Basu, Susanto; Wang, J. Christina

Working Paper

Technological progress, the "User Cost of Money," and the real output of banks

Working Papers, No. 13-21

Provided in Cooperation with:
Federal Reserve Bank of Boston

Working Papers

FEDERAL RESERVE BANK OF BOSTON

No. 13-21

Technological Progress, the "User Cost of Money," and the Real Output of Banks

Susanto Basu and J. Christina Wang

Abstract:
Financial institutions provide their customers a variety of unpriced services and cover their costs through interest margins—the interest rates they receive on assets are generally higher than the rates they pay on liabilities. In particular, banks pay below-public-market interest rates on deposits while charging above-public-market rates on loans. Various authors have suggested that this situation allows one to measure the real quantity of financial services provided without explicit prices as proportional to the real stocks of financial assets held by households. We present a general-equilibrium Baumol-Tobin model where households need bank services to purchase consumption goods. Bank deposits are the single medium of exchange in the economy. The model shows that financial services are proportional to the stocks of assets only under restrictive conditions, including the assumption that either all technologies are constant or banks' technology grows at the same rate as technology in the nonfinancial economy while relative technologies of other financial institutions possibly decline. In contrast, measuring real financial output by directly counting the flow of actual services is a robust method unaffected by unbalanced technological change.

JEL Classifications: D24, E41, O47, D91

Susanto Basu is a professor in the department of economics at Boston College and a research associate at the National Bureau of Economic Research. His email address is susanto.basu@bc.edu. J. Christina Wang is a senior economist and policy advisor in the research department of the Federal Reserve Bank of Boston. Her email address is christina.wang@bos.frb.org.

This paper, which may be revised, is available on the web site of the Federal Reserve Bank of Boston at http://www.bostonfed.org/economic/wp/index.htm.

We thank Barry Bosworth, Erwin Diewert, Dennis Fixler, Robert Inklaar, Joe Peek, Marshall Reinsdorf, Kevin Stiroh, Thijs ten Raa, Jack Triplett, Frank Wykoff, and participants at a number of conferences and seminars for comments.

This version: December 31, 2013

Introduction

Accurate measurement of service output has become increasingly important for correctly measuring gross domestic product (GDP) and productivity. Services now account for nearly 60 percent of U.S. GDP, and the share continues to grow. But measuring service output, especially with adequate quality adjustment, remains challenging. Within the service sector, financial services are among the most difficult to measure, since it is not even clear how to measure *nominal* output, let alone *real* output.[1] The main reason for this is that financial firms often do not charge explicit fees for their services. Instead, they routinely earn substantial income in the form of a positive interest margin—the spread between interest received and interest paid. This measurement problem is made even more challenging nowadays by rapid and massive expansion in the range and features of financial instruments offered by financial institutions.

It is generally agreed that financial institutions provide their customers a variety of real services, and recoup their costs by earning a positive interest margin—generally higher interest rates received on assets than paid on liabilities. The most prominent case is arguably that of banks: the interest rates paid by banks on deposit balances are routinely lower than those paid on market securities with comparable risk (and the rates charged on bank loans are higher than those charged on comparable market securities). Depositors and borrowers are willing to accept these nonmarket rates because they value the services they receive. In this paper, we study the issue of measuring financial services that are priced implicitly. To make the exposition intuitive, the services we model most closely resemble banks' services to depositors. However, we emphasize that much of the paper's logic carries over to analyzing bank services to borrowers and also applies to analyzing the services of nonbank financial institutions, such as insurance companies.

When financial institutions are compensated via an interest margin, one can measure the nominal output (akin to "gross margin"[2]) of depositor services as the interest that depositors forgo by accepting an interest rate lower than the yield on market instruments with the most comparable risk profile. In other words, the nominal output can be imputed as the product of the interest rate spread and the current value of deposit balances. Various authors have gone further

[1] Triplett and Bosworth (2004, ch. 7) provide a clear summary and critique of several of the existing measures of bank output, and discuss their preferred measure.

[2] This value can be construed as the nominal value of a bank's "gross output"; that is, inclusive of compensation for intermediate inputs (for example, stationery and utilities) used in producing depositor services, but not the actual

and suggested that one can measure the *real* quantity of financial services provided (but priced implicitly) as linearly proportional to the real balance of deposits—which implies, of course, that the price index is linearly proportional to the interest rate spread. That interest rate spread (in general, many spreads, if there are a variety of monetary assets) is often termed "the user cost of money."

The literature that provides the theoretical foundation for this measurement method starts with Barnett (1978, 1980) and Donovan (1978). These papers, as well as those that follow, assume as a primitive that monetary assets enter consumers' utility function directly.[3] This assumption follows the shortcut to modeling money demand pioneered in Sidrauski's classic paper (1967). It has been clear to monetary economists from the start of that literature that the presence of money in the utility function (MIUF) is a simplified representation of a more complex reality, where money somehow aids consumers by making transactions easier.[4] But without an explicit derivation giving rise to MIUF, it remained unclear whether this shortcut could ever be rigorously justified, and if so under what conditions.

Nearly 20 years later, such a derivation was finally provided. Feenstra (1986) uses the tools of duality to show "functional equivalence" between money in the utility function and a class of general transaction cost functions in the budget constraint. Furthermore, he shows that those cost functions can be derived from a variety of money demand models in which money reduces the transaction costs of purchasing consumption goods.

Importantly, Feenstra (1986) implicitly shows that the MIUF formulation combines the consumer's primitive preferences (which depend *only* on final consumption) with a technology for making transactions, which is assumed to be a function of consumption and real (money) balances. Both the consumer's preferences and the transaction technology are assumed to have time-invariant functional forms. In the economics of measurement, it is standard to assume that preferences are stable over time. But it is definitely not common to assume that technologies are also constant. Indeed, one main objective of the measurement literature is to measure total factor productivity (TFP) growth, which is typically understood to be a measure of technology change.

---

funds borrowed and lent. Since such purchased inputs account for a tiny share in financial firms' "gross margin," we ignore them and use "output" synonymously with "value added" throughout the paper.

[3] We use the terms "consumers" and "households" interchangeably. It seems more natural to call these agents consumers, in the context of discussing consumption and utility.

[4] Not least to Sidrauski himself, who wrote of his own paper "…it is incomplete and the assumptions on which it is based are relatively crude abstractions." (1967, p. 534).

So Feenstra's derivation leads to a natural question: in this era of massive financial innovations and deregulation, is it innocuous to assume that such transaction technologies are indeed stable over time? If the transaction technology changes through technological progress in the financial sector, what are the consequences for the inspired shortcut of measuring real financial output as proportional to real balances? And if technology does vary over time, is Feenstra's functional equivalence result sufficient to guarantee that the implicitly priced real output of financial institutions will be proportional to the easily-observed real balances of assets and liabilities?

To address these issues, one needs a model where (1) there are transactions costs, (2) financial institutions provide services to reduce these costs, (3) providing services is costly, and (4) that cost is recouped via an interest margin. In addition, the model needs to be set in a general-equilibrium model in order to understand how technological changes on the firm side affect the functional equivalence result on the consumer side. We thus present a general-equilibrium model of the demand for monetary assets that follows the seminal work of Baumol (1952) and Tobin (1956). We can then easily compare our results to those of Feenstra (1986), since the Baumol-Tobin model is one of the specific cases he analyzes.

To summarize our results, we find that if either there is constant technology for all sectors, or there is constant *relative* technological progress in banks (bank technology and goods-production technology grow at the same rate) but relative regress in other financial institutions, then there will be a one-to-one relationship between the real, implicitly priced, service output of banks and the real balances of financial assets held by consumers for transaction purposes. In this case, real bank service output can be measured as proportional to real deposit balances. Unfortunately, this result is not robust. We show that there is no stable MIUF representation if financial sector technology changes at a different rate than the technology for producing goods, in which case the ratio of bank service output to deposit balances becomes unstable as well. It is therefore no longer valid to use real deposit balances to construct an index of real financial output. This means that the approach proposed in studies such as Fixler and Reinsdorf (2006), although easy to implement and thus appealing, is valid only under conditions that are likely to be too restrictive. For example, one could never use output data from the asset-based approach to calculate TFP growth in financial services, since the output measure would be valid only if the relative TFP growth in banking were zero.

Furthermore, we find that Feenstra's (1986) functional equivalence result is *not* sufficient to guarantee a stable bank-service-to-deposit ratio. We show that there are conditions under which there is a stable MIUF representation—that is, functional equivalence holds—yet the ratio of real bank services to real deposit balances varies over time. Consequently, it is not generally valid to treat money in the utility function as a primitive and hence as the theoretical foundation for using asset balances to measure financial services output. This contrasts with the general validity of taking preferences over goods as a primitive.

