector That Produces Real Services

### 7.2.1 Resolving Asymmetric Information I: Nonbank Financial Institutions

Now we assume, more realistically, that information is in fact asymmetric. Entrepreneurs know their idiosyncratic productivity and actual output, but households cannot observe them directly. In this case, as we know from Akerlof (1970), the financial market will become less efficient and may break down altogether.

We introduce two new institutions into our model. The first is a rating agency, which screens potential borrowers and monitors those who default to alleviate the asymmetric information problems. The other is a bond market; that is, a portfolio of corporate debt. The two combined fulfill the function of channeling funds from households to entrepreneurs so that the latter can invest. Both institutions have real-world counterparts, which will be important when we turn to our model's implications for output measurement.

The purpose of introducing these two new institutions will become clear in the next subsection when we compare them with banks. There, we will show that a bank can be decomposed into a rating agency plus a portfolio of corporate debt, and the real output of banks—informational services—is equivalent to the output of the agency alone. Thus, it makes sense to understand the two pieces individually before studying the sum of the two. Understanding the determination of bond market interest rates is particularly

[20] This is assuming there is no deposit insurance. See Wang (2003a) for a full treatment of banks' capital structure with risk and deposit insurance. Of course, in the real world, taxes and transactions costs break the pure irrelevance result of Modigliani-Miller. But the basic lesson—that the reference rate must take risk into account—is unaffected by these realistic but extraneous considerations.

important when we discuss measurement, because we shall argue that corporate debt with the same risk-return characteristics as bank loans provides the appropriate risk-adjusted reference rate for measuring bank output.

We discuss rating agencies first. These are institutions with specialized technology for assessing the quality (i.e., productivity) of prospective projects, and they are also able to assess the value of assets if a firm goes bankrupt. Thus, these institutions are similar to the rating agencies found in the real world, such as Moody's and Standard and Poor's, which not only rate new issues of corporate bonds but also monitor old issues.

The technology of each rating agency for screening ($S$) and monitoring ($M$) is as follows:

$$Y_t^{JA} = A_t^J (K_t^{JA})^{\beta^J} (N_t^{JA})^{1-\beta^J},\ J = M \text{ or } S. \tag{12}$$

We use the superscript $A$ to denote prices and output of the agency. The variables $K_t^{JA}$ and $N_t^{JA}$ are the capital and labor, respectively, used in the two activities. The variables $A_t^M$ and $A_t^S$ differ when the pace of technological progress differs between the two activities. Difference between output elasticities of capital $\beta^M$ and $\beta^S$ means that neither kind of task can be accomplished by simply scaling the production process of the other task.

We assume there are many agencies in a competitive market, so the price of their services equals the marginal cost of production. The representative rating agency solves the following value maximization problem:

$$E_0\left[\sum_{t=0}^{\infty}\left(\prod_{t=0}^{t} R_t^{SV}\right)^{-1}(f_s^{SA}Y_t^{SA} + f_t^{MA}Y_t^{MA} - W_t N_t^A - I_t^A)\right], \tag{13}$$

$$Y_t^{SA} = A_t^S (K_t^{SA})^{\beta^S} (N_t^{SA})^{1-\beta^S}, \tag{14}$$

$$Y_t^{MA} = A_t^M (K_t^{MA})^{\beta^M} (N_t^{MA})^{1-\beta^M}, \text{ and } Y_0^{MA} = 0, \tag{15}$$

$$N_t^A = N_t^{SA} + N_t^{MA}, \text{ and } K_t^A = K_t^{SA} + K_t^{MA}, \tag{16}$$

$$K_{t+1}^A = K_t^A(1 - \delta) + I_t^A. \tag{17}$$

In equation (13), $Y_t^{SA}$ and $Y_t^{MA}$ are the rating agency's respective output of screening and monitoring services. The variables $f_t^S$ and $f_t^M$ are the corresponding prices (mnemonic: *f*ees), and as assumed, are equal to the respective marginal cost. The variable $W_t$ is the real wage rate, and $N_t^A$ is the agency's total labor input. Equations (14) and (15) are the production functions for screening and monitoring, respectively, with the inputs defined as in equation (12). Total labor and capital inputs are given in equation (16), and equation (17) describes the law of motion for the agency's total capital.

The agency is fully equity funded. Thus, the discount rate for the agency's value maximization problem (i.e., $R_t^{SV}$ [$SV$ standing for services]) is exactly its shareholders' required rate of return on equity. The variable $R_t^{SV}$, analogous to $R_{t+1}^H$ in equation (6), thus is determined by the systematic risk of the agency's entire cash flow. According to the pricing equation (6), and equivalently equation (4), $R_t^{SV}$ equals

$$(18)\quad R_{t+1}^{SV} = R_{t+1}^{f}\left\{1 - \operatorname{cov}_t\left[m_{t+1}, \frac{f_{t+1}^{SA}Y_{t+1}^{SA} + f_{t+1}^{MA}Y_{t+1}^{MA} - W_{t+1}N_{t+1}^{A} + (1-\delta)K_{t+1}^{A}}{K_{t+1}^{A}}\right]\right\}.$$

The denominator ($K_{t+1}^A$) in the covariance is the agency's capital used in production at time $t + 1$, funded by its shareholders at time $t$. The numerator is the ex post return on that capital, consisting of its operating profits (revenue minus labor costs), plus the return of the depreciated capital lent by the stockholders at time $t$.[21]

Even though the agency is paid contemporaneously for its services, the fact that it must choose its capital stock a period in advance creates uncertainty about the cash flow accruing to the owners of its capital. This uncertainty arises fundamentally because the demand for screening and monitoring is random, driven by the stochastic process for aggregate technology, $A_{t+1}$. Thus, the implicit rental rate of physical capital in period $t$ for this agency is $(R_t^{SV} - 1 + \delta)$,[22] where $R_t^{SV}$ will generally differ from the risk-free rate.

Because a rating agency is of little use unless one can borrow on the basis of a favorable rating, we assume that a firm can issue bonds of the appropriate interest rate in the bond market once it is rated. That is, once an agency finishes screening a firm's project, it issues a certificate that reveals the project's type (i.e., $z^i$). Armed with this certificate, firms sell bonds to households in the market, offering contractual rates of interest $R_{t+1}^i$ that vary according to each firm's risk rating. The variable $R_{t+1}^i$ depends on households' required rate of return on risky debt, but $R_{t+1}^i$ is not the required return per se. The two differ by the default premium, as discussed in subsection 7.1.2. (Determining the appropriate interest rate to charge an entrepreneur of type $i$ is a complex calculation, in part because the probability of default is endogenous to the interest rate charged. We thus defer this derivation to the appendix.)

There is an additional complication: because entrepreneurs are born without wealth, they are unable to pay their screening fees up front. Instead, they must borrow the fee from the bond market, in addition to the capital they plan to use for production next period, and must dash back to the rating agency within the period to pay the fee they owe. In the second period, they must pay the bondholders a gross return on the borrowed productive capital, plus a same return on the fee that was borrowed to pay the agency.

[21] The payoff to the shareholder depends, of course, on the marginal product of capital. The assumption of constant-returns, Cobb-Douglas production functions allows us to express the result in terms of the more intuitive average return to capital. Note that the capital return in equation (18) is actually an average of the marginal revenue products of capital in screening and monitoring, with the weights being the share of capital devoted to each activity.

[22] Recall that all $R$ variables are gross interest rates, so the net interest rate $r = R - 1$.

In the second period of his or her life, after his or her productivity is determined by the realization of $A_{t+1}$, an entrepreneur may approach his or her bondholders and inform them that his or her project was unproductive and that he or she is unable to repay his or her debt with interest. The households cannot assess the validity of this claim directly. Instead, they must engage the services of the rating agency to value the firm (its output plus residual capital). The agency charges a fee equal to its marginal cost, as determined by the maximization problem in equations (13) through (17). We assume that the agency can assess the value of the firm perfectly. Whenever a rating agency's services are engaged, the bondholders get to keep the entire value of the project after paying the agency its monitoring fee.[23] The entrepreneur gets nothing. Under these circumstances, the entrepreneur always tells the truth and only claims to be bankrupt when that is in fact the case.

Note that in this asymmetric information environment, entrepreneurs require additional inputs of real financial services from the agencies to obtain capital. The production function for gross output for a firm of type *i* is still given by equation (9). But now, entrepreneurs have two additional costs. In the first period, when they borrow capital, they must buy certain units of certification services. The amount of screening varies with the size of the project (see the appendix for a detailed discussion of the size dependence of these information processing costs). A project of size $K^i_{t+1}$ needs $\upsilon^S_t(K^i_{t+1})$ units of screening services. Then, in the second period, a firm is required to pay for $Z^M_{t+1}\upsilon^M_t(K^i_{t+1})$ units of monitoring services, where $Z^M$ equals one if the firm defaults and equals zero otherwise. Functions $\upsilon^S(.)$ and $\upsilon^M(.)$ determine how many units of screening, and possibly monitoring, are needed for a project of size $K^i$. Either $\upsilon^S(.)$ or $\upsilon^M(.)$ is strictly convex, and this effectively leads firms to have diminishing returns to scale.[24] Thus, it is no longer optimal to put all the capital at the most productive firm, and the equilibrium involves production by a strictly positive measure of firms.

Given these two additional costs, firm *i* producing in period $t + 1$ maximizes

$$E_t U\{\max[A_{t+1}z^i(K^i_{t+1})^{\alpha}(N^i_{t+1})^{1-\alpha} - (R^i_{t+1} + \delta)K^i_{t+1} - W_{t+1}N^i_{t+1} \\ - \upsilon^S_t(K^i_{t+1}) - Z^M_{t+1}\upsilon^M_t(K^i_{t+1}), 0]\}.$$

The variable $R^i_{t+1}$ is the contractual interest rate appropriate for a project of type *i*—the analogue to the full information contractual rate for the highest productivity project in equation (10). As in the situation with perfect information, either the entrepreneur gets positive residual profits, or he or she declares bankruptcy and gets nothing.

[23] We assume that a project always has a *gross* return large enough to pay the fee. This assumption seems reasonable—even Enron's bankruptcy value was high enough to pay similar costs (amounting to over one billion dollars).

[24] A convex cost of capital is needed to obtain finite optimal project scale; we discuss this issue further in the appendix.

We define $\tilde{R}^{Ki}_{t+1}$ as the ex post gross return on capital for the project. It is the project's total output net of labor cost and depreciation, $(\tilde{R}^{Ki}_{t+1} - 1)K^i_{t+1} = Y^i_{t+1} - W_{t+1}N^{i*}_{t+1} - \delta K^i_{t+1}$, where $N^{i*}_{t+1}$ is the optimal quantity of labor. Thus, the ex ante required rate of return on the bonds issued by firm *i*, $R^{Li}_{t+1}$, is the required return implied by the asset-pricing equation

(19)

$$E_t\left\{m_{t+1}\cdot\frac{R^i_{t+1}[K^i_{t+1}+f^S_t\upsilon^S(K^i_{t+1})](1-Z^{Mi}_{t+1})+[\tilde{R}^{Ki}_{t+1}K^i_{t+1}-f^M_{t+1}\upsilon^M(K^i_{t+1})]Z^{Mi}_{t+1}}{[K^i_{t+1}+f^S_t\upsilon^S(K^i_{t+1})]}\right\}=1.$$

So, as usual, $R^{Li}_{t+1}$ depends on the conditional covariance between the cash flow and the stochastic discount factor. The expression in the numerator of the fraction is the state-contingent payoff to bondholders. If the realization of technology $(A_{t+1})$ is sufficiently favorable, then the project will not default (i.e., $Z^M = 0$), and the bondholders will receive the contractual interest promised by the bond—$R^i_{t+1}[K^i_{t+1} + f^S_t\upsilon^S(K^i_{t+1})]$. Otherwise, if the realization of technology is bad enough, the firm will have to declare bankruptcy, and bondholders will receive the full value of the firm, net of the monitoring cost—$[\tilde{R}^{Ki}_{t+1}K^i_{t+1} - f^M_{t+1}\upsilon^M(K^i_{t+1})]$. The contracted interest rate on the bond issued by a project $(R^i_{t+1})$ depends on its ex ante required rate of return $R^{Li}_{t+1}$, which in turn depends on the risk characteristics of that project. For details, see the appendix.

The denominator of equation (19) is the total amount of resources the firm borrows from households. The variable $K^i_{t+1}$ is the capital used for production, while $f^S_t\upsilon^S(K^i_{t+1})$ is the screening fee. As discussed previously, entrepreneurs need to borrow to pay the screening fees, because they have no endowments in the first period of their lives.

In general, households will hold a portfolio of bonds, not just one. For comparison in the next subsection with the case of a bank, it will be useful to derive the required return on this portfolio. Because each bond return must satisfy equation (19), we can write the return to the portfolio as a weighted average of the individual returns. Then, for a large portfolio of infinitesimal projects, the required rate of return is set by the equation

(20)

$$E_t\left(m_{t+1}\cdot\frac{\int_{i:K^i_{t+1}>0}\{R^i_{t+1}[K^i_{t+1}+f^s_t\upsilon^s(K^i_{t+1})](1-Z^{Mi}_{t+1})+[\tilde{R}^{Ki}_{t+1}K^i_{t+1}-f^M_{t+1}\upsilon(K^i_{t+1})]Z^{Mi}_{t+1}\}}{\int_{i:K^i_{t+1}>0}[K^i_{t+1}+f^S_t\upsilon^S(K^i_{t+1})]}\right)=1.$$

where the integral is taken over all firms whose bonds are in the investor's portfolio.[25]

[25] To illustrate the derivation, consider an example of discreet projects. Suppose a lender holds bonds from *N* firms. Equation (19) holds for every firm *i* and can be rearranged by pulling the denominator $K^i_{t+1} + f^S_t\upsilon^S(K^i_{t+1})$ outside the expectations sign, because it is known at time *t*. Then, multiply each firm's equation (19) by the firm's share in the aggregate resources borrowed (i.e., $[K^i_{t+1} + f^S_t\upsilon^S(K^i_{t+1})]/\sum^N_{i=1}[K^i_{t+1} + f^S_t\upsilon(K^i_{t+1})]$), and add up the *N* resulting equations.

### 7.2.2 Resolving Asymmetric Information II: Banks That Produce Real Services

We are finally ready to discuss bank operations. Now the banking sector performs real services, unlike the accounting device in subsection 7.1.5. We assume that banks assess the credit risk of prospective borrowers and lend them capital, and if a borrower claims to be unable to repay, banks investigate, liquidate the assets, and keep the proceeds. That is, in our model—and in the world—banks perform the functions of rating agencies and the bond market under one roof. As important, especially for measurement purpose, note that banks, rating agencies, and the bond market all coexist, both in the model and in reality. Our banks are completely equity funded.[26] They issue stocks in exchange for households' capital. Part of the capital is used to generate screening and monitoring services, with exactly the same technology as in equation (12). The rest of the capital is lent to qualified entrepreneurs. At time $t$, a bank must make an ex ante decision to split its total available capital into in-house capital (used by the bank for producing services in period $t + 1$, denoted $K^B_{t+1}$) and loanable capital (lent to entrepreneurs, used to produce the final good in period $t + 1$). Because the banking sector is competitive, banks price their package of services at marginal cost.

The exact statement of the bank's value maximization problem is tedious and yields little additional insight, so it, too, is deferred to the appendix. In summary, entrepreneurs are shown to be indifferent between approaching the bank for funds or going to a rating agency and then to the bond market,[27] given that banks have the same screening and monitoring technology as the agency (production functions in equations [14] and [15]).

Instead, in the rest of this section, we illustrate the intuition of the model's conclusion—a bank's cash flow equivalent to that of a rating agency plus a bond portfolio—and its implication for bank output measure.

First, we describe a bank's total cash flow. At any time $t$, banks cannot charge explicit fees for the service of screening young entrepreneurs' applications for funds, because the applicants have no initial wealth. Instead, banks have to allow the fees to be paid in the next period and must obtain additional equity in the current period to finance the production costs of

The right-hand side clearly sums up to one, while $\sum_{i=1}^{N}[K^i_{t+1} + f^S_t\upsilon(K^i_{t+1})]$ becomes the common denominator for the left-hand side. Consequently, we find that $E_t(m_{t+1} \cdot \sum_{i=1}^{N}\{R^i_{t+1}[K^i_{t+1} + f^S_t\upsilon^S(K^i_{t+1})](1 - Z^{Mi}_{t+1}) + [\tilde{R}^{Ki}_{t+1}K^i_{t+1} - f^M_{t+1}\upsilon^M(K^i_{t+1})]Z^{Mi}_{t+1}\}/\sum_{i=1}^{N} [K^i_{t+1} + f^S_t\upsilon(K^i_{t+1})]) = 1$. That is, the weighted average of the $N$ firms' conditions equals the sum of the numerators over the sum of the denominators.

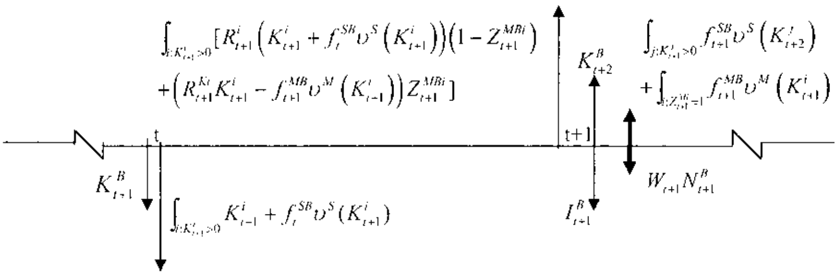
[26] Again, our assumption that the bank does not issue debt is irrelevant for our results. See the discussion of the Modigliani-Miller (1958) theorem at the end of section 7.1.5.

[27] We assume that in equilibrium, both the banking sector and agencies/the bond market get the same quality of applicants on average. In equilibrium, entrepreneurs will be indifferent about which route they should take to obtain their capital, so assigning them randomly is an innocuous assumption.

screening. Upon concluding the screening process, banks will lend the appropriate amount of capital to each firm. The firm must either repay the service fees and the productive capital with interest in period $t + 1$ or declare bankruptcy. In case of a default, the bank monitors the project and takes all that is left after deducting fees, exactly as if the firm had defaulted on a bond. At the same time, the bank also gets the fees, so unlike a bondholder, a bank truly gets the full residual value of the project!

Next, it is illuminating to partition the bank's cash flow as if it were produced by two divisions. The first, which we term the service division, does the actual production of screening and monitoring services, using capital chosen in the previous period ($K^B_{t+1}$) and labor hired in the current period. Monitoring services are paid by firms that have declared bankruptcy. But because the entrepreneurs have no resources in the first period of life, the fees for the screening services are paid by the other part of the bank, which we call the loan division. (Ultimately, of course, the bank will have to obtain these resources from its shareholders, as we will show next.) Once the screening is done, the loan division lends to entrepreneurs the funds it received as equity capital. The cash inflow of the loan division comes solely from returns on loans—either their contractual interest or the bankruptcy value of the firm, net of monitoring costs—exactly as in the case of bondholders. See figure 7.1 for a diagram showing the cash flows through a bank in any pair of periods.

The key to understanding our decomposition of a bank's cash flow is to realize that each period, the bank's shareholders must be paid the full returns



$\int_{i:K^i_{t+1}>0}[R^i_{t+1}\left(K^i_{t+1}+f^{SB}_t\upsilon^S\left(K^i_{t+1}\right)\right)\left(1-Z^{MBi}_{t+1}\right)$
$+\left(R^{Ki}_{t+1}K^i_{t+1}-f^{MB}_{t+1}\upsilon^M\left(K^i_{t-1}\right)\right)Z^{MBi}_{t+1}]$
$K^B_{t+2}$
$\int_{j:K^j_{t+1}>0}f^{SB}_{t+1}\upsilon^S\left(K^j_{t+2}\right)$
$+\int_{i:Z^{Mi}_{t+1}=1}f^{MB}_{t+1}\upsilon^M\left(K^i_{t+1}\right)$
t
t+1
$K^B_{t+1}$
$\int_{i:K^i_{t+1}>0}K^i_{t-1}+f^{SB}_t\upsilon^S(K^i_{t+1})$
$I^B_{t+1}$
$W_{t+1}N^B_{t+1}$

**Fig. 7.1 Cash flows for a bank's shareholders who invest in $K^A_{t+1}$ and generation-*t* firms' capital**

*Notes:* The bank's shareholders invest in both the bank's productive capital $K^B_{t+1}$ and generation-$t$ firms' productive capital $\int_{i:K^i_{t+1}>0}K^i_{t+1}$, as well as the associated screening fee $\int_{i:K^i_{t+1}>0}f^{SB}_t\upsilon^S(K^i_{t+1})$ at the end of period $t$. The variable $K^B_{t+1}$ is used in the bank's production in period $t + 1$, while $\int_{i:K^i_{t+1}>0}K^i_{t+1}$ is used in firms' production. From the bank's operation (i.e., screening generation-$t + 1$ and monitoring generation-$t$ projects), the shareholders receive a variable profit of $\int_{j:K^j_{t+1}>0}f^{SB}_{t+1}\upsilon^S(K^j_{t+2}) + \int_{i:Z^{Mi}_{t+1}=1}f^{MB}_{t+1}\upsilon^M(K^i_{t+1}) - W_{t+1}N^B_{t+1}$. The variable $I^B_{t+1}$ is investment, and $K^B_{t+2} - I^B_{t+1} = (1 - \delta)K^B_{t+1}$; that is, part of the shareholders' gross return is the initial bank capital, net of depreciation. At the end of period $t + 1$, the shareholders either receive the contracted interest rate $R^i_{t+1}$ from a firm or pay the necessary monitoring fee $f^{MB}_{t+1}\upsilon^M(K^i_{t+1})$ and receive all the residual payoff.

on their investment in the previous period. The intuition is the no-arbitrage condition as follows: suppose an investor chooses to hold the bank's stock for only one period; then he must be fully compensated for his entire initial investment when he sells the stock at the end of the period.[28] Because investors always have the option of selling out after one period, this condition must hold, even when investors keep the stock for multiple periods; otherwise, arbitrage would be possible.

This principle of shareholders receiving the full return on their investment every period is most important for understanding the cash flow associated with screening. At time *t*, a group of investors invest in a bank's equity, conditional on the expected return at time $t + 1$. It is these time-*t* shareholders who implicitly pay the fees for the bank's screening of new projects at time *t*, because screening enables them to invest in worthy projects and thus earn the returns at time $t + 1$.