We suspect that, in this era of rapid and pervasive financial industry innovations, there are a multitude of reasons why the relationship between real asset balances and real financial service consumption is unlikely to be stable. Therefore, instead of using real asset balances, one should construct quantity indices for financial services using the same methods used to measure other service sector outputs in general. In particular, it should be recognized that the services underlying financial transactions are qualitatively similar to professional services such as consulting and accounting. Real quantity indices can then be constructed for precisely defined financial transactions. In fact, real output of various bank services to borrowers and depositors as traditionally measured by the Bureau of Labor Statistics (BLS) are exactly such quantity indices (see the Technical Note updated in Royster 2012).

Such a set of real quantity indices of financial services immediately implies a set of price deflators, given nominal output. We show that if the relative technology for producing financial services changes, then these deflators are unlikely to be proportional to interest rate differentials on the associated financial instruments, such as the interest rate spread between a bank's deposits and money market securities. Thus, easily observed interest rate spreads unfortunately cannot be used as price deflators for real financial services.

In the remainder of the paper, we first present our general-equilibrium Baumol-Tobin model of financial transactions in Section I. Section II uses the model to analyze the conditions under which financial balances evolve in fixed proportion to financial services, and applies the results to measurement of financial sector output. Section III contains more general reflections on this set of issues. Section IV presents concluding observations and suggests future research.

I. A General-Equilibrium Baumol-Tobin Model

In this section we study a modified version of the well-known Baumol-Tobin model to analyze the relationship between real balances of money-like assets and transaction services when transaction costs are lowered by holdings of a medium of exchange. We start with this model, whose features best resemble payment services, in order to both build intuition and compare directly with related previous studies, particularly Feenstra (1986). We then show that its conclusions apply more generally to most other financial transaction services as well.

Furthermore, since our focus is to uncover the conditions under which there is a constant as well as proportional relationship between real balances and transactions, we consider the case without uncertainty, in order to highlight the key intuition underlying those conditions. Without risk, all financial assets must offer an identical rate of return, and so the quantity of any individual class of assets is in fact indeterminate, unless additional constraints are imposed.[5] In contrast, the quantity of each type of transaction services is determined by its relationship with consumption along with its production technology, as we demonstrate below.[6]

For the same purpose of focus and tractability, we also abstract from real-world complexity concerning bank behavior to keep the nonessential features of the model simple, in particular we assume perfect competition in all markets.[7] Additionally, we focus on solutions for steady states so that the effects of changes in the technology for producing financial services can be derived analytically.[8] We begin by introducing the optimization problem faced by the four types of agents in the economy: households, goods-producing firms, market index mutual funds, and banks.

[5] In this model without risk, the interest rate spread between deposits and all other assets is trivially defined. In the real world with risk, the spread should be calculated vis-à-vis market securities with the same risk profile but offering no transaction services, which would be off-the-run (meaning securities issued before the most recent offering of same-maturity Treasuries), maturity-matched Treasuries for insured deposits in the United States, as we have shown in earlier studies (for example, Basu, Inklaar, and Wang 2011).

[6] Fischer (1983) makes a similar argument in a model of transaction services with perfect foresight.

[7] For instance, banks likely possess market power by offering differentiated services, and banks sometimes bundle transaction services with other services and subsidize the former using revenue from the latter.

[8] Previous general-equilibrium models that explicitly consider transaction costs, such as Romer (1986) and Jovanovic (1982), also focus on the steady-state solutions.

A. Consumers

To facilitate comparison, we formulate the consumer's problem similarly to Feenstra (1986), although we work in continuous time.[9] One major difference is that Feenstra implicitly treats the demand for money—the medium of exchange in general—narrowly, as a demand for currency, whereas we abstract from currency altogether and model bank deposits as the sole form of money, since currency is now used in a rather small fraction of payments and accounts for an even tinier fraction of monetary assets in a modern economy. The consumer's problem in this model thus becomes one of choosing whether to keep her assets in a mutual fund that pays a higher rate of return or to hold some assets as bank deposits, which can be used to purchase consumption goods but pay a lower interest rate. It is worth emphasizing from the beginning, however, that the lower interest rate paid by banks in our model does not stem from rents enjoyed by the banks due to their monopoly power, but from their implicit contract with consumers, which stipulates that consumers compensate the banks indirectly for the payment services via forgone interest income. This will become clear later as we explain the model setup.

Specifically, households supply labor inelastically; at the same time, they own the financial intermediaries and, indirectly, the goods-producing firm. There is a continuum of households with mass 1 (a unit measure of households). Each household, indexed by *i*, maximizes with perfect foresight the present value of discounted utility over an infinite horizon:

$$\int_0^\infty e^{-\rho t} \frac{C_{it}^{1-\sigma}}{1-\sigma} dt, \tag{1}$$

subject to the constraints:

$$\dot{A}_{it} = W_t + \Pi_{it} + r_t E_{it} + r_t^B E_{it}^B + r_t^D D_{it} - C_{it} - 1_i(t) P_t^N, \tag{2}$$

$$A_{it} = E_{it} + D_{it} + E_{it}^B, \text{ and} \tag{3}$$

$$D_{it} \geq \int_{it}^{it^{+1}} C_i(s)\, ds \text{ with } 1_i(it^{+1}) = 1 \text{ and } 1_i(it + \varepsilon) = 0, \forall \varepsilon \in [0,\ it^{+1} - it), \tag{4}$$

$$\text{given the initial values of } E_{i0},\ D_{i0}, \text{ and } E_{i0}^B. \tag{5}$$

[9] Romer (1986) also studies a continuous-time model but in an overlapping generations setting. Feenstra analyzes a number of other models of the demand for money in addition to the Baumol-Tobin model and proves functional equivalence for a generalized transactions technology, which covers the Baumol-Tobin model.

In the objective function (1), $\rho > 0$ is the discount factor and C_t is consumption.[10] We assume conventional first- and second-derivative properties for the utility function, which implies that $\sigma > 0$. The instantaneous budget constraint (2) is expressed in real terms; that is, all the interest rates are real rates and all the other variables are in units of real consumption.[11] The price of consumption is normalized to one in every period. Specifically, A_t is the (real balance of) consumers' total assets, W_t is the wage (for the one unit of labor each consumer supplies inelastically), and Π_{it} is economic profit (if any) from the ownership of the firms.[12] Equation (3) describes the composition of A_t, which comprises E_t, E_t^B, and D_t, which stand for the consumer's real holdings of the market portfolio in an index mutual fund, equity in a bank, and bank deposits, respectively. Consumers can exchange their holdings from E_t or E_t^B to D_t any time, although we further assume that consumers hold their bank stocks indirectly through mutual funds so that all exchanges must be conducted by mutual funds.[13] The real rate of return on equity held in the mutual fund is r_t, on equity in a bank is r_t^B, and the interest rate paid by banks on deposits is r_t^D. We will show shortly that $r_t^D < r_t^B = r_t$ and that the spread between the rates represents the implicit charge paid by consumers for banks' payment services.

We now discuss in detail the unusual-looking expenditure term $1(t)P_t^N$ in equation (2). P_t^N denotes the charge faced by consumers each time they transfer assets from the mutual fund to the bank. $1(t)$ is an indicator function that takes on the value 1 if the consumer transfers assets out of the mutual fund at time t, and is zero otherwise. These terms result from our assumption about the technology available for consumers to realize consumption. Specifically, we assume that all consumption must be purchased with the aid of transaction services that only banks can provide. As a normalization, we assume that one unit of consumption requires one unit of payment services. For concreteness, think of an economy where all consumption goods are purchased

[10] In the discussion we drop the *i* subscript except where doing so would cause confusion.

[11] This setup can be interpreted as implicitly assuming that all contracts in the economy are indexed for inflation, so banks pay a real rather than a nominal interest rate on deposits. Note that contracts can still be denominated in dollars, which serve as a unit of accounts. This assumption simplifies the notation, but creates no substantive change in the analysis.

[12] In this constraint and in subsequent expressions in this paper, we use the conventional notation of a dot over a variable to indicate the partial derivative with respect to time.

[13] Following Baumol and Tobin, who assume consumers must exchange outside assets, we also rule out direct exchanges of bank shares for deposits. Allowing exchanges between bank equity and deposits and charging for them as for mutual-fund-to-bank transfers will only complicate banks' optimization problem without changing any of the results qualitatively.

with debit cards, and the banks supply all the necessary processing, such as bookkeeping and transferring the appropriate sum from the buyer's account to the seller's account.

To establish the link between transaction services and bank deposits, we make the realistic assumption that banks only provide such services to consumers who maintain a balance of deposits with them. Purchases are paid out of the deposit balance. Under this arrangement, deposits constitute a form of money, which we define as any asset used as the medium of exchange. In fact, deposits are the only form of money in this model economy; we abstract from currency for simplicity.

The bundling of transaction services and deposit balances remains a standard practice in the real world, even though today it is technically feasible for a consumer to have her purchases processed without maintaining a balance of assets at the processing institution.[14] This practice is intrinsically tied to another real-world feature: that banks do not charge explicitly for many, if not most, of their services. But banks must be compensated somehow, since they have to expend real resources to process transactions.[15] For various reasons, the convention that has been adopted is that banks instead receive implicit compensation by paying depositors an interest rate lower than the rates paid on comparable-market fixed-income securities.

This prevalent practice of charging indirectly for bank services is at the core of this study, whose focus is understanding the relationship between the real quantity of services provided and the real balance of the associated assets. To this end, we make the simple but realistic assumption that banks do not impose any fees but instead recoup their cost entirely by paying a below-market rate on deposits. Therefore, the indirect service revenue can be imputed from the interest rate spread between deposits and market debt of the same risk profile. It is worth noting that, unlike the case with currency, the forgone interest on deposits is not a deadweight loss. In this model, given perfect foresight, there is a single interest rate on all market securities, which is r in equation (2), because all these securities are identical in their payoff structure. So, the equilibrium relationship between the deposit interest rate and the market rate will be $r_t^D < r_t$ (derived precisely below).

[14] For example, customers can have bills paid automatically without maintaining a deposit account with a merchant.

[15] The question of how banks should charge for their services has attracted increased interest in recent years because of the near-zero short-term market interest rates. This zero-lower-bound environment means that banks cannot adequately cover expenses incurred in providing services to depositors. And yet they have found it hard to impose explicit fees on deposit accounts. One reason particular to this episode is the public's aversion to paying banks, because of the public perception that bankers' greed caused the subprime mortgage fiasco.

Since bank deposits pay a lower interest rate than the rate paid on assets held in mutual funds, the consumer will want to minimize her holding of assets in the form of deposits.[16] On the other hand, she will not want to frequently convert small amounts of her mutual fund holdings if she faces a discrete cost for each transfer. One obvious cost is per-transfer charges levied by mutual funds, which is denoted P_t^N in equation (2). Note that, unlike banks, mutual funds charge explicitly for their services.[17] Another cost is the opportunity cost of time and attention on the consumer's part to carry out a conversion. This latter cost is likely more individual-specific, and may, for example, increase as one's opportunity cost of time, and hence income, increases. We ignore such opportunity cost for simplicity, even though it would strengthen our results.

Note that we have assumed that the size of the charge (P_t^N) is independent of the fund transfer amount. This implies that consumers in the model economy balance the flow costs of forgone interest against the fixed costs of making transfers between accounts. This is exactly the same logic as in the optimization problem for financial transactions embodied in the money demand models of Baumol (1952) and Tobin (1956). Although this assumption of flat fees is made for simplicity, adding an element of cost that depends positively on the transfer amount will not alter the analysis qualitatively. All that we need is for the cost faced by consumers not to be exactly linear in the amount of the transfer. This seems a realistic requirement, since real-world examples abound where the cost of trading depends less than linearly on the amount of assets traded.[18]

We further assume that each consumer must maintain a deposit balance sufficient to cover her purchases between two asset transfers from the mutual fund. At the same time, we rule out borrowing. This simplifying assumption allows us to focus on analyzing the relationship between a predetermined balance of deposits and the quantity of payment services. This condition is formalized in equation (4): at any instant, the consumer must hold a sufficient deposit balance to finance her total (flow of) consumption until her next asset transfer from her mutual fund

[16] In fact, this is not necessarily the unique outcome under perfect competition among banks if consumers consider the equilibrium condition that the forgone interest is not a deadweight loss, but instead just enough payment for the transaction services. Then, there can actually be a continuum of equilibria where depositors who maintain a higher balance receive a commensurately higher interest rate. But this complication is ruled out by the standard logic that each agent maximizes her own objective without taking into account market equilibrium conditions.

[17] In reality, some components of mutual fund fees are also indirect. The results here regarding banks apply to such mutual fund charges as well. For simplicity, we abstract from indirect mutual fund charges.

[18] For example, nowadays almost all retail brokerage accounts charge a flat fee for each stock trade. This is in fact a real-world example where the asset balance bears no definite relationship to the amount of transaction services.

account. This can be termed a deposit-balance-in-advance constraint, which is analogous to a cash-in-advance constraint. In our notation, $1_i(t)$ equals one at a moment when consumer i transfers mutual fund balances to the bank, and zero otherwise. t_{it}^{+1} denotes the next moment when she does so again.

B. The Optimal Solution to the Consumer's Problem

We begin by ignoring the transaction frictions and solving for optimal consumer behavior. As is well known, in the usual optimal growth model the first-order conditions for the consumer comprise a budget constraint like (2) and the Euler equation for consumption[19]:

$$\frac{\dot{C}_{it}}{C_{it}} = \frac{1}{\sigma}\left(r_t - \rho\right). \tag{6}$$

Our interest will be on the steady state, where consumption is constant. We will show that the relevant interest rate for intertemporal tradeoffs is the rate of return on equity, r. Since bank deposits pay a lower interest rate, deposits are used as a medium of exchange, but not as a store of value. In the steady state without growth, $r_t = r = \rho$, implying a constant level of optimal consumption flow, identical across consumers, because of the utility function; this will be derived explicitly later in the section on the equilibrium. That is, in the steady state, we have

$$C_{it} = \bar{C}.^{20}$$

Now we solve for the consumer's optimal use of banking and mutual fund services. In order to simplify the problem, however, we assume that consumers keep their consumption at a constant flow level between trips to the bank.[21] This is always satisfied in the steady state, at the level $\bar{C}$. Thus, in the steady state over an interval of time of length 1, if the consumer makes N transfers from the mutual fund to the bank, her average deposit balance will be $\bar{C}/2N$ because her starting balance must be sufficient to cover the total flow of consumption. The consumer's decision is then to choose an optimal number of transfers that solve the following problem (we drop subscripts to denote the steady-state value of each variable):

[19] There is also a transversality condition of the usual form, which we do not discuss further. We also do not impose a no-Ponzi-game condition explicitly, since it is unnecessary in this model that lacks borrowing.

[20] In general, we drop both the i and the t subscripts in equations to denote steady-state values. The bar over C here is to emphasize that C is constant in the steady state.

[21] This assumption is often used in dynamic versions of the Baumol-Tobin models; see, for example, Jovanovic (1982). Rotemberg (1984) studies a discrete-time version of the model. Romer (1986) presents a general-equilibrium Baumol-Tobin model where consumption varies optimally between injections of liquid assets.

$$\underset{N}{\text{Min}} \;\; P^N N + \frac{\left(r - r^D\right)\bar{C}}{2N}. \tag{7}$$

As in the standard Baumol-Tobin model, the optimal number of trips (here, transfers) minimize the total cost of fixed-fee transactions plus forgone interest on bank deposits, the average balance of which is $\bar{C}/2N$. (Here, the assumption of constant-flow consumption between asset transfers simplifies the analysis.) The first-order condition gives the well-known Baumol (1952) square-root rule regarding the optimal number of trips, denoted N^*:

$$N^* = \sqrt{\frac{\left(r - r^D\right)\bar{C}}{2P^N}}. \tag{8}$$

Since equation (8) gives the optimal number of transfers per unit of time, the amount of time between two trips to the bank is given by

$$\Delta t_1^* = \frac{1}{N^*} = \sqrt{\frac{2P^N}{\left(r - r^D\right)\bar{C}}}. \tag{9}$$

From the symmetry of the problem faced by each household, both N^* and Δt_1^* will be identical across consumers. Recall, however, that there is a continuum of households, not a single representative household. We make use of this feature and further assume that the distribution of households making asset transfers between mutual funds and banks is uniform over the interval [t, $t + \Delta t_1^*$] for all t. Thus, at each instant a mass of households of $1/\Delta t_1^*$ transfer assets in the amount $\bar{C}\Delta t_1^*$ from mutual funds to banks, and the number of transfers and the amount of each transfer are constant over time. This assumption facilitates our steady-state analysis by ensuring that both the mutual fund industry and the banking industry face a constant demand and thus provide a constant flow of services over time.[22]

We can now show that the Euler equation (6), which was derived without taking account of financial frictions, continues to hold in our model's steady state, where financial transactions are costly. Assuming that P^N, $\bar{C}$, and Δt_1^* are constant over time in the steady state, along with profits and market prices (which we will verify later), we can rewrite the budget constraint as

$$\dot{A}_{it} = W + \Pi + rA_{it} - \left[1 + \frac{P^N}{\Delta t_1^*} + \frac{r - r^D}{2}\Delta t_1^*\right]\bar{C}.$$

[22] Romer (1986) uses a continuous-time overlapping-generations model to achieve the same objective.