We now demonstrate the equivalence between a bank and a rating agency plus a bond portfolio. We use a superscript *B* to denote bank decision variables. We denote by $R^H$ the rate of return that the households require in order to hold a bank's equity. Then, $R^H$ will be determined by the following asset-pricing equation:

$$\begin{aligned}(21)\quad & E_t[m_{t+1}(\{\mathbf{[}f^{SB}_{t+1}Y^{SB}_{t+1} + f^{MB}_{t+1}Y^{MB}_{t+1} - W_{t+1}N^B_{t+1} + (1-\delta)K^B_{t+1}\mathbf{]} + \\ & \mathbf{(}\textstyle\int_{i:K^i_{t+1}>0}\{R^{Bi}_{t+1}[K^i_{t+1} + f^{SB}_t\upsilon^S(K^i_{t+1})](1 - Z^{Mi}_{t+1}) + [\tilde{R}^{Ki}_{t+1}K^i_{t+1} - f^{MB}_{t+1}\upsilon^M(K^i_{t+1})]Z^{Mi}_{t+1}\}\mathbf{)}\} \\ & \qquad \div \{K^B_{t+1} + \textstyle\int_{i:K^i_{t+1}>0}[K^i_{t+1} + f^{SB}_t\upsilon^S(K^i_{t+1})]\})] = 1.\end{aligned}$$

The numerator equals the bank's total cash flow in period $t + 1$. It is organized into two parts (in bold brackets and parentheses) to correspond to the cash flows of the two hypothetical divisions in order to facilitate the comparison of a bank with a rating agency plus a bond portfolio. The first part is the cash flow of the service division, which does all the screening and monitoring; every term there is defined similarly to its counterpart in the numerator of equation (18)—the cash flow for the rating agency. The second part is the cash flow of the loan division, equal to the interest income, summed over all the entrepreneurs to whom the bank has made loans, net of the monitoring costs. Every term is defined similarly to its counterpart in equation (20), which is the return on a diversified portfolio of many bonds, each of which has a payoff similar to the numerator of equation (19).

The denominator of equation (21) is the sum of bank capital, which

[28] Alternatively, one can think of the bank paying off the full value of its equity each period—returning the capital that was lent the previous period, together with the appropriate dividends—and then issuing new equity to finance its operations for the current period. Of course in practice, most of the bank's shareholders at time $t + 1$ are the same as the shareholders at time *t*, but the principle remains the same.

comprises the amount the bank uses for screening and monitoring ($K^B$), the amount it lends to entrepreneurs, and the screening fees put up by this period's shareholders—which is best conceptualized as a form of intangible capital.[29]

Note that in order to derive the respective cash flows of the two divisions in the numerator, we deliberately add monitoring income $f^{MB}Y^{MB}$ to the first term and subtract monitoring costs $\int f^{MB}\upsilon^{M}Z^{Mi}$ from the second. But this manipulation on net leaves the bank's overall cash flow unchanged, because

$$(22) \qquad Y^{MB}_{t+1} = \int_{i:K^i_{t+1}>0}[\upsilon^{M}(K^i_{t+1})Z^{Mi}_{t+1}].$$

The reason is that the monitoring services produced generate income for the service division, and those are exactly the services the loan division must buy in order to collect from defaulting borrowers.

We have so far accounted for all of the cash inflow and outflow of the loan division and for the cash inflow corresponding to the provision of monitoring services for the service division. The next component is the cash inflow from providing screening services by the service division. According to the logic of fully compensating shareholders every period (discussed earlier), these screening services are implicitly paid for by time-$t + 1$ shareholders, and the fees constitute part of time-$t$ shareholders' return. They are the analogue of the screening fees in the denominator, which amount to $f^{SB}_tY^{SB}_t$ (for a reason similar to equation [22]), and were paid by time-$t$ shareholders to compensate time-($t$ – 1) shareholders. The final component of the capital return for the service division is the return of the depreciated capital to shareholders. (Depreciated capital is in the capital return of the loan division implicitly, because we use gross rates of return in that part of the numerator.)

### 7.2.3 Equilibrium with Asymmetric Information

We do not solve for the full set of equilibrium outcomes for all the variables, because we need only a subset of the equilibrium conditions to make the important points regarding bank output measurement. A major use of general equilibrium in our model is that it allows us to derive asset prices (and risk premia) endogenously in terms of the real variables (in particular, the marginal utility of consumption). Thus, in the context of this model, it is clear where everything comes from in the environment facing banks.

The first step toward proving the nature of the equilibrium is to note that the cash flow of any bank can be thought of as coming from two assets that households can choose to hold separately, each corresponding to equity

[29] That is, even though not recorded on balance sheets, the screening fees are nonetheless part of the overall investment funded by the investors today, and they expect to benefit from the payoff of that investment in the subsequent period.

claims on just one division of the bank. For the purpose of valuing an asset, it is immaterial whether the asset actually exists. Thus, it is immaterial whether the bank actually sells separate claims on the different streams of cash flows coming from its different operations; no bank does. But investors will still value the overall bank as the sum of two separate cash flows, each discounted by its own risk-based required rate of return. To take an analogy, Ford Motor Company shareholders in the United States certainly make different forecasts for the earnings of its Jaguar, Volvo, and domestic divisions, and they know that exchange rate risk applies to earnings from the first two but not to the third. Shareholders then add these individual discounted components to arrive at their valuation of the entire company.

It is important to note that no asset-pricing theory implies a unique way to split up a bank's—or indeed any firm's—cash flow generated by its various operations. Investors can choose to think of a bank as comprising the sum of any combination of its operations that adds up to the entire bank's cash flow. The crucial point is that the asset-pricing equation (4) must apply to *any* and *all* subsets of a bank's overall cash flow. But the service versus loan division is the most meaningful way of partitioning a bank's operations for the purpose of understanding real bank output, because it separates the bank's production of real output from its holding of assets on behalf of its investors. Moreover, this division generates two entities that both have real-world counterparts (i.e., rating agencies and bond markets). Therefore, this division is most useful, both for understanding and for measuring bank output (see our discussion of measurement next). This argument is made formally in the working paper version of this chapter (Wang, Basu, and Fernald 2004).

In conclusion, in any equilibrium, the service division of the bank must have a required rate of return on capital of $R^{SV}$, and each loan that the bank makes must have the same required return, $R^{Li}$, as it would have were it made in the bond market.[30]

### 7.2.4 The Model Applied to Measurement

We have presented the essential features of a simple DSGE model with financial intermediation. The model shows that because banks perform several functions under one roof, investors view a bank as a collection of assets—a combination of a bond mutual fund (of various loans) and a stock mutual fund (one that holds the equities of rating agencies). Investors value the bank by discounting the cash flow from each asset using the relevant risk-adjusted required rate of return for that asset. But in general, all of the

[30] We have shown that in any equilibrium that exists, households demand the same rate of return on each division of the bank as the rate on the rating agency and the bond portfolio, respectively. We have not claimed that an equilibrium must exist in this model or that the equilibrium previously described is unique—there may be multiple equilibria, with different asset prices associated with each one.

cash flows will have some systematic risk, and thus none of the required rates will be the risk-free interest rate.

In the context of the model, it is clear that proper measurement of nominal and real bank output requires that we identify the actual services banks provide (and are implicitly compensated for) and recognize that these services are qualitatively equivalent to the (explicitly priced) services provided by rating agencies. So, it is logical to treat bank output the same as the explicit output of those alternative institutions.

Another benefit of our approach—and a different intuition for its validity—is that the measure of bank output it implies is invariant to alternative modes of operation in banks. The prime example is the securitization of loans, which has become increasingly popular in recent years, where banks originate loans (mostly residential mortgages) and then sell pools of such loans to outside investors, who hold them as they would bonds. In this case, a bank turns itself into a rating agency, receiving explicit fees for screening (and servicing over the lifetime of the loan pool). Securitization should not change a reasonable measure of bank output, because banks perform similar services, regardless of whether a loan is securitized. Our model, which counts service provision as the only real bank output, indeed will generate the same measure of bank output, regardless of whether loans are securitized. But if one follows SNA93, then a bank that securitizes loans will appear to have lower output on average, because it will not be credited with the output, which is actually the transfer of the risk premium to debt holders. Thus, under SNA93, an economy with increasing securitization will appear to have declining bank output, even if all allocations and economic decisions are unchanged.

### 7.2.5 Different Capital Structures for Banks

The previous subsections of section 7.2 have all assumed that banks are 100 percent equity financed. This is unusual, in that we are used to thinking of banks as being financed by debt (largely deposits). But we will show next that the MM (1958) theorem holds in our model, so all of our previous conclusions are completely unaffected by introducing debt (deposit) financing. Of course, there is a large literature in corporate finance discussing how differential tax treatment of debt and equity and information asymmetry (between banks and households, that is) cause the MM theorem to break down. But we have deliberately avoided such complications in order to exposit the basic intuition of our approach. Once that intuition is clear, it will be simple to extend the model to encompass such real-world complications.

We have an environment where information is symmetric between banks and households, so there is no need for screening and monitoring when raise funds from (i.e., sell equity shares to) households. Thus, we reasonably assume that there are no transaction costs of any kind between banks and households. We also assume that interest payments and dividends receive the same tax treatment. In this setting, banks' capital structure is irrelevant, in that the required rate of return on banks' total assets is the same, with or without debt. When banks are leveraged, the required rates of return on the bank's debt and equity are determined by the risk of the part of the cash flow promised to the debt and the equity holders, respectively. Because debt holders have senior claim on the bank's cash flow, the ex ante rate of return they require is almost always lower than the rate required by shareholders. But the rate of return on the bank's total assets is the weighted average of the return on debt and equity, and it equals the return on the assets of an unlevered (i.e., all-equity) bank. This result is a simple application of the MM (1958) theorem.

The implication of this result is that all the preceding analysis of the imputation of implicit bank service output remains valid, even when banks are funded partly by deposits. We discuss the extension to deposit insurance in section 7.3.2; Wang (2003a) analyzes it fully.

The preceding overview of the model and the intuition for the key results should equip the reader for analytical details of the model in the appendix.[31] Alternatively, a reader more interested in the measurement implications now has the theoretical background for the measurement discussion that follows in section 7.3.

## 7.3 Implications for Measuring Bank Output and Prices

This model yields one overarching principle for measurement: focus on the flow of actual services provided by banks. This principle applies equally to measuring both nominal and real banking output, and thus the implied (implicit) price deflator. Following theories of financial intermediation, we model banks as providing screening and monitoring services that mitigate asymmetric information problems between borrowers and investors. Because screening and monitoring represent essential aspects of financial services in general, we would want any measure of bank output to be consistent with the model's implications. But the SNA93 recommendations for measuring implicit financial services—and the NIPAs implementation—are not. They generally do not accurately capture actual service flows.

The model highlights three conceptual shortcomings of the SNA93/NIPAs framework. First, the model shows that the appropriate reference rate for measuring nominal bank lending services must incorporate the borrower's risk premium, which is not part of bank output. Intuitively, the

[31] The appendix focuses on solving analytically the joint determination of the optimal contractual interest rate and each entrepreneur's choice of capital and labor in production. In particular, it spells out (a) the exact terms of the debt contract for entrepreneur's projects (including the interest rate charged) that is consistent with bank profit maximization and (b) each entrepreneur's utility-maximizing choice of capital and labor, given the debt contract.

borrowing firm must pay that premium, determined by the intrinsic risk profile of its cash flow, regardless of whether it obtains the funds through a bank or through the bond market (after getting certified by a credit rating agency). The return on the risky assets—loans or bonds—is part of the cost of capital to the borrowing firm and is income to households.