In the steady state, financial frictions act as a tax that boosts the effective price of consumption in excess of 1, which is the price of the consumption goods themselves. However, as long as this "tax" is levied at a constant proportional rate, the real return on equity mutual funds gives the appropriate interest rate for determining the time path of consumption. Thus, consumption will be constant at a steady state when $r = \rho$ and when the financial sector prices, P^N and $r - r^D$, are constant. Of course, the level of utility-relevant consumption, C, will be lower than in the steady state without financial frictions. One interpretation, offered by Feenstra (1986), is that the total flow of expenditure on consumption gross of any required services is the same across steady states with and without frictions, but with frictions some of the expenditure goes to purchase financial services rather than to C itself.

We now analyze the markets for the financial services expenditures that are necessary to enable the household to purchase and consume the goods and services.

C. Goods Production

There are a large number of competitive firms producing a homogeneous consumption good with constant-returns, Cobb-Douglas technology. The production function for the representative firm is:

$$Y_t = G_t \left(K_t^G \right)^{\gamma} \left(L_t^G \right)^{1-\gamma}, \tag{10}$$

where $\gamma < 1$ and K^G and L^G are capital and labor used in the production of goods. A_t is the goods technology. As usual in continuous time, the production function (10) gives the instantaneous *flow* of goods output at each point in time, given the instantaneous capital and labor inputs K and L.

The firm solves the following problem:

$$\underset{K^G, L^G}{\text{Max}} \ \Pi_t = Y_t - R_t K_t^G - W_t L_t^G$$

subject to production technology in (10), where R is the rental rate of capital. Note that the price of goods, like that of consumption, is normalized to one. The first-order conditions are

$$R_t = \gamma G_t \left(K_t^G \right)^{\gamma - 1} \left(L_t^G \right)^{1-\gamma}, \tag{11}$$

$$W_t = (1-\gamma) G_t \left(K_t^G \right)^{\gamma} \left(L_t^G \right)^{-\gamma}. \tag{12}$$

D. Mutual Funds

Mutual funds are financial intermediaries that manage all household assets other than bank shares and deposits. They are fully equity funded and own most of the capital stock in the economy, the majority of which they rent to the goods-producing firms. For concreteness, think of these as stock index funds. Specifically, we assume that they receive, on behalf of households, all the income created in this economy (that is, wages, interest, dividends, and profit) except for the income accruing to the part of the capital stock funded by bank deposits. Mutual funds invest all this income (including automatic reinvestment of all dividends and profit) and thus add to households' capital holdings. This assumption enables us to study the relationship between transactions and asset balances with only one type of inter-institution asset transfer (from mutual funds to banks); we obtain the same qualitative results with minimal complication. For the same reason, we assume that the receipt of income does not incur transaction costs, nor do any transactions related to investing in or renting out productive capital.[23] At a household's direction, the fund manager also periodically transfers some of the household's mutual fund assets to its bank. These transfers *do* incur transaction costs. Mutual funds are also price-takers in both the output and the factor markets. Thus, the representative fund maximizes the present value of profits to its shareholders:

$$\underset{I,K,L^Z}{\text{Max}} \int_0^\infty e^{-\int_0^t r_s ds}\left(R_t K_t^{GZ} + P_t^N N_t - W_t L_t^Z - I_t^{G+Z}\right)dt \tag{13}$$

subject to the constraints:

$$K_t^{G+Z} = K_t^{GZ} + K_t^Z, \tag{14}$$

$$\dot{K}_t^{G+Z} = I_t^{G+Z} - \delta K_t^{G+Z}, \tag{15}$$

$$N_t = \int_0^1 1_i(t)di, \tag{16}$$

and

$$N_t = Z_t (K_t^Z)^\gamma (L_t^Z)^{1-\gamma}. \tag{17}$$

Since mutual funds own the capital stock, K_t^{G+Z}, they also carry out the investment, I_t^{G+Z}, on behalf of households. The price of investment is the price of the final good, which is normalized to one. Equation (14) states that each fund allocates its total capital stock K_t^{G+Z} to

[23] Costless receipt of income is a customary assumption, implicit in the models of Baumol (1952) and Tobin (1956),, among others.

two uses: K_t^{GZ} is the capital rented to the goods-producing firms, and K_t^Z is used by the mutual fund in its own production. Equation (15) describes the law of motion for the overall capital stock, where δ is the depreciation rate. There is no capital adjustment cost.

N_t is the number of consumers making transfers between equity shares and deposits at time *t*, and P_t^N is the price charged for each transfer, taken as given by the fund.[24] Equation (17) characterizes the technology of the mutual fund industry for producing the fund transfer services, which is normalized to equal the number of transfers N_t. K_t^Z, and L_t^Z are the capital and labor, respectively, employed in the production process. Note that we have assumed the same Cobb-Douglas shares for capital and labor as in the goods-producing sector. This assumption simplifies analysis—making relative prices a function of only relative technology—without altering the results qualitatively.

Denote the co-state variable (that is, the shadow price) for the state variable K_t in (15) as η_t, and substitute (17) along with (14) into (13); then the Hamiltonian for this problem is written as:

$$H_t^Z = e^{-\int_0^t r_s ds}\left[R_t\left(K_t^{G+Z} - K_t^Z\right) + P_t^N Z_t (K_t^Z)^{\gamma}(L_t^Z)^{1-\gamma} - W_t L_t^Z - I_t^{G+Z}\right] + \eta_t\left(I_t^{G+Z} - \delta K_t^{G+Z}\right). \quad (18)$$

Let $\Theta_t \equiv \int_0^t r_s ds$; then maximizing H_t^Z with regard to the choice variables *I*, K^Z, and L^Z and the state variable K^{G+Z} yields:

$$-e^{-\Theta_t} + \eta_t = 0, \quad (19)$$

$$e^{-\Theta_t}\left(P_t^N \gamma N_t / K_t^Z - R_t\right) = 0, \quad (20)$$

$$e^{-\Theta_t}\left[P_t^N (1-\gamma) N_t / L_t^Z - W_t\right] = 0, \quad (21)$$

and

$$-\left(e^{-\Theta_t} R_t - \eta_t \delta\right) = d\eta_t / dt. \quad (22)$$

Maximizing H_t^Z with regard to the co-state variable η_t recovers the constraint (15).

Condition (19) states that the shadow price of capital in current value is simply one, the price of investment goods, and thus its date-zero value is the cumulative discount factor. This is intuitive in that the payoff from investing one more unit of output, which raises the capital stock by one unit, should equal its price. This is an outcome of competition and free entry, which

[24] It should be understood that (16) means that the integral (of the number of transfers) is taken over the depositors of an individual bank, which are a set of households drawn randomly from the population and distributed uniformly between zero and one.

ensures that the owners of capital—the households—receive all of the net marginal product of the capital supplied and yet there is not pure economic profit. Note that the investment is financed with all the household income received by the mutual fund on the household's behalf, net of the funds transferred to banks. When we discuss the equilibrium, we will show that this condition is satisfied.

Condition (22) describes the evolution of the shadow price of capital, which declines over time at a rate equal to the rental price net of depreciation. Using condition (19) to substitute out η_t in (22) yields the following intuitive result: the rental price on capital equals the financing cost plus the depreciation rate; that is,

$$R_t = r_t + \delta . \tag{23}$$

Conditions (20) and (21) are the analogs to conditions (11) and (12) above, measured in date-zero value. Together with the production function (17), these two conditions imply that

$$P_t^N = \frac{R_t^{\gamma} W_t^{1-\gamma}}{Z_t} . \tag{24}$$

Condition (24) is the familiar profit-maximization condition for competitive firms; that is, the mutual fund prices its services at marginal cost.[25] This is important in what follows; we discuss its implications later at length. For now, we note that P^N is not necessarily constant, and in general will change if any of the factor prices (R or W) or technology (Z) changes. This in turn will affect the number of transfers consumers choose to make and therefore the amount of deposits they choose to maintain at banks.

E. Banks

Banks receive deposits and provide payment services.[26] Banks rent out their deposits in the form of capital to the goods-producing firms, thereby earning the same rate of return on their assets—the net return on capital, *r*,—as mutual funds do on behalf of the households.

Before we state the banks' problem as we actually model it, we show that the fundamental economics of the banking sector are essentially the same as those of the mutual fund sector

[25] This also implies another familiar relationship: the relative price of mutual fund services equals the inverse ratio of its technology parameter Z relative to that for producing the numéraire good G.

[26] For brevity, we call these institutions banks, but it is equally valid to interpret them as money market mutual funds, or any other financial intermediaries that provide payment services while holding a balance of customers' assets.

discussed in Section I.D. Let *S* represent payment services (recall that the consumer needs to conduct all consumption goods purchases through banks). Then, a bank that could charge explicit fees (P^S) for its services would pay a competitive market return on its deposits. In fact, bank deposits and equity would pay the same rate of return here because the model features perfect foresight, which implies that the Modigliani-Miller theorem holds trivially. Note that, as stated earlier, all the income accruing to shareholders is assumed to be received by mutual funds on the shareholders' behalf. This removes the need to consider the exchange of bank stocks for bank deposits (to fund consumption) and the associated transaction cost.

In short, each bank would solve the following problem, which is to maximize the present value of profit to their shareholders while taking prices as given, including the interest rate on deposits determined by a competitive deposit market:

$$\underset{K,I,L^B,D}{\text{Max}} \int_0^{\infty} e^{-\int_0^t r_s^B ds} \left(R_t K_t^{GB} + P_t^S S_t + \dot{D}_t - W_t L_t^B - I_t^{G+B} - r_t^D D_t \right) dt , \tag{25}$$

subject to

$$K_t^{G+B} = K_t^{GB} + K_t^B , \tag{26}$$

$$\dot{K}_t^{G+B} = I_t^{G+B} - \delta K_t^{G+B} , \tag{27}$$

$$D_t \geq \int_0^1 \left(\int_{it}^{it^{+1}} C_{i\tau} d\tau_i \right) di , \tag{28}$$

$$\dot{D}_t = \int_0^1 1_i(t) D_{it}^T di + r_t^D D_t - C_t , \tag{29}$$

$$S_t = C_t = \int_0^1 C_{it} di , \tag{30}$$

$$S_t = B_t (K_t^B)^{\gamma} (L_t^B)^{1-\gamma} . \tag{31}$$

The setup of this hypothetical bank's problem closely resembles that of the mutual fund. K_t^{GB} is the capital that banks rent out to the goods-producing firms. K_t^B, L_t^B are the capital and labor, respectively, used in the bank's own production, characterized in (31). Equations (26) and (27) are the bank analog to equations (14) and (15) for mutual funds, describing the composition and law of motion, respectively, for the overall productive capital owned by the bank. Equation (28) simply states that the overall deposit balance at the bank is the sum over all depositors' balances, each of which must satisfy the deposit-in-advance constraint.

Equation (29) is the law of motion for the deposit balance: the increment at any instant equals the inflow of transfers from mutual funds plus accrued interest and net of the outflow of consumption expenditures. D_{it}^T denotes the amount of funds each depositor chooses to transfer from her mutual fund account to the bank each time. Because of the constant flow of consumption between any two transfers, D_{it}^T equals twice the average balance over that time interval. Since the timing of transfers from mutual funds to banks is uniformly distributed across households, we know that the deposit balance at a bank level echoes the pattern of the aggregate balance in the economy, which evolves smoothly instead of displaying the sawtooth pattern for each individual depositor.[27] Furthermore, it will be constant in a static steady state. Hence, in all later discussions about whether there is a constant ratio between bank services and deposit balances, we refer to the aggregate balance, which equals the average balance in an individual account, instead of the individual balance, at a point in time. Equation (30) states that the bank provides services (S_t) to each depositor continuously, and the amount of services is normalized to equal the total flow of consumption carried out by all its depositors at any moment.

If we substitute (26), (29), and (31) into the objective function (25), and denote the co-state variable for the state variable K_t^{G+B} as ψ_t, then we have the following Hamiltonian:

$$H_t^B = e^{-\int_0^t r_s^B ds}\left[R_t\left(K_t^{G+B}-K_t^B\right)+P_t^S B_t (K_t^B)^{\gamma}(L_t^B)^{1-\gamma}+D_t^T-W_t L_t^B-I_t^{G+B}-C_t\right]+\psi_t\left(I_t^{G+B}-\delta K_t^{G+B}\right),$$

with $D_t^T \equiv \int_0^1 1_i(t) D_{it}^T di$. Since banks take consumers' asset transfer D_t^T and consumption C_t as given, this unconventional bank faces an optimization problem identical to the mutual fund's problem. It should also choose the same solution. In particular, it should set the price of payment services equal to the marginal cost of production:

$$P_t^S = \frac{R_t^{\gamma} W_t^{1-\gamma}}{B_t}. \tag{32}$$

Likewise, the rate of return on bank equity must satisfy the following relationship:

$$r_t^B = R_t - \delta = r_t. \tag{33}$$

Recall that r_t is the rate of return on mutual fund shares. This result is intuitive: without uncertainty, all assets held only for pecuniary purpose must offer the same rate of return.

[27] It should be understood that a single representative bank's average balance can be any finite fraction of the aggregate balance; it is in fact indeterminate because of the constant-returns-to-scale banking technology.

Similarly, the rate paid on deposits r^D will be determined by consumers' intertemporal Euler equation and competition among banks. This should equal r_t as well, since banks earn r_t on their capital and pay shareholders r_t.

With deposits paying the same rate of return as other assets, households will want to hold all their wealth in deposits because this saves them the cost of transferring assets from mutual funds to banks. Since they are endowed with bank shares E_0^B and mutual fund shares E_0 initially, this means they will choose not to increase the holding of either but instead will accumulate all wealth in the form of deposits. To the extent that the deposit balance D_{it} is sufficient to cover C_{it} at any instant [meaning (28) is satisfied], households can avoid the fund transfer charges entirely. Note that returns on E_0^B and E_0 are used to (partly) fund investment, which is assumed to incur no transaction cost.

In this case, measurement of bank output would be easy. It is clear that nominal output is $P^S S$ and real output is S, while $r^D D$ is a transfer of asset income to consumers according to the United Nations System of National Accounts. Note that the direct way to measure output would be to count the number of transactions processed, S. The more commonly used method is to construct a deflator, P^S, and use it to deflate the total service charge. These two approaches are trivially equivalent in this simplest case with a single type of bank payment services. In the real world, however, the latter is more practicable for constructing an aggregate index of a disparate variety of transaction services. In either case, there is generally no fixed proportional relationship between S and D. In the equilibrium for the above optimization problem for banks, we would have $S = C$ as usual because of the normalization made in Section I.A, and we would also have D be a fixed fraction of total K in the economy. But there is no reason why the ratio between consumption (C) and capital (K) should be constant outside the steady state or across different steady states.

However, as we know, actual institutions differ from this straightforward arrangement. Specifically, banks generally do not pay depositors the full return on their assets. For certain types of accounts, such as demand deposits, this is in part stipulated by regulations (such as Regulation Q) prohibiting or restricting the payment of interest. Consequently, banks have to remunerate depositors in kind, by providing services (S) "for free." The two parties essentially strike a barter agreement. More broadly, the prevalence of the banking practice of paying depositors a below-market rate is at least partly a vestige of optimal choices made subject to

previously relevant constraints. One important category of constraint is technology. For example, communication used to be more costly, even prohibitively so, and thus the only way to know whether a customer had enough wealth to afford his purchases was to hold the necessary funds directly. Then it would be natural to deduct the service charges directly from the balances.

We incorporate this real-world feature into our model; that is, we assume that banks charge for services implicitly by offering a below-market rate of return on deposits. Furthermore, for simplicity and without any loss of intuition, we study the polar case where banks levy no fees at all but instead are compensated for all productive services via the interest rate spread.[28] Under this assumption, and applying the result (explained earlier) that banks have a constant level of deposits $\bar{D}_t$ because households uniformly stagger their transfers over time, the representative bank solves the following optimization problem:

$$\underset{K,I,L^B,D}{\text{Max}} \int_0^{\infty} e^{-\int_0^t r_s^B ds} \left(R_t K_t^{GB} + \dot{D}_t - W_t L_t^B - I_t^{G+B} - r_t^D \bar{D}_t \right) dt , \tag{34}$$

subject to the same set of constraints as characterized by (26) and (31).

From the point of view of individual banks, $\bar{C}$ and N^*, and hence $\bar{D}_t = \bar{C}/2N_t^*$, are exogenous variables that they take as given. Combined with perfect competition, this implies that banks must set the rate on deposits to be just low enough so that they recoup the cost of providing payment services needed for consumption: $\left(r_t - r_t^D \right) \bar{D}_t = \hat{P}_t^S \bar{C}$, where we have inserted the equilibrium condition that the net return on renting out capital equals r_t, and $\hat{P}_t^S$ denotes the shadow price of the payment services, which equals the marginal cost [as shown in (32)]. Solving for $\bar{D}_t$ then yields:

$$r_t^D = r_t - \frac{2N_t^* R_t^{\gamma} W_t^{1-\gamma}}{B_t} . \tag{35}$$

As with the mutual funds, equation (35) should be interpreted as an equilibrium condition ensuring that banks make zero profits.