Second, the model shows that the timing of bank cash flows often does not match the timing of actual bank service output. As a prime example, screening is typically done before the loan generates income. This problem does not necessarily disappear, even when the origination fees are explicitly paid up front (ruled out in the model), because generally accepted accounting principles (GAAP) often require banks to artificially smooth these revenues over the lifetime of the loan, thus inadvertently reinstating the problem.

Third, the *measured* value of implicit output of services such as monitoring, whose *expected* value is incorporated in advance into the interest rate charged, most likely deviates from the *actual* output. This is because the realization of such services is contingent on loans' ex post realized return, which almost surely deviates from the ex ante expected return. Our model suggests that when the observed bank interest income is the realized return, the most suitable reference rate is actually the matched holding-period rate of return, not the ex ante expected rate, on a portfolio of debt with comparable risk and no financial services.

We now discuss further implications of these issues in the context of nominal and then real output.

### 7.3.1 Nominal Bank Output

Nominal bank services should correspond to the value of *service flows* provided by banks. It should exclude the value of any revenue that might flow through a bank that does not, in fact, correspond to actual financial services provided by the bank. This principle is embedded in the key first-order condition for a bank's optimal choice of contractual loan interest rate, $R^{Bi}_{t+1}$ (i.e., equation [A11] in the appendix):[32]

$$(23)\quad [(1 - p^i_t)R^{Bi}_{t+1}K^i_{t+1} + p^i_t E_t(Int^i_{t+1} \mid Default^i_{t+1})] - R^{Li}_{t+1}K^i_{t+1}$$
$$= R^{Li}_{t+1} f^S_t \upsilon^S_t(K^i_{t+1}) + p^i_t E_t[f^M_{t+1}\upsilon^M_{t+1}(K^i_{t+1})].$$

The variable $p^i_t$ is the probability that the borrower defaults, $(1 - p^i_t)$ otherwise. The expected interest payment by the borrower in case of default is denoted as $E_t(Int^i_{t+1} \mid Default^i_{t+1})$. Thus, the terms in square brackets on the left-hand side is the expected interest from lending $K^i_{t+1}$ to a borrower of risk type $i$; that is, the default probability-weighted average interest income the bank expects to receive. The variable $R^{Li}_{t+1}$ is the required rate of return that the bond market would charge borrower $i$ for a debt of the same size (and

[32] Note the distinction between the contractual rate and the required rate of return for a defaultable loan (section 7.1.1).

by our reasoning in the preceding sections, it is also the return that bank shareholders demand for financing such a bank loan). The variable $\upsilon_t^S(K_{t+1}^i)$ is the amount of screening services, the price per unit of which is $f_t^S$, and the variables $\upsilon_{t+1}^M(K_{t+1}^i)$ and $f_{t+1}^M$ are the counterparts for monitoring services.

Thus, the left-hand side is the difference between the expected bank income from loans of risk type *i* and the (hypothetical) income on a bond of the same size with the same risk characteristics. The right-hand side is the nominal value of the bank's *expected* services of screening and monitoring that loan.[33]

Equation (23) incorporates our three main points regarding measurement. First, consider reference rates. We define $F(R_{t+1}^{Bi}) \equiv [p^i\, R_{t+1}^{Bi} K_{t+1}^i + (1 - p^i)E_t(Int_{t+1}^i \mid Default_{t+1}^i)]/K_{t+1}^i$ as the interest rate the bank expects to receive, net of defaults, on loans to borrowers of type *i*. Then, the left-hand side of equation (23), an interest margin, can be expressed as an interest spread multiplied by the loan size, $K_{t+1}^i$:

$$[F(R_{t+1}^{Bi}) - R_{t+1}^{Li}]K_{t+1}^i. \tag{24}$$

As Fixler, Reinsdorf, and Smith (2003; hereafter FRS) suggest, one can use "interest margins as values of implicit services of banks" (34). The key issue is deciding what reference rate(s) to use. Equation (24) makes clear that $R_{t+1}^{Li}$ is the appropriate reference rate for imputing the implicit value of bank output.

Importantly, this reference rate must be *risk adjusted* (i.e., contain a risk premium reflecting the systematic risk associated with the loans). In sharp contrast, U.S. and other national accounts stipulate a reference rate that explicitly *excludes* borrower risk. The 2003 benchmark revisions of the U.S. NIPAs define the reference rate as the average rate earned by banks on U.S. Treasury and U.S. agency securities.[34] As FRS argue, "If a highly liquid security with no credit risk is available to banks, the banks forego the opportunity to earn this security's rate of return . . . when they invest in loans instead" (34). That's true; but it's also true that banks forego the opportunity to invest in high-risk/high-yielding junk bonds!

Our model clarifies the apparent ambiguity inherent in the opportunity cost argument by incorporating modern asset-pricing theories (the CCAPM, specifically). Indeed, by combining theories of asset pricing and financial intermediation, our model (and Wang's [2003a]) extends and generalizes the user cost framework to take account of uncertainty and asymmetric information.

[33] The potential monitoring cost is not known in advance but must be expected, because it depends on wages and productivity that will be realized in period $t + 1$. The variable $E_t$ is the expectations operator, conditional on time-$t$ information.

[34] This average rate is not, in fact, a risk-free rate, even in nominal terms. In particular, U.S. agency securities have a positive and time-varying interest spread, reflecting credit risk, over Treasuries of matching maturities.

Asset-pricing theories imply that an asset's required rate of return depends (positively) on its systematic risk. In two special cases, the required return equals the risk-free rate: if there is no systematic risk (i.e., only idiosyncratic risk, which creditors can diversify away) or if investors are risk neutral. In such a world, there would be no risk premia. Otherwise, the correct reference rate (i.e., opportunity cost of funds) for imputing bank lending services must be adjusted for systematic risk. Thus, our model makes clear that the current NIPAs implementation of the user cost approach—with a *risk-free* reference rate for lending services—is not appropriate in the realistic world with uncertainty.

What is the intuition for risk-adjusted reference rates? When a bank keeps loans on its balance sheet and charges implicitly for services, it sets each loan rate to cover both the services provided *and* the riskiness of the loan. In equilibrium, the loan interest rate, net of implicit service charges, must compensate the ultimate suppliers of funds (i.e., households in this model) for the disutility arising from the risk. Conversely, the borrower could (at least conceptually) go to a rating agency, get certified, and then issue bonds at the risk-adjusted rate. Adjusting reference rates for risk thus preserves neutrality with respect to economically identical institutional arrangements for obtaining external funds.

Comparison with securitization further illustrates the rationale for risk-adjusted reference rates.[35] When banks securitize loans, they receive explicit payments for services; household optimization implies that the securitized asset yields a service-free interest rate that reflects each loan's risk properties. However, banks perform the same kinds of screening and monitoring services and arrive at much the same optimization rule, which governs the service-free interest rates (along with other choice variables), whether loans are held on the balance sheet or securitized. The principle of neutrality across economically identical lending arrangements implies that bank output measurement should be invariant. Otherwise, measured bank output fluctuates with the share of securitized loans, even if actual bank services are constant over time.

Securitization thus provides a useful conceptual benchmark against which to judge the validity of any measure of implicitly priced bank services. Our model, in effect, imputes implicit bank output to be invariant, regardless of whether loans are securitized. That is, the nominal value of bank services equals total bank interest income minus the amount of service-free pure interest—corresponding to the rate that would be charged on a securitized loan pool with matching risk (i.e., if the loans were securitized).

Furthermore, the model implies that the NIPAs mismeasure the oppor-

[35] By the end of 2003, over 80 percent of residential mortgage loans and 30 percent of consumer credit were securitized (Federal Reserve Board 2009) and are an increasing share of loans to businesses.

tunity cost of banks' own funds (i.e., financial assets minus liabilities). The SNA93 recommends that the opportunity cost of a bank's own funds be netted out of its imputed service output; the version implemented in the 2003 NIPAs revision uses the risk-free rate as the user cost of banks' own funds (see FRS 2003, 36). Our model can be construed as implying a similar netting-out principle, because any reference rate can be expressed as a weighted average of the respective user cost of banks' debt and own funds. However, the model makes clear that both rates ought to be risk adjusted, according to the different risk on banks' debt and own funds, respectively.

More importantly, our model-implied measure of implicit bank service output does not depend on the bank's capital structure, which is but a coincidental state variable. That is, the opportunity cost of funds for a loan needs to be risk adjusted according to the same asset-pricing theories, regardless of whether the lending is financed by intermediation (i.e., deposit taking) or by banks' own funds.

Counting the risk premium as part of bank output also overstates GDP. In the model, GDP is not mismeasured, because financial services are an intermediate input into nonfinancial firms' production. An SNA93-based measure misallocates some value added to banks. But the logic of the model applies to consumer loans (e.g., mortgages and credit cards), which also involve risk and risk-assessment services. So, final services to consumers and GDP would be overstated if we include the risk premium in bank output.

The second general issue that equation (23) highlights is the timing mismatch between the provision of screening services and the resulting cash flow. Banks screen borrowers in period $t$, but these services are not compensated until period $t + 1$. The borrower's (future) payment of $R^{Li}_{t+1} f^{S}_{t} \upsilon^{S}_{t}(K^{i}_{t+1})$ for screening services thus exceeds the contemporaneous nominal value of the services,$f^{S}_{t} \upsilon^{S}_{t}(K^{i}_{t+1})$. Ideally, one would attribute those services to period $t$, when the bank screens and originates the loans, rather than to $t + 1$.

In principle, if banks charge explicit origination fees upfront—rather than rolling these fees into the interest rate—then the timing mismatch becomes less important. In practice, firms often do pay explicit origination fees. But GAAP require that banks amortize the origination fee over the life of a loan. So, the reported income stream is artificially smoothed relative to the timing of service provision. If true screening services vary over time, then accounting data might not properly reflect this variation. In this case, direct quantity data (such as counting the number of loans originated) can help ensure correct timing.

The third general point from equation (23) is that *actual* monitoring output differs from *expected* (i.e., $p^{i}_{t} E_{t}[f^{M}_{t+1} \upsilon^{M}_{t+1}(K^{i}_{t+1})]$ on the right-hand side of equation [23]), the value of which is included in the expected interest margin (i.e., the left-hand side of equation [23]). That is, the contractual rate covers expected monitoring services based on ex ante probability of default, but monitoring takes place only when a borrower actually defaults ex post. In fact, *neither* ex ante *nor* ex post interest margins match the actual value of monitoring, while the two margins almost surely do not equal each other. We suspect that in good times, banks do less monitoring than expected, while enjoying higher-than-expected interest margins; in bad times, they do more monitoring than expected, while suffering lower-than-expected interest margins. Thus, in a boom, ex post interest margins exceed ex ante margins, and in turn exceed the value of banks' actual service flows. In a recession, ex post interest margins fall short of ex ante margins, and in turn fall short of the actual value of service flows.[36]

In general, adjusting the ex post interest margin for the actual rate of default yields more accurate measurement of nominal bank service output. Such adjustments can be implemented; for example, Wang (2003b) uses bank holding company data to adjust the ex post interest income for the default realizations.[37]

This problem of mismeasuring monitoring services is unlikely to disappear, even when one averages over a large number of loans, unless there was no aggregate risk. This nondiversifiable deviation of actual from expected cash flow is precisely the *reason* why there is a risk premium. That is, in good times, when output and consumption are high (so marginal utilities are low), banks generate more residual cash flow for shareholders; in bad times, when output and consumption are low, banks generate less residual cash flow.