Note that the interest rate paid on deposits is declining in the number of transfers to the bank, N^*. The reason is that, for a given level of consumption, a larger number of transfers

[28] We ignore the complication that the equilibrium deposit rate may be negative but is constrained by the zero lower bound for psychological reasons. We also assume away the possibility that the equilibrium deposit rate may exceed the regulatory rate cap since we do not model such regulatory constraints explicitly. The latter constraint is unlikely to bind in the low-rate environment prevailing over the last decade or so, whereas the former is more likely to do so.

implies a lower average deposit balance, and thus lower bank income. But if the level of consumption is fixed, then so are bank operating expenses. On net, therefore, the bank has lower net income to distribute back to depositors in the form of interest payments.

F. The Optimal Number of Transfers N^*

Now we solve for N^*, the optimal number of transfers between a mutual fund and a bank account. Recall that we have shown in the household's optimization problem that, in steady state, the optimal number of trips follows the familiar Baumol (1952) square-root rule (the variables without time subscripts denote their steady-state values):

$$N^* = \sqrt{\frac{\left(r - r^D\right)C}{2P^N}}\,.$$

Substituting in the equilibrium conditions for $r - r^D$ from (35) and P^N from (24) gives:

$$N^* = \sqrt{\frac{\frac{2NR^\gamma W^{1-\gamma}}{B}C}{2\frac{R^\gamma W^{1-\gamma}}{Z}}} \Rightarrow N^* = \frac{Z}{B}C\,. \tag{36}$$

Thus, the optimal number of mutual fund transfers carried out by the consumer will depend linearly on both consumption (C) and relative financial technology—the technology for making transfers relative to that for making payments (that is, Z/B). This last result is a deviation from the usual square-root formula because of the conventional bank practice adopted in our model, so in equilibrium $r - r^D$ depends negatively on N, via the dependence of the average deposit balance on N.

G. Equilibrium and Steady State

We begin by solving the consumer's optimization problem, using some of the results derived in previous subsections. Substituting the optimal choices for P^N, $r - r^D$, and N from equations (24), (35), and (36) as well as the "deposit in advance" constraint (3) into the budget constraint (2) yields:

$$C_t\left(1 + \frac{2R_t^\gamma W_t^{1-\gamma}}{B_t}\right) + \dot{E}_t + \dot{E}_t^B + \dot{D}_t = W_t + \Pi_t + r_t\left(E_t + E_t^B + D_t\right), \tag{37}$$

with the i subscript omitted since all consumers are identical.

This simpler budget constraint says that from the consumer's point of view, the need for financial services to purchase consumption goods acts as a tax on consumption. Thus, the effective price of consumption is not just the price of output, which is normalized to one, but also includes the prices of financial services, which are a function of the financial sector's technology and factor prices. Note that the factor prices in turn depend on the goods-producing technology G_t. It is easy to confirm that in equilibrium the expenditure on mutual funds' transfer services equals the expenditure on banks' payment services, explaining why the "tax" on each unit of consumption equals *twice* the (shadow) price of a unit of banking services.

Using the simplified budget constraint, we now solve for the intra- and intertemporal first-order conditions characterizing optimal consumption behavior:

$$U'(C_t) = \lambda_t \left(1 + \frac{2R_t^{\gamma} W_t^{1-\gamma}}{B_t}\right), \text{ and} \tag{38}$$

$$\dot{\lambda}_t / \lambda_t = \rho - r_t . \tag{39}$$

λ_t is the current-value co-state variable for the state variable of total assets in the Hamiltonian. It has the usual interpretation of representing the marginal utility of wealth. However, the marginal utility of consumption exceeds λ_t because consumption also entails expenditures on financial services. On the other hand, its rate of change over time still follows the usual rule, equal to the difference between the subjective discount rate and the net return on capital.

Market-clearing conditions for capital and labor, respectively, are as follows:

$$L_t^G + L_t^B + L_t^Z \equiv L_t = 1, \tag{40}$$

$$K_t^G = K_t^{GZ} + K_t^{GB}, \text{ and } K_t^G + K_t^Z + K_t^B \equiv K_t = D_t + E_t + E_t^B . \tag{41}$$

Recall that the deposit balance D_t for the economy as a whole does not vary over time, barring changes in any of the parameters, even though the balance in each individual deposit account follows a sawtooth pattern. This is because the time of fund transfer is staggered uniformly across consumers so that at any instant a fixed fraction of consumers deposit funds in their bank accounts. Moreover, total inflow into deposits exactly offsets total outflow, which equals the constant flow of consumption *C*. This ensures that the capital stock used in goods production, K_t^G , also remains constant.

The overall deposit balance D_t equals the average balance maintained in each deposit account between any two fund transfers because of averaging across consumers. Given a constant flow of consumption (such as in the steady state), the average balance equals half the amount transferred into a deposit account each time. In equilibrium, the amount deposited each time is just enough to cover the consumer's consumption until the next transfer balance; that is, constraint (4) holds with equality. Combining this with consumers' optimal choice of the number of transfers to conduct within each unit of time, described in equation (8), we obtain the overall balance of deposits as follows:

$$D_t = \frac{\bar{C}_t}{2N_t^*} = \sqrt{\frac{\bar{C}_t P_t^N}{2\left(r_t - r_t^D\right)}}. \tag{42}$$

The rest of the K_t is funded with E_t and E_t^B. The ratio between them is in fact indeterminate because of the absence of uncertainty. We can pin down E_t^B with a capital requirement or assume that they evolve proportionally to their initial amounts, E_0 and E_0^B, respectively.

Perfect competition in goods and factor markets ensures zero profit:

$$\Pi_t = 0, \quad \forall t.$$

Combining first-order conditions (11) and (23) yields the condition for the real interest rate:

$$r_t = \gamma G_t \left(K_t^G\right)^{\gamma-1} \left(L_t^G\right)^{1-\gamma} - \delta. \tag{43}$$

Capital follows the following law of motion:

$$\dot{K}_t = I_t - \delta K_t. \tag{44}$$

Having described the equilibrium, we now characterize the steady state. Assume (for now) that all technologies are constant. Under this assumption (and without population growth), the steady state of this model implies zero growth in all aggregate and per-capita variables.[29] We now solve for the equations characterizing the steady state. In what follows, we again use variables without time subscripts to denote steady-state values.

Inserting the definition of the real interest rate (43) into the Euler equation (39) and imposing the condition that the marginal utility of wealth stays constant in the steady state produces:

$$\gamma G\left(K^G\right)^{\gamma-1}\left(L^G\right)^{1-\gamma}-\delta=r=\rho\Rightarrow\frac{K^G}{L^G}=\left(\frac{\rho+\delta}{\gamma G}\right)^{1/(\gamma-1)}\equiv k\left(G,\gamma,\delta,\rho\right). \qquad (45)$$

Note that this equation pins down the capital-labor ratio in goods production as a function of technology (G, γ, δ) and preferences (ρ). This ratio is constant in the static steady state. Otherwise, it would grow at the same rate as $G^{1/(1-\gamma)}$, which is the technology expressed in labor-augmenting form.

Since all sectors have the same capital and labor shares, they must also have the same capital-labor ratio, which is a function of just the goods-producing technology parameter G, because the goods are the numéraire in this economy:

$$\frac{K^G}{L^G}=\frac{K^B}{L^B}=\frac{K^Z}{L^Z}=k\left(G,\gamma,\delta,\rho\right). \qquad (46)$$

Goods output must be used for consumption or investment. The law of motion (44) implies that investment must just replace depreciation in the steady state in order to maintain a constant capital stock. Therefore, steady-state goods production equals consumption plus depreciation:

$$C+\delta K=G\left(K^G\right)^{\gamma}\left(L^G\right)^{1-\gamma}. \qquad (47)$$

From the results in the previous sections, we also know that

$$S=C=B\left(K^B\right)^{\gamma}\left(L^B\right)^{1-\gamma}, \qquad (48)$$

and

$$N=C\frac{Z}{B}=Z\left(K^Z\right)^{\gamma}\left(L^Z\right)^{1-\gamma}. \qquad (49)$$

Equations (40)–(43) and (46)–(49) comprise 12 equations [expression (46) actually contains three equations, while (48) and (49) each contain two equations] that determine 12 endogenous variables: C, K, K^G, L^G, K^Z, L^Z, K^B, L^B, D, r (hence the rental price R), S, and N. Equations (12), (24), and (35) then determine the remaining price variables—the wage rate W, the price of fund transfer services P^N, and the deposit interest rate r^D, respectively. The exact solutions can be found in the appendix.

[29] The logic can be easily understood from the fact that the marginal product of capital is diminishing: given a constant technology with enough capital, the marginal product of capital (net of depreciation) would become negative, which surely cannot be a choice that optimizing households would make.

II. "Money" in the Utility Function and the Technology for Producing Financial Services

In this section, we show first that, under one strong assumption, our model is able to rigorously justify a setup where consumers are assumed to derive direct utility benefits from bank deposits. That is, Feenstra's (1986) celebrated "functional equivalence" theorem is nested as one restricted version of our model. The restriction is that the technology for producing mutual fund services (asset transfers, in our model) must be constant.

This result is interesting, for two reasons. First, the money-in-the-utility-function setup is used to justify measuring the services of the financial system as proportional to the stock of assets (money) held by the system. Money is defined as an asset that is dominated in financial rate of return, so it corresponds to bank deposits in our model. Thus, it is interesting to see whether there are any conditions where functional equivalence holds in our model with explicit technologies for making financial transactions. The second reason is that according to our findings functional equivalence is *not* a sufficient condition for bank deposits to proxy for the intermediation services produced by banks. This result is problematic for a long literature that takes the holdings of money, or financial assets generally, as a proxy for real liquidity or intermediation services produced by the financial system.

Finally, we show that in our model a necessary and sufficient condition for deposits to proxy for real bank (or financial sector) output is that the technologies for making transactions by banks—but not by mutual funds—must be constant over time relative to the goods-producing technology. In other words, there must be relative technological progress in the banking sector relative to the mutual fund sector, unless all technologies stay constant. Given the huge relative changes in financial sector technology that have taken place over the last few decades, especially outside traditional banks, we view this result as a negative one for using bank deposits to proxy for bank service output.

A. A Rigorous Foundation for "Money" in the Utility Function

If we express N as $N = \bar{C}/2\bar{D}$ using the relationship between the (constant) flow of consumption ($\bar{C}$) and the average balance of deposits ($\bar{D}$) between any two fund transfers and substitute it into a consumer's expenditure on mutual fund services $P^N N$, we can rewrite the budget constraint as:

$$C_t + P_t^N \frac{C_t}{2\bar{D}_t} + \dot{E}_t + \dot{E}_t^B + \dot{D}_t = W_t + \Pi_t + r_t\left(E_t + E_t^B\right) + r_t^D \bar{D}_t . \qquad (50)$$

Now suppose, for this subsection only, that P_t^N is constant for all *t*, and denote that value as θ. This condition arises as long as the technology for making transfers grows at the same rate, including zero, as the goods-producing technology (which will be referred to as the numéraire technology), as can be seen easily from equation (A.5) in the appendix. The original optimization problem for the consumer with this assumption can be expressed as maximizing the original objective function (1):

$$\int_0^\infty e^{-\rho t} \frac{C_{it}^{1-\sigma}}{1-\sigma} dt ,$$

subject to:

$$C_t + \theta \frac{C_t}{2\bar{D}_t} + \dot{E}_t + \dot{E}_t^B + \dot{D}_t = W_t + \Pi_t + r_t\left(E_t + E_t^B\right) + r_t^D \bar{D}_t , \qquad (51)$$

where we have also substituted identity (3) into the budget constraint (2).

Now define a variable *X* as

$$X_t \equiv C_t + \theta \frac{C_t}{2\bar{D}_t} \equiv C_t + \phi(C_t, \bar{D}_t) . \qquad (52)$$

In Feenstra (1986), as in earlier transaction cost models, $\phi(C_t, \bar{D}_t)$ is referred to as a "liquidity cost"—the cost consumers incur because they must first exchange their wealth into a liquid asset that pays a lower rate of return than bonds or equity and use it to procure consumption. Feenstra (1986) therefore interprets *X* as "gross consumption"; that is, all expenditures related to consuming, which in that model equal the sum of expenditures on consumption proper plus those on "liquidity services."

Note, however, there is a qualitative distinction between this so-called liquidity cost $\phi(.,.)$ and the actual bank services in our model. Here, $\phi(C_t, \bar{D}_t)$ in fact comprises only a household's real expenditures on mutual fund transfer services but not its real expenditures on bank services, $\left(r_t - r_t^D\right)\bar{D}_t$. Thus, *X* as defined in (52) is just the sum of actual consumption and expenditures on mutual fund services; it does not include expenditures on bank services and therefore is not a

complete measure of gross consumption in this model.[30] The intuition for this result is simple: the liquidity cost $\phi(.,.)$ in other models contains the explicit expenditure needed to realize consumption, *other than* the interest income households forgo by holding money. In our model, the liquidity cost corresponds to mutual fund transfer fees while the forgone interest is in fact implicit payment for bank services. This observation is also key to understanding our result that functional equivalence is not sufficient to guarantee that financial services can be measured by financial balances. This is so because, even if functional equivalence holds, it will only guarantee a fixed inverse relationship between deposit balances and *mutual fund* services (given the level of consumption); it will not guarantee a fixed ratio between deposits and *bank* services.

Now, consider an alternative problem for the consumer, which is to maximize

$$\int_0^\infty e^{-\rho t} V\left(X_t, \bar{D}_t\right) dt, \tag{53}$$

subject to:

$$X_t + \dot{E}_t + \dot{E}_t^B + \dot{D}_t = W_t + \Pi_t + r_t\left(E_t + E_t^B\right) + r_t^D \bar{D}_t. \tag{54}$$

Constraints (54) and (51) are clearly equivalent. So the original optimization problem defined by (1) and (51) will be equivalent to the second problem defined by (53) and (54); that is, they differ only in their functional notations if an indirect utility function V exists such that

$$U\left(C_t\right) \equiv V\left(X_t, \bar{D}_t\right).$$

Note that real expenditures on bank services in this model are indirectly captured by the presence of $\bar{D}_t$ in the alternative utility function V and the fact that deposits offer a return $r_t^D < r_t$.

Proposition: If P^N is constant, then problems (1)–(51) and (53)–(54) are equivalent.

Proof: The original problem (1)–(51) satisfies Feenstra's (1986) assumption 1. Thus, the proof follows directly from Feenstra's (1986) proposition 1. Furthermore, the function V so defined has a variety of useful and intuitive properties, which follow from Feenstra's (1986) assumption 2.[31]

[30] The equivalent object to expenditure on bank services in Feenstra's model is the nominal interest cost of holding money. However, in our model, that expenditure corresponds to the production of real output using capital and labor.

[31] Although Feenstra's problem is formulated in discrete time and ours is set in continuous time, his proof shows equivalence at each point in time and has no intertemporal dimension. Thus, his proof applies to our setting as well.

Notice that the derivation of the reformulated problem (53)–(54) justifies rigorously a representation of the consumers' problem that has real bank deposit balances ("money balances") in consumers' (indirect) utility functions. Money-in-the-utility-function is the starting assumption of Barnett (1978, 1980), who is commonly credited with establishing the theoretical foundation for the "user-cost-of-money" approach to measuring financial service output. In this approach, quantity is measured using an index of the real balances of assets, and the price thus corresponds to the relevant interest rate spreads.[32] Nominal output of the financial services sector then simply equals the product of the nominal balance and the corresponding interest rate spread (for example, forgone interest for depositors).

B. Additional Conditions Needed for a Fixed Ratio between Bank Services and Balances

Given the importance of this approach in the literature, it is crucial to note the following result: our model in fact shows that functional equivalence is *not* enough to establish a time-invariant relationship between the real service output of banks and the real level of deposits. In our model, for bank services *S* to be in fixed proportion to deposit balances *D* is equivalent to having each consumer make a constant number of fund transfers within each unit of time. This is easily seen by noting that $N = \bar{C}/2\bar{D}$. Then the sufficient conditions for deposits to move in a one-to-one ratio with the real output of bank services are the same as those for *N* to be constant. These can be classified into two cases: the first, one might call the static case, where there is no technology change in any sector, while the second case is a dynamic one that allows for positive technological growth in the goods-producing sector.

In the static case, *all* technologies—for producing bank services, mutual fund transfers, and goods (*B, Z,* and *G*)—are constant and scalar multiples of one another. Of course this means that both financial technologies are constant relative to the numéraire technology. Then all factor prices are also constant. Under these conditions, there is a constant ratio between bank services *S* and deposit balances *D.* This can be shown trivially by inspecting the steady state solution for *N*, which is equation (A.4) in the appendix and expanded below:

$$N = Z\frac{Gk^{\gamma} - \delta k}{B + 2G} = zG\frac{\kappa G^{1/(1-\gamma)}}{(b+2)G} = \frac{z\kappa}{b+2}G^{1/(1-\gamma)},$$

[32] See, for example, Hancock (1985) and Fixler and Zeischang (1992).

where $Z = zG$, $B = bG$ in a static economy, and $\kappa \equiv \left(\gamma/\rho+\delta\right)^{\gamma/(1-\gamma)} - \delta\left(\gamma/\rho+\delta\right)^{1/(1-\gamma)}$.