We conclude this section by discussing how to extend the model to include bank depositor services (e.g., direct transaction and payment services, safe deposit boxes, etc.). Conceptually, they raise fewer complications than lending services, especially regarding the treatment of risk. Without the service component, deposits are simply fixed-income securities. A straightforward extension of the model implies that nominal depositor services equal the margin between interest paid and interest imputed, using reference rates based on market debt securities with comparable risk. For balances covered by deposit insurance, the correct reference rate is the risk-free rate, as used in the NIPAs. For balances not covered or without deposit insurance, however, depositors would demand a higher expected return that depends on the default risk of a bank's asset portfolio and its capital structure. So, the NIPAs measure is appropriate only for insured balances and is unlikely to remain correct for countries without deposit insurance, such as New Zealand.

[36] See appendix 2 in Wang, Basu, and Fernald (2004; hereafter WBF) for technical derivations of these results and for how to adjust for actual default.

[37] Going forward, more relevant data is likely to be generated in the coming implementation of the Basel II accord for capital requirement, which encourages banks to develop internal risk management systems.

### 7.3.2 Is Risk Assumption a Service?[38]

One interpretation of the NIPAs choice of the reference rate is that they construe risk bearing as an additional service provided by banks.[39] Our model, on the contrary, considers only screening and monitoring services to be bank output, produced using capital and labor. Presumably, one could write down alternative internally consistent *accounting* systems that are consistent with any given *economic* model. So, one could probably write down another accounting system, also internally consistent, where bearing risk is treated as service output in all transactions.[40]

Nevertheless, at least two intuitive criteria help in choosing between different, internally consistent accounting frameworks. First, one wants to choose an accounting framework where the quantities measured have natural economic interpretation. Second, the framework should treat identical market transactions identically. The system we propose meets these two criteria. The current system, in contrast, does not.

We have already discussed several examples that illustrate these criteria. For example, if firms are indifferent between borrowing from banks or from the bond market, then we would want to treat them identically with respect to their marginal decisions. The current national accounts do not do so.

More generally, the current system does not treat risk bearing consistently across alternative market arrangements. Indeed, the current accounting system leads to very peculiar outcomes when applied to outside banks narrowly defined. Consider mutual funds. The account holders of mutual funds are owners of the assets—shareholders. Because the current system credits bank shareholders with the risk premium for assuming risk, mutual fund shareholders should be treated in the same way. Thus, the NIPAs framework would seem to imply that the mutual fund management industry should be credited with producing services equal to actual asset returns in excess of the risk-free return (multiplied by the market value of the assets).

We do not think it appropriate to credit the mutual fund industry with producing trillions of dollars of value added, corresponding to the difference between average stock returns and risk-free interest. Our framework would say that we should credit mutual funds only with providing the services that people think they are buying from mutual funds—transactions, bookkeeping services, and sometimes financial advice. We think this corresponds much more closely to the economic reality.

[38] We thank Paul Schreyer, whose comment on this chapter stimulated us to add this section.

[39] For example, FRS (2003) say, "The spread between the reference rate of return and the lending rate is the implicit price that the bank receives for providing financial services to borrowers, which include the cost of bearing risk."

[40] We are not aware of any fully worked out models that explore the full implications of treating risk assumption as a service output.

Finally, counting risk assumption as a bank service causes conceptual difficulties when the resulting measure of output is used in productivity studies. Suppose one bank turns down very risky loans, whereas another seeks out high-risk projects and lends to them at high interest rates. Suppose also that both banks use the same amount of capital and labor to provide exactly the same processing services, such as screening and monitoring. They have the same output and productivity by our definition. It is undesirable to claim instead that the bank making more risky loans—which the other bank could have made but declined—is the more productive bank, *solely* because of the riskiness of its loan portfolio.

### 7.3.3 Real Bank Output

It is clear that one wants to measure real output as the actual service flow provided by banks. The model aims to focus on the issue of risk and bank output measurement; hence, it considers only bank lending activities, which essentially involve processing information—specifically, financial and credit data. These services are qualitatively similar to other information services, such as accounting and consulting.

Banks provide many distinct types of services. The model captures this by the different production functions for screening and monitoring. Screening depends on the number of *new* loans issued, whereas monitoring depends on the number of *outstanding* loans (in the model, inherited from last period). There is also heterogeneity within either activity. The amount of screening and monitoring, respectively, that is needed for a loan depends on many factors that differ across loans.[41] The model captures this in the form of size differences across loans. The multiproduct nature of bank services implies that aggregate bank output should be defined as a Törnqvist or Fisher index of the quantity (index) of each distinct service type.

Measuring the real value of monitoring services presents the same difficulty that affects nominal value measurement: measured output based on deflated nominal value (assuming both the risk premium and the cost of screening are properly accounted for) generally differs from *both* the actual and the expected output of monitoring. Thus, in a downturn, productivity analysts would see a banking sector experience lower imputed output, despite absorbing as much (if not more) primary or intermediate inputs. Then, measured banking total factor productivity (TFP) would fall sharply in a downturn, even if actual TFP did not change.

The way around this difficulty is to measure real monitoring services using

[41] For example, a loan's denomination, the borrower's industry and geographic location, and his or her previous interaction with the bank are all relevant factors. In practice, the amount of screening and monitoring needed differ more across commercial and industrial loans than across (conforming) residential mortgage and consumer loans. See WBF (2004) for a more detailed discussion of the heterogeneity in information services across different categories of loans.

direct quantity indicators. For instance, one can make use of the number of loans overdue or delinquent in each period to gauge the actual amount of monitoring performed; one may be able to collect data also on the associated costs of restructuring and foreclosure to estimate the quality-adjusted output of monitoring different loans.

How do these conceptual issues relate to what the national accounts actually measure (or attempt to measure)? The national accounts base their estimates of real output on a real index of banking services calculated by the Bureau of Labor Statistics (BLS). In terms of lending activities, the BLS (1998) tries to count activities such as the number of loans of various types (commercial, residential, credit card, etc.). Within each category, different loans are weighted by interest rates, the presumption being that loans that bear a higher interest rate involve more real services. Across categories of services, output is then aggregated using employment weights.

As the BLS technical note makes clear, limitations on the availability of appropriate data force many of their choices. Conceptually, at least, we highlight a few of the issues suggested by the model.

First, one should try to distinguish new loans from the stock of old loans, because they involve different services (that is, screening and monitoring, respectively). In particular, the timing of when each type of service is undertaken differs. Second, interest rates are probably not the right weights to use within loan category. Relative interest rates contain the compensation for (a) systematic risk, (b) screening services, and (c) expected monitoring services (tied to expected default probability). Thus, the relative interest rate weights are probably correlated with the proper weights–but imperfectly and certainly not linearly. Third, nominal output, instead of employment requirement, should be used as the weight for aggregating output across categories of services. Last, as noted earlier, one should try to measure real monitoring output more directly. Even using the number of outstanding loans—as the BLS does, on the grounds that existing and new loans require some services—will not capture the likely countercyclical pattern of actual monitoring services. (In fact, the number of outstanding loans is more likely to be procyclical.)

### 7.3.4 Price Deflators for Bank Output

Conceptually, what do we mean by the price of financial services? We use what seems like a natural definition of the price deflator: the nominal value of services divided by the real quantity index. So, the deflator is directly implied by the preceding discussions of both nominal and real output measures.

Our definition, although natural and intuitive, differs entirely from the common meaning of prices for financial instruments. The latter often refer to interest rates themselves, as in “the interest rate is the price of money,” or as in “pricing a loan,” which refers to setting the proper interest rate.

Similarly, the user cost approach refers to the interest rate *spread* (between the loan rate and a reference rate, scaled by a general deflator such as the Consumer Price Index [CPI]) as the user cost price of a loan.

This sometimes loose reference to financial prices can be appropriate in the context of discussing rates of return on financial *instruments.* But the model makes clear that neither the interest rate nor the interest rate spread is the price for financial *services,* even though banks often charge for these services indirectly via an interest rate spread. Similarly, the book value of loans is not the right quantity measure of lending services. (Bank efficiency studies often inappropriately treat loans' book value—deflated with a general price deflator, such as the CPI—as the quantity of bank output and treat interest rate as the price.)

As an explicit example, consider depositor services. Depositors implicitly pay for the services they receive by accepting a lower interest rate. Suppose a depositor decides to purchase *fewer* financial services by putting the *same* deposits in an Internet bank that offers a higher interest rate. The natural interpretation is that the nominal quantity of services falls (as measured by a lower interest margin) because the real quantity of services falls. It would clearly be mistaken to claim that nominal output falls because the price (i.e., the interest rate spread) falls, while the quantity (measured by the dollar value of deposits) is fixed.

In summary, the model implies the proper price of financial services by providing theoretical guidance for measuring the nominal and real values of such services. As important, we now discuss how to meet the practical challenges of implementing the model's implied nominal and real output measures.

### 7.3.5 Implementing the Model's Recommendations in Practice

To properly measure the value of nominal bank services, we must first estimate and remove the risk premium on bank loans. The risk premium on comparable market securities (i.e., commercial papers, mortgage-backed securities, etc., that are subject to the same systematic risk) serves as a good proxy. Such proxies are readily available. Wang (2003b) suggests some securities one may use and provides a preliminary estimate of bank service output, free of the risk premium. (Her estimate suggests that on average, the risk premium may amount to 20 to 25 percent of imputed bank service output.)

Arguably, a better alternative is a rate that is adjusted for the risk, as assessed according to each bank's internal risk-rating system. Indeed, Basel II requires that banks assess their risks even more carefully than they already do—offering an opportunity for improving the accuracy of the estimate of risk premia. This should then lead to a more accurate estimate of (a Törnqvist or Fisher index of) aggregate real bank output, where the nominal output share of each distinct type of bank service serves as the aggregation weight.

Second, we need the timing of measured output—screening, in particular—to match when services are rendered rather than when services generate revenue. If explicit origination fees are available for a type of loan, national accountants can collect cash-based accounting data on total origination income for that type of loan and estimate the true screening output by deflating the income with the explicit fee. The fee can also serve as a proxy for the price of similar charges that are implicit. Or, one can derive quantity indexes from direct counts of distinct activities (e.g., the number of new credit card loans made), and the weighted sum of the growth rates of these indexes gives the growth of aggregate service output.

Third, to address the issue that actual monitoring services (both nominal and real) are likely to differ from both the expected and the measured value, one can make use of bank data on actual loan default rates, as noted previously. In addition, because the correct reference rate equals the rate of return on market securities with comparable risk-return characteristics as bank loans, one can use ex post returns on such matched fixed-income market securities to more accurately infer bank service flows.

Finally, consider depositor services. It seems easier to define a product for depositor services than for lending services, because depositor services are more homogeneous across banks and in terms of product characteristics.[42] Conceptually, each distinct type of transaction should be viewed as one depositor service output. Thus, each ATM or teller-assisted transaction is presumably a composite good of several distinct activities. But for practical reasons, we can define each visit to an ATM or a teller as one unit of a service product. Similarly, without data on the number of each distinct type of transaction, we can treat maintaining each account of a given type as one product and use the number of deposit accounts of different types to measure output. This amounts to assuming that each account of a given type requires the same amount of bookkeeping, payment processing, and so forth, every period.

## 7.4 Further Implications for Measurement

The model's framework helps clarify several other issues in the literature. These include the use of assets/liabilities themselves as a measure of bank output, the question of whether to include capital gains as part of bank output, and how to measure other financial services/instruments provided by banks.