If the above conditions for a constant ratio between bank services and balances are satisfied, then one can infer bank service output, *S*, from the quantity of deposits, *D*. One can, in turn construct an index of bank services as proportional to the average deposit balance $\bar{D}$ over time. The price index for bank services will then be proportional to the rate spread $r - r^D$ multiplied by the price index for consumption goods, exactly as in Fixler and Reinsdorf (2006).[33]

While one would obtain equivalent measures of bank services using either an index based on actual counting of *S* or an index based on real balances, $\bar{D}$, it is clearly much easier to simply obtain deposit balances and interest rate spreads. Thus, if Fixler and Reinsdorf's (2006) procedure can be applied generally, it is preferable on practical grounds to the procedures suggested by Wang, Basu, and Fernald (2009).

C. "Money" in the Utility Function with Time-Varying Technologies

Now drop the assumptions that all technologies are always constant and scalar multiples of one another. Equivalently, consider situations where the relative prices, P^N, the price of mutual fund transfers, and the shadow price of bank services, $\hat{P}^S$, are not constant over time. After all, in Section I we showed that these prices depend on the technologies in the two financial services sectors and on economy-wide real wages and rental prices [see equations (24) and (31)]. There is no economic reason why the growth rate of either technology must bear a specific relationship to the rate of growth of the technology for producing goods. Feenstra (1986), however, does not consider possible changes in transaction technologies. Normalizing the numéraire technology to be constant, we can show the consequences for measuring bank output if financial sector technologies vary relatively over time.

Suppose first that the economy was initially in the steady state. Now suppose that *Z*, mutual funds' technology for producing asset transfer services, doubles once and for all. It is easily seen from equations (A.3) through (A.6) in the appendix that, in the new steady state, the price for

[33] It is easy to identify "the" interest rate spread in our nonstochastic environment, where there is only one nonbank interest rate. In a stochastic environment where there are many interest rates, Wang (2003a) and Wang, Basu, and Fernald (2009) show that the reference rate (here, *r*) needs to be corrected for risk.

fund transfers (P^N) halves and the optimal number of transfers doubles, while optimal consumption remains the same.[34]

The proof follows from inspecting equations (47)–(49) in conjunction with (36). Suppose that total capital, *K*, consumption, *C*, and the distribution of capital and labor across sectors remain unchanged even as *Z* doubles. Clearly, equations (47)–(48) are satisfied, since no exogenous variables have changed in those equations. With unchanged inputs of *K* and *L* in the mutual fund sector but with technology that is twice as productive, the output of that sector (*N*) must double as well. But this is exactly what is implied by the optimal demand for transfers, (36). Therefore, the initial conjecture of unchanged *K* and *C* in the new steady state is verified. It then implies that the average deposit balance, $\bar{D}$, will be halved, since $\bar{D} = \bar{C}/2N$.

Now consider the proposal to construct an index of bank service output by equating the growth rate of services to the growth rate of real deposit balances. Since *C* remains the same, bank services, *S*, have to remain constant as well. Measuring bank service output directly, by counting the number of transactions, would reveal this fact. But constructing an index of real output as proportional to the real balance of deposits would show—incorrectly—that real service output had been cut in half.

Second, consider changes in bank technology, *B*, holding both *Z* and *G* constant. For example, suppose that *B* increases once and for all. Then, in the new steady state, equation (A.4) in the appendix shows that the optimal number of transfers N^* would fall. But, by (42), a decline in N^* means that the ratio *S/D* would fall. Hence, the time path of deposits would not give a correct index of the banks' real service output in this case, either. Note that this is true even when all the conditions of our proposition are satisfied. That is, deposits are not necessarily a correct index of bank service production even in a case where functional equivalence holds.

More generally, the optimal solution for *N* also makes clear that when the goods technology (*G*) grows at a positive rate, the technology for bank services (*B*) must grow at the same rate, whereas the technology for a mutual fund to make transfers (*Z*) must *decline* relative to *G*. So the common condition for a constant ratio between bank services and deposit balances, whether *G* is static or is growing over time, is that *B* relative to *G* must remain constant. That is,

[34] In this model, the general-equilibrium analysis of the new steady state gives the same results as analyzing the change in P^N and *N* in partial equilibrium, holding consumption constant.

$$Z = zG^{-\gamma/(\gamma-1)} \text{ and } N = Z\frac{Gk^{\gamma} - \delta k}{B+2G} = zG^{-\gamma/(1-\gamma)}\frac{\kappa G^{1/(1-\gamma)}}{(b+2)G} = \frac{z\kappa}{b+2}.$$

The intuition is that, to minimize the overall transaction cost, consumers optimally choose to spend the same amount on mutual fund transfers and bank payment services. Then, for *N* to stay constant, the per-transfer fee P^N, which is inversely proportional to *Z*/*G*, needs to grow at the rate of consumption and bank services *S*. These latter quantities, and in turn deposits *D*, grow at the rate $1/(1-\gamma)$. The interest rate spread $r - r^D$ remains constant. Hence, the expenditure on mutual fund transfers stays the same as that on bank services, and both grow at the rate $1/(1-\gamma)$.

Contrasting the condition of P^N rising at a particular rate with the condition for functional equivalence, which is a constant P^N, it is obvious that only in the static case is functional equivalence sufficient for a constant ratio between bank services and deposit balances. Whenever there is positive growth in the technology for producing goods, the conditions needed for a constant services-balances (*S*/*D*) ratio are more stringent than those needed for a money-in-utility-function formulation. Hence, functional equivalence is generally insufficient to ensure constant proportionality between the flow of services and the stock of deposits.

Our analysis also points out that, even in our deliberately simple setup, changes in relative financial sector technologies are only one reason why the *S/D* ratio may change over time. In fact, it is clear from equation (36) that, even in partial equilibrium, only proportional changes in both mutual fund and bank technologies (that is, Z_t and B_t, respectively) will leave N^* and in turn $\bar{D}$ unchanged. Otherwise, changes in either Z_t or B_t alone have symmetric but opposite effects on N^* and in turn on $\bar{D}$. Furthermore, it is easy to see that changes in P^N and $\hat{P}^S$ due to other exogenous factors (with regard to the model), such as changes in sector-specific value-added taxes, or even changing capital market regulations, can break down even the equivalence as well. These are quite realistic conditions. We will elaborate on them later when we discuss the implications for measurement, and so we just note here that a constant ratio between service output and real balances is most likely a rare coincidence in real-world situations.

III. Implications for Measurement

We have shown that in realistic cases—changes in relative bank payment technologies—the proposal to measure implicit financial sector output as proportional to financial balances will fail

to give accurate results. This failure can occur even when functional equivalence holds and there exists an indirect utility function of consumption and real balances. Moreover, even when banks' relative technology remains constant, relative technological change in nonbank financial institutions must occur at a specific *negative* rate to obtain a constant ratio between bank services and balances. In our model, banks constitute the part of the financial sector that actually performs the transactions needed for households to consume, while nonbank financial institutions only transfer assets to banks. Our example, however, demonstrates only one mechanism, albeit arguably the most relevant, through which a constant service-to-balance ratio breaks down.

In fact, one real-world condition in particular may have evolved in the opposite direction to what is required for a constant services-balances ratio: the price for nonbank financial services, P^N in the model, has likely fallen over time instead of rising relative to the (shadow) price of bank payment services. We can infer that the relative price of nonbank services has fallen from the fact that consumers are making an increasing fraction of their payments through nonbank accounts, especially at money market mutual funds. Likewise, the cost of securities trading, another price of nonbank services, has also fallen. The cost of trading market securities can fall either because the technologies for order execution, clearing, and settlement improve, or because better capital market regulations mitigate the costs stemming from asymmetric information problems.

Another realistic situation in which both functional equivalence and a constant ratio between services and balances would fail to hold is if the model were extended to include currency as an asset used in making payments and the inflation rate varied over time. In fact, as noted by Feenstra (1986) himself, only by treating inflation as a tax levied at the end of each period does one obtain a stable transaction cost function from the Baumol-Tobin model. Otherwise, transaction costs will vary with the inflation rate.

Finally, even the small modeling change of allowing the mutual fund and banking sectors to have different capital shares of production would mean that the service-to-deposit ratio would change over time as factor prices (and hence relative marginal costs of the two sectors) change, not just if either technology changes.

The general conclusion is that there exists no fundamental theory stipulating any definite relationship between the quantity of service output and the real balance of financial assets. Any

mapping between service flow and asset balance depends entirely on features of the transaction technology and its relation to other technologies in each specific model.

Not only can one easily develop models where transaction services are wholly separate from any financial assets, but real-world examples of such services abound. Fischer (1983) makes exactly the same argument "…[it is] possible to imagine institutions that make transactions without requiring any corresponding asset holding. The postal giro system is the most important example. A company making C.O.D. deliveries is another (p. 6)." More such cases emerge exist today, such as web portals (www.mvelopes.com, for example) that make payments to any account a customer designates (bank deposit, mutual fund, utilities, etc.), as well as goods- and services-providing companies that offer their own online payment option.

The bottom line, therefore, is that using the real balance of financial assets to measure financial service output is most likely a reduced-form approximation at best. It