First, the model provides no theoretical support for the widespread practice of using the dollar value of interest-bearing assets (loans plus market

[42] For instance, safe deposit box rentals are a relatively homogeneous activity, as are wire transfers, money orders, and cash withdrawals. To a lesser degree, so are cashing checks and opening accounts of a specific type.

securities) on bank balance sheets deflated by, for example, the GDP deflator, as real bank output.[43] This practice is standard in the empirical microeconomic literature on bank cost and profit efficiency.[44] Our model suggests a simple counterexample, in the spirit of the bank that does nothing. Suppose a bank has accumulated a loan portfolio by doing prior screening and monitoring but originates no new loans and does not need to monitor any old ones at a particular point in time. Then, our model makes it clear that the bank has zero service output in that period. But the microliterature would conclude that the bank's output is arbitrarily large, depending on the size of its existing loan portfolio.

Second, although the model does not explicitly consider capital gains, it provides a guiding principle for answering the question of whether capital gains should be counted in banking or financial output. Capital gains and interest income are two often interchangeable ways of receiving asset returns, with the former related more often to unexpected returns and the latter related more to expected returns. If interest income is often employed as implicit compensation for financial services provided without explicit charge, then in principle, capital gains can be used in place of interest for the same purpose. By design, such capital gains will be *expected* gains, because the service provider expects to be compensated. These gains should be recognized as implicit compensation for real financial services. Otherwise, capital gains should not be recognized.

To illustrate this principle, we use the same example of screening services in lending. Suppose that instead of holding loans on its balance sheet, a bank sells them after its shareholders have put up the initial funding, consisting of both the productive capital lent to the firms and the screening fees. Also assume that the bank only records the value of the capital lent, but not the screening fees, as assets on its balance sheet.[45] Accordingly, the loans' contractual interest rates are quoted with respect to just the capital lent, although the expected value of the interest will cover the screening fees as well. Then, when the bank sells these loans (i.e., debt claims on the firms' cash flows), it will enjoy a capital gain equal to the value of the screening fees, because the present value of those claims exceeds the book value by exactly the amount of the fees. Clearly, the capital gain in this case is qualitatively the same as the extra interest income the bank would receive in compensation for its services if it kept the loans. So, this capital gain should be counted as bank output.

[43] Some existing studies also use deposit balance to measure depositor services, implicitly assuming that the service flow is in fixed proportion to the account balance. But Wang (2003a) shows, in realistic settings, that the relationship between the quantity of services and the account balance is likely to be highly nonlinear and time varying.

[44] See, for example, Berger and Mester (1997) and Berger and Humphrey (1997) for surveys of the literature.

[45] This is a quite likely scenario, because the fees are like intangible assets, which are often poorly or simply not accounted for on balance sheets.

On the other hand, following the same principle, capital gains or losses purely due to the random and unexpected realization of asset returns should not be counted as financial output. This can be seen in the model from the fact that the ideal reference rate is an ex post rate. The economic intuition is fairly clear, although it is best illustrated with multiperiod debts. Suppose we modify the model so that entrepreneurs and their projects last three periods. Then firms would borrow two-period debt, which would be screened and monitored in the usual way. Suppose also that aggregate technology is serially correlated. Then, a favorable realization of technology would lead to a capital gain on all bonds and bank loans that have yet to mature, because a good technology shock today raises the probability of good technology in the next period, which reduces the probability of bankruptcy in that period. But these capital gains do not reflect any provision of bank services—in fact, loans one period from maturity would be past the screening phase and would not yet require monitoring—and thus, the capital gains should not be counted as part of output. Intuitively, the only exception to this rule would arise if the capital gains on the loans are due to the provision of some banking service. For example, if banks provide specialized services to firms that make them more productive, which leads to an appreciation in the value of their assets, one would want to count some of that gain. This seems unlikely in the context of banks but may be realistic for venture capital firms.

Third, our model can be readily applied to value implicit services generated by banks when they create financial instruments other than loan contracts, and it can also be applied to measure implicit services generated by other financial institutions that create a wide variety of complex financial instruments.

The general applicability of our method stems from the fact that a loan (i.e., bond) subject to default risk is equivalent to a default-free loan combined with a short position in a put option[46] (i.e., giving the borrower the option of selling the project to the lender at a prespecified price). We denote the contractual interest rate as $R^i$ and a project's actual rate of payoff as $R^A$. Then, the payoff on a defaultable loan equals min $(R^i, R^A)$; a lender receives either the promised interest or the project's actual payoff, whichever is less. We can rewrite the risky loan's payoff as:

$$\min(R^i, R^A) = R^i - \max(0, R^i - R^A). \tag{25}$$

The first term describes the payoff from a riskless loan guaranteed to pay $R^i$; the second term, max(.), is the payoff to a put option on the project with a strike price of $R^i$. When the project pays less than $R^i$, the option holder would exercise the option (selling the project and receiving $R^i$) and earn a net

[46] Put options, in general, offer the holder the option to sell an asset (real or financial) at a prespecified price to the party that offered (i.e., shorted) the option contract.

return of $R^i - R^A$.[47] When the project pays more than $R^i$, the option holder would not exercise the option and thus earn zero return. The negative sign in front of the second term means the lender of the defaultable loan is shorting (i.e., selling to the borrower) the put option. More generally, equation (25) describes the fact, well known in corporate finance, that a firm's bondholders essentially write a put option to the firm's shareholders.

In banks' cases, this means that issuing a loan is qualitatively the same as writing (i.e., holding a short position in) a put option to the borrower. The processing costs incurred should be the same as well, because all the risk in a defaultable loan lies in the embedded put option. So, screening and monitoring is only needed for that risky component, whereas the other component—the riskless loan—should involve little information processing. Therefore, the implicit services that banks produce in the process of underwriting a loan can be viewed as equivalent to services generated in the process of creating a financial derivatives contract.

This means the measure of implicit bank services implied by our model can be applied equally well to similar services that financial institutions generate in creating other types of financial instruments. The general principle is the same: apply asset-pricing theories to price the financial instrument by itself; the difference between that value and the contract's actual value yields the nominal value of the implicit services. Measurement issues similar to those related to lending, as previously discussed at length, will no doubt arise; our recommendations for implementing the output measure in practice apply then as well.

## 7.5 Conclusions

We develop a dynamic stochastic GE model to address thorny issues in measuring financial service output. Financial institutions perform screening and monitoring services to resolve asymmetric information. Measuring real output involves measuring the flow of actual financial services produced; measuring nominal output requires measuring the income that correspond to these services. Equilibrium asset-pricing conditions help resolve some of the perplexing conceptual issues in the literature.

A key result, as in Wang (2003a), is that the risk premium on loans is not part of banks' nominal output, because it does not correspond to the screening and monitoring services provided by banks. The risk premium is part of the capital income transfer from the final users to the ultimate suppliers of loanable funds (i.e., from the borrowing firms to households). The

[47] This is, in effect, one way to describe default: a defaulted borrower's zero total payoff can be decomposed into two pieces—a negative net worth of $R^A - R^i$ exactly offset by a positive payoff of $R^i - R^A$ from holding the option.

rationale is intuitive enough: one wants to measure the output of economically similar institutions the same way. In the model and in the world, bank services to borrowers essentially combine the services of a rating agency with funding through the bond market. But the bond market is clearly just a conduit for transferring funds from households to firms; equally clear, the return on those funds, including any risk premium, is not the output of the rating agency!

Conversely, our implied output measure also satisfies the intuitive principle that a firm's output is invariant to the specific institutional source of external funding, as long as its liabilities have the same risk-return profile and incur the same informational services. The firm pays the same risk premium and the same service charges (implicit or explicit), regardless of whether the funds flow through a bank or through the bond market.

The model highlights the conceptual shortcomings in the existing national accounting measure of bank output. By counting the risk premium as part of nominal bank output, the current SNA93 and NIPAs measures treat economically identical alternative funding institutions differently and alter the output of the borrowing firm, depending on its source of funding. At the same time, the model makes clear that the book value of financial instruments on the balance sheet of banks, commonly used as the measure of bank output in the large body of bank efficiency studies, generally does not correspond to the true bank output, nominal or real.

In addition, we highlight two practical problems. First, the timing of cash flows often does not match the timing of actual bank services, because the bank screens in advance and then generates income over time. Second, expected bank net interest income incorporates the ex ante expected value of providing monitoring services; but ex post the quantity and nominal value of these services do not match the realized net interest income of the bank. We have discussed ways to resolve these problems.

More generally, we advocate a model-based approach to measurement for conceptually challenging areas of financial services and insurance.[48] We suggest that researchers write down an explicit optimizing model of what each firm/industry does. A model clarifies what we *want* to measure and thus what the ideal data set is. Only after we know how to do measurement in principle can we begin to compromise in practice. And if the shadow costs of the data availability constraints are too high, the measurement community can call for additional data collection projects.

Our approach suggests several priorities for extending theory and collecting data. Our method applies directly to bank services produced in the process of generating financial instruments other than loans (e.g., lines

[48] For recent studies, see, for example, Schreyer and Stauffer (2003), who consider an extensive set of services provided by financial firms; chapter 6 in Triplett and Bosworth (2004b) discusses the measurement of insurance output.

of credit, derivatives). Likewise, our model applies to the production of financial services by nonbank intermediaries. Thus, our work serves as a template for measuring financial service output of the financial sector more generally. Also, our method connects financial measurement to the vast body of research on asset pricing and corporate finance. Thus, conclusions from these literatures on some real-world complexities (e.g., realistic tax treatment of interest and capital gains) can be readily incorporated. Wang (2003a) discusses some of these issues in depth, such as the effects of deposit insurance.

On data collection, Basel II reporting requirements can generate data on the risk profiles of banks' assets. Also, constructing an index of real bank output requires improved surveys; for example, direct quantity counts for a wider variety of bank activities would be useful, and data on how marginal costs of originating and monitoring loans vary with size and other attributes would help with quality adjustment.

We conclude by summarizing the answers to the four questions posed in the abstract. First, the correct reference rates on loans must incorporate risk. Second, one does not want to use an ex ante measure of the risk premium on bank funds in each reference rate—using an ex post holding return on bonds of comparable riskiness comes closer to measuring the actual production of bank services. But the timing mismatch and other problems mean that in general, no single reference rate provides a perfect measurement of the nominal value of implicit service output. Third, the price deflator for financial services generally is not the overall price level. Financial services are an information product, qualitatively similar to other information processing services (e.g., consulting); in general, the price of financial services relative to final output will not be constant. Fourth, we should count capital gains as part of financial service output only if the gains are expected as implicit compensation for actual services provided.

## Appendix

### *Financial Intermediation under Asymmetric Information and Bank Output*

This appendix solves a bank's and its borrower's joint optimization problem to derive analytically why and how implicit bank output can be measured by decomposing a bank's overall cash flow.[49] The key equation underlying the decomposition is the one that sets the optimal interest rate on a bank loan.

[49] This is a summary of appendix 1 in WBF (2004), which the reader is urged to consult for more detailed derivations.

### Screening and Monitoring

In the model, each project (operated by a firm that is owned by an entrepreneur) spans two periods. Banks' first function is to screen each project in the first period to uncover its credit risk, which determines the loan's interest rate. We assume that banks' screening technology can fully discern a project's type, denoted $\theta^i$, to avoid unnecessary complications. Because entrepreneurs have no initial wealth, banks price the implicit fee into the interest charged, to be paid next period. Firms then use the loans to purchase capital.

In the second period, each firm uses the capital to produce the single homogeneous final good of the economy and is liquidated at period end. The lending bank takes no further action unless a firm defaults, in which case the bank incurs a cost to monitor the firm and extracts all the residual payoff.[50]

In summary, banking service output consists of screening the *new* projects born in each period and monitoring the *old* projects that fail. Screening and monitoring have different production functions. They parsimoniously represent the myriad of tasks performed by banks in their general role as information processors in the credit market. So, the analysis here can be readily adapted to study (implicit) bank output in creating other financial instruments, such as derivatives contracts.

### Bank Cost Functions for Screening and Monitoring

A loan's interest rate depends in part on the bank's cost of screening and monitoring. So, we first detail properties of these two cost functions. Banks have the same CRS technology as the rating agency for screening and monitoring, respectively (see equation [12]).[51] The cost of screening or monitoring varies with each loan's attributes, most likely in a nonlinear fashion. The model represents loan attributes with a single dimension of size. Then, the cost of screening ($S$) or monitoring ($M$) a *single* loan of size $L^i$ in time $t$ can be written as

$$\text{(A1)} \qquad \begin{aligned} c_t^J &= \upsilon^J(L_t^i) f^J(W_t, R_t^{SV} - 1 + \delta) \\ &= \frac{\upsilon^J(L_t^i)}{\mathrm{A}^J}\left(\frac{W_t}{1-\beta^J}\right)^{1-\beta^J}\left(\frac{R_t^{SV} - 1 + \delta}{\beta^J}\right)^{\beta^J}, \quad J = S, M. \end{aligned}$$

The term $f_t^J(.)$ is the cost of processing a numeraire loan (whose size is normalized to one). So, it only depends on factors common to all ($S$ and

[50] That is, we adopt the standard costly state verification setup from Townsend (1979). See WBF (2004) for a discussion of its distinction from the monitoring function in Diamond (1991).

[51] That is, referring to the constant marginal cost of processing each *additional loan* of given attributes. The CRS assumption is made for simplicity, given that the degree of returns to scale does not matter for deriving the right measure of bank output.

$M$)s output: input prices (the wage rate $W_t$ and the shadow rental price of bank capital $R_t^{SV} - 1 + \delta$; see section 7.1), output elasticities ($\beta^J$), and the technology parameter ($A^J$). The other term $\upsilon^J(L^i)$, which depends solely on size, then scales the numeraire cost across loans of different sizes. Given perfect competition for both screening and monitoring, $f_t^J$ will also be the price (relative to the price of the final good) of the respective numeraire service, and $\upsilon^J(L^i)$ will be the weight for aggregating services.

We assume, intuitively, that the cost of monitoring a loan grows more slowly than linearly in loan size; that is, $\upsilon^{M\prime} > 0$, and $\upsilon^{M\prime\prime} < 0$. Aggregate monitoring output then depends on not only the sum but also the distribution of loan sizes. On the other hand, $\upsilon^S(.)$ is assumed to be convex for technical reasons (explained next).

### Terms of the Loan Contract for Entrepreneurs' Projects

We now describe terms of the loan contract, which will enter a bank's optimization problem in the next section. For a penniless entrepreneur $i$ born in period $t$ (called generation-$t$) to purchase capital $K_{t+1}^i$ for his or her project, he or she must borrow $K_{t+1}^i$ (the subscript denoting the period in which the capital is used in production), plus the screening fee $f_t^S\upsilon^S(K_{t+1}^i)$, and must pay off everything at the end of period two from the project's payoff.

Project $i$, arriving in period $t$, pays $\theta^i R_{t+1}^K$ for every unit of investment, where $R_{t+1}^K$ is the average ex post gross return (to be realized in period $t + 1$) across all potential projects, while $\theta^i$ is the project-specific risk parameter of $i$ (i.e., type) uncovered by the bank screening process. The variable $\theta^i$ depends on the random draw of $i$ from the distribution of project productivities, $z^i$.[52] So, $\theta^i$ is independently and identically distributed across time and projects. We denote its cumulative distribution function as $G(\theta)$, with $E(\theta) = 1$. The variable $R_{t+1}^K$ represents the aggregate risk and thus depends on the realization of the aggregate productivity shock in period $t + 1$ (i.e., $A_{t+1}$).[53] We denote the conditional cumulative distribution function as simply $F(R_{t+1}^K)$, which is assumed to be differentiable over a nonnegative support. The variable $\theta^i$ is uncorrelated with $R_{t+1}^K$, because $z^i$ and $A_{t+1}$ are uncorrelated.

Because project payoff is borrowers' sole source of income for repayment, it is intuitive to map the contractual rate for loan $i$ (call it $R_{t+1}^i$) into a (unique) threshold value of the aggregate return $R_{t+1}^K$(call it $R_{t+1}^{Ki}$), such that $R_{t+1}^i[K_{t+1}^i + f_t^S\upsilon^S(K_{t+1}^i)] = \theta^i R_{t+1}^{Ki} K_{t+1}^i$. So, $F(R_{t+1}^{Ki})$ is the endogenous default probability of $i$. The lender's expected gross return is $F(R_{t+1}^{Ki})\theta^i K_{t+1}^i$, where $\Phi(R_{t+1}^{Ki}) \equiv [1 - F(R_{t+1}^{Ki})]R_{t+1}^{Ki} + \int_0^{R_{t+1}^{Ki}} R_{t+1}^K dF(R_{t+1}^K)$—the two terms being the expected rate of return, conditional on no default and default, respectively.

[52] Section 1.G of appendix 1 in WBF (2004) derives the exact mapping between $\theta^i$ and $z^i$ for given $A_{t+1}$: $\theta^i = [\Upsilon(z^i)^{1/\alpha} + (1 - \delta)]/[\Upsilon\kappa + (1 - \delta)]$, where $\kappa \equiv \int_{z_{\min}}^{\infty} z^{1/\alpha} K_{t+1}^i d\vartheta(z)/K_{t+1}^{NF}$ and $\Upsilon \equiv (A_{t+1})^{1/\alpha}\,\alpha[(1 - \alpha)/W_{t+1}]^{(1/\alpha - 1)}$. It is omitted here due to space constraint.

[53] Wang, Basu, and Fernald (2004) explain in detail why omitting project-specific noises in each project's realized return does not alter the model's implications for output measurement.

### Financial Intermediaries' Optimization Problem

This subsection solves banks' optimal production plan and loan interest rate. The representative bank maximizes the present value of cash flows by choosing $R^{Ki}_{t+1}$ (conditional on $K^i_{t+1}$), $N^S_t$, $N^M_t$, and $I^B_t$:

$$\text{(A2)}\quad V^B_0 = E_0\left[\sum_{t=1}^{\infty}\left(\prod_{\tau=1}^{t} R^H_\tau\right)^{-1}\left\{\int_0^{\hat{\theta}}[\theta K^i_{t+1}R^K_{t+1} - f^M_{t+1}\upsilon^M(K^i_{t+1})]dG(\theta)\right.\right.$$
$$+ \int_{\hat{\theta}}^{\infty}\theta K^i_{t+1}R^{Ki}_{t+1}dG(\theta) - \int_0^{\infty}K^i_{t+2}dG(\theta) + \int_0^{\infty}f^S_{t+1}\upsilon^S(K^i_{t+2})dG(\theta)$$
$$\left.\left.+ \int_0^{\hat{\theta}}f^M_{t+1}\upsilon^M(K^i_{t+1})dG(\theta) - W_{t+1}N^B_{t+1} - I^B_{t+1}\right\}\right],$$

subject to the constraints:

$$\text{(A3)}\qquad R^{Ki}_{t+1}(\hat{\theta}) = R^K_{t+1},$$

$$\text{(A4)}\qquad \int_0^{\infty}\upsilon^S(K^i_{t+1})dG(\theta) = A^S_t(K^S_t)^{\beta^S}(N^S_t)^{1-\beta^S},$$

$$\text{(A5)}\qquad \int_0^{\hat{\theta}}\upsilon^M(K^i_{t+1})dG(\theta) = A^M_{t+1}(K^M_{t+1})^{\beta^M}(N^M_{t+1})^{1-\beta^M},$$

$$\text{(A6)}\qquad N^B_t = N^S_t + N^M_t, \text{ and } N^M_0 = 0,$$

$$\text{(A7)}\qquad K^B_{t+1} = K^B_t(1-\delta) + I^B_t, \text{ where } K^B_t = K^S_t + K^M_t;\text{ Given } K^B_0 = K^S_0,$$

$$\text{(A8)}\qquad K^{NF}_{t+1} = K^{NF}_t(1-\delta) + I^{NF}_t, \text{ where } K^{NF}_t = \int_0^{\infty}K^i_t dG(\theta);\text{ Given } K^{NF}_0,$$

$$\text{(A9)}\qquad K^S_{t+1} + K^M_{t+1} + \int_0^{\infty}[K^i_{t+1} + f^S_t\upsilon^S(K^i_{t+1})]dG(\theta) = V^B_t.$$

Expectations in equation (A2) are taken over the distribution of $R^K_{t+1}$. The first two integrals are the overall interest (net of monitoring fees $f^M_{t+1}\upsilon^M[K^i_{t+1}]$) the bank will receive in period $t + 1$. The third integral is the productive capital the bank passes on to generation-$t + 1$ entrepreneurs after screening them. So, the sum of the three terms constitutes the cash flow of the loan division. The remaining terms form the cash flow of the bank's services division, whose implicit outputs of screening ($Y^S_{t+1}$) and monitoring ($Y^M_{t+1}$) are $\int_0^{\infty}\upsilon^S(K^i_{t+2})dG(\theta)$ and $\int_0^{\hat{\theta}}\upsilon^M(K^i_{t+1})dG(\theta)$, respectively. The variables $f^S_{t+1}$ and $f^M_{t+1}$ are the respective shadow prices, and $W_{t+1}$ is the wage rate in period $t + 1$; $N^B_{t+1}$ is the bank's total labor input, and $I^B_{t+1}$ is its total investment. Bank shareholders both pay (as debtholders of nonfinancial firms) and receive (as owners of the bank) the monitoring fees, so the two flows exactly offset each other in the bank's overall cash flow.

In equation (A3), $\hat{\theta}$ identifies the type of borrowers who are just able to pay their loan interest, given the realized $R^K_{t+1}$. Equations (A4) and (A5) are the production functions for screening in period $t$ and monitoring in period

$t + 1$, respectively. Total labor input is given in equation (A6), and $N_0^M = 0$ (and $K_0^M = 0$), given no monitoring at $t = 0$. Equations (A7) and (A8) describe the motion of the bank's and nonfinancial firms' capital, respectively.

Equation (A9) is the bank's balance sheet: the value of equity ($V_t^B$) equals the value of assets, consisted of productive capital to be used in screening ($K_{t+1}^S$) and monitoring ($K_{t+1}^M$) next period, funds [$\int_0^\infty K_{t+1}^i dG(\theta)$] transferred to borrowing firms, and this period's screening fees, which can be thought of as an intangible asset that will generate income in the next period, because it will be repaid by borrowing firms, on average.

The variable $R^H$ in equation (A2) needs elaboration. It is bank shareholders' required rate of return, equivalent to the return on total assets for a fully equity-funded bank. Thus, $R^H$ is determined by the risk profile of *total* bank cash flow, according to households' Euler equation (6). Section 7.1 has shown that $R^H$ is the weighted average of (implicit) required rates on the two partial cash flows generated by the loan division and the services division—$R^L$ and $R^{SV}$, respectively. Correspondingly, equation (A2) can be decomposed into two terms, as follows:[54]

$$\text{(A10)}\quad \mathrm{E}_0\Bigg[\left(1-\frac{K_1^B}{V_0^B}\right)\sum_{t=1}^{\infty}\left(\prod_{\tau=1}^{t}R_\tau^L\right)^{-1}\Bigg\{\int_0^{\hat{\theta}}[\theta K_{t+1}^i R_{t+1}^K - f_{t+1}^M \upsilon^M(K_{t+1}^i)]dG(\theta) + \int_{\hat{\theta}}^{\infty}\theta K_{t+1}^i R_{t+1}^i dG(\theta) - K_{t+2}^{NF}\Bigg\} + \left(\frac{K_1^B}{V_0^B}\right)\sum_{t=1}^{\infty}\left(\prod_{\tau=1}^{t}R_\tau^{SV}\right)^{-1}\Bigg\{\int_0^{\infty}f_{t+1}^S \upsilon^S(K_{t+2}^i)dG(\theta) + \int_0^{\hat{\theta}}f_{t+1}^M \upsilon^M(K_{t+1}^i)dG(\theta) - W_{t+1}N_{t+1}^B - I_{t+1}^B\Bigg\}\Bigg].$$

This partition maps into a bank's cash flow under securitization: banks receive origination fees up front and servicing fees over the lifetime of the loan pool. (See section 7.3 for more discussions.) Investors then receive the residual interest payments. This also maps into a rating agency plus a bond issue (section 7.1).

### The Determination of the Contractual Interest Rate

The loan division's optimal decision (the first component in equation [A10]) sets the contractual interest rate. It contains all the relevant cash flows—including the processing cost—for the debtholders. It expresses the condition that the interest rate charged must generate an expected return (net of the monitoring cost) equal to the ex ante rate of return required by households on their investment. This condition must hold for every loan

[54] This is under the implicit assumption that bank services are paid first, before shareholders receive the residual interest.

to avoid arbitrage. So, the optimal rate ($R^{Ki}_{t+1}$) on a loan to a generation-$t$ entrepreneur ($i$) must satisfy:[55]

$$\text{(A11)}\quad [1 - F(R^{Ki}_{t+1})]\theta^i R^i_{t+1} K^i_{t+1} + \int_0^{R^{Ki}_{t+1}} \theta^i R^K_{t+1} K^i_{t+1}\, dF(R^K_{t+1}) - E_t[f^M_{t+1}]\upsilon^M(K^i_{t+1})F(R^{Ki}_{t+1}) - R^{Li}_{t+1} f^S_t \upsilon^S(K^i_{t+1}) = R^{Li}_{t+1} K^i_{t+1}.$$

This is the key first-order condition from the bank's maximization problem set up in equations (A2) through (A9). Note that the relevant discount rate for the risky debt return is $R^{Li}$ but not $R^{Hi}$. As intuition suggests, equation (A11) implies that the higher the screening or monitoring costs, the worse the project types, and lower means of $R^K_{t+1}$ all lead to higher $R^{Ki}_{t+1}$.[56] To ensure a finite scale of operation at each firm, we assume that $R^{Ki}_{t+1}$ falls with loan size ($K^i_{t+1}$).

## Optimal Choice of Capital by Nonfinancial Firms

Entrepreneur $i$ chooses $K^i_{t+1}$ to maximize the expected utility of his residual return:[57]

$$\text{(A12)}\qquad \max E_t(U^i_{t+1}) = \max \int_{R^{Ki}_{t+1}}^{\infty} U[(R^K_{t+1} - R^{Ki}_{t+1})\theta^i K^i_{t+1}]\, dF(R^K_{t+1}),$$

subject to the constraint in equation (A11). The variable $U(.)$ is the usual concave utility function, as defined in equation (8). The first-order condition for $K^i_{t+1}$ is:

$$\text{(A13)}\quad \int_{R^i_{t+1}}^{\infty} U'(.)\left[(R^K_{t+1} - R^{Ki}_{t+1}) - \left(\frac{\partial R^{Ki}_{t+1}}{\partial K^i_{t+1}}\right)K^i_{t+1}\right]\theta^i dF(R^K_{t+1}) = 0.$$

The implicit relationship between $R^{Ki}_{t+1}$ and $K^i_{t+1}$ is represented by $\partial R^{Ki}_{t+1}/\partial K^i_{t+1}$, embedded in equation (A11). Clearly, $R^{Ki}_{t+1}$ and $K^i_{t+1}$ are jointly determined by the bank and the firm's optimization problems.

Equation (A13) makes it clear that for given $F(R^K_{t+1})$, the contractual loan rate needs to rise in the size of the loan (i.e., $\partial R^{Ki}_{t+1}/\partial K^i_{t+1} > 0$) to obtain a finite optimal $K^i_{t+1}$.[58] For individual $K^i_{t+1}$ to be determinate, an upward-sloping supply curve for funds is also necessary (i.e., the optimal solution of $K^i_{t+1}$ rises in the mean of $R^K_{t+1}$), given that firms' technology is CRS. In fact, this means that production will not happen just at the most efficient firm

[55] Note that $f^M_{t+1}$ is not known when $R^{Ki}_{t+1}$ is chosen in period $t$, and hence the expectations operator.

[56] See WBF (2004) for derivations of these and all the other comparative statics, and if relevant, the conditions under which they are obtained. None of the conditions affect the model's conclusion regarding bank output measurement.

[57] This formulation is consistent with equation (10) in the text, except that here, the entrepreneur's payoff is expressed all in terms of his or her residual return on capital, which has already accounted for the cost of labor and bank information services implicitly. Section G of appendix 1 in WBF (2004) shows the exact mapping between the two formulations.

[58] Wang, Basu, and Fernald (2004) discuss in detail the conditions under which this result arises. In general, it seems to call for more than simple processing costs. But the exact mechanism matters not for our purpose—deriving the proper output measure.

(i.e., with the highest $\theta^i$, corresponding to $\bar{z}$ in section 7.1). Instead, banks lend to a group of firms with a descending order of $\theta^i$ until aggregate capital stock is all utilized; the more efficient a firm, the larger its capital size. All else equal, the more capital available, the larger the set of firms that invest. On the other hand, given the aggregate capital stock, higher screening or monitoring cost means a larger set of firms will invest, and the efficiency level of the marginal firm will be lower.

## References

Akerlof, G. 1970. The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84 (3): 488–500.

Allen, F., and A. M. Santomero. 1998. The theory of financial intermediation. *Journal of Banking and Finance* 21 (11/12): 1461–85.

———. 1999. What do financial intermediaries do? FIC Working Paper no. 99-30-B. University of Pennsylvania, Wharton Financial Institutions Center.

Barnett, W. A. 1978. The user cost of money. *Economic Letters* 1 (2): 145–9.

Berger, A. N., and D. B. Humphrey. 1997. Efficiency of financial institutions: International survey and directions for future research. *European Journal of Operational Research* 98 (2): 175–212.

Berger, A. N., and L. J. Mester. 1997. Inside the black box: What explains differences in the efficiencies of financial institutions? *Journal of Banking and Finance* 21 (7): 895–947.

Bernanke, B. S., and M. Gertler. 1989. Agency costs, net worth, and business fluctuations. *American Economic Review* 79 (1): 14–31.

Bernanke, B. S., M. Gertler, and S. Gilchrist. 1999. The financial accelerator in a quantitative business cycle framework. In *Handbook of macroeconomics,* vol. 1C, *Handbooks in economics,* vol. 15, ed. B. S. Bernanke, M. Gertler, and S. Gilchrist, 1341–93. New York: Elsevier Science.

Cochrane, J. H. 2001. *Asset pricing.* Princeton, NJ: Princeton University Press.

Diamond, D. W. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51 (3): 393–414.

———. 1991. Monitoring and reputation: The choice between bank loans and directly placed debt. *Journal of Political Economy* 99 (4): 688–721.

Diewert, W. E. 1974. Intertemporal consumer theory and the demand for durables. *Econometrica* 42 (3): 497–516.

Fixler, D. J. 2004. Discussion of output measurement in the insurance and the banking and finance industries. In *Productivity in the U.S. services sector: New sources of economic growth,* ed. J. E. Triplett and B. P. Bosworth, 217–30. Washington, DC: Brookings Institution.

Fixler, D. J., M. B. Reinsdorf, and G. M. Smith. 2003. Measuring the services of commercial banks in the NIPAs: Changes in concepts and methods. *Survey of Current Business* 83 (9): 33–44.

Fixler, D. J., and K. D. Zieschang. 1992. User costs, shadow prices, and the real output of banks. In *Studies in income and wealth,* vol. 56, *Output measurement in the service sector,* ed. Z. Griliches, 219–45. Cambridge, MA: National Bureau of Economic Research.

Hancock, D. 1985. The financial firm: Production with monetary and nonmonetary goods. *Journal of Political Economy* 93 (5): 859–80.

Leland, H. E., and D. H. Pyle. 1977. Informational asymmetries, financial structure, and financial intermediation. *Journal of Finance* 32 (2): 371–87.

Modigliani, F. F., and M. H. Miller. 1958. The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48 (3): 261–97.

Ramakrishnan, R., and A. V. Thakor. 1984. Information reliability and a theory of financial intermediation. *Review of Economic Studies* 51 (3): 415–32.

Schreyer, P., and P. Stauffer. 2003. Financial services in national accounts: Measurement issues and progress. Paper presented at the meeting of the Organization for Economic Cooperation and Development (OECD) Task Force on Financial Services in the National Accounts, October.

Townsend, R. M. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21 (1): 265–93.

Triplett, J. E., and B. P. Bosworth. 2004a. Measuring banking and finance: Conceptual issues. In *Productivity in the U.S. services sector: New sources of economic growth,* ed. J. E. Triplett and B. P. Bosworth, 177–211. Washington, DC: Brookings Institution.

———. 2004b. Price, output, and productivity of insurance: Conceptual issues. In *Productivity in the U.S. services sector: New sources of economic growth,* ed. J. E. Triplett and B. P. Bosworth, 123–77. Washington, DC: Brookings Institution.

United Nations, Eurostat, International Monetary Fund, Organization for Economic Cooperation and Development, and World Bank. 1993. *System of National Accounts, 1993.* New York: United Nations.

U.S. Department of Labor, Bureau of Labor Statistics. 1998. Technical note on commercial banks, SIC 602: Output components and weights. Manuscript, December. Washington, DC: GPO.

U.S. Federal Reserve Board. 2009. Z.1: Flow of funds accounts of the United States. Available at http://www.federalreserve.gov/releases/z1/current/data.htm.

Wang, J. C. 2003a. Loanable funds, risk, and bank service output. Federal Reserve Bank of Boston, Working Paper no. 03-4, July. Available at http://www.bos.frb.org/economic/wp/wp2003/wp034.htm.

———. 2003b. Service output of bank holding companies in the 1990s, and the role of risk. Federal Reserve Bank of Boston, Working Paper no. 03-6, September. Available at http://www.bos.frb.org/economic/wp/wp2003/wp036.htm.

Wang, J. C., S. Basu, and J. F. Fernald. 2004. A general-equilibrium asset-pricing approach to the measurement of nominal and real bank output. Federal Reserve Bank of Boston, Working Paper no. 04-7, October. Available at http://www.bos.frb.org/economic/wp/wp2004/wp047.htm.

## Comment Paul Schreyer

### Introduction

The topic of banking output has long been a thorny issue for national accountants and analysts of banking performance and productivity. Christina Wang, Susantu Basu, and John Fernald (see chapter 7 of this volume;

Paul Schreyer is head of the National Accounts Division of the Organisation for Economic Cooperation and Development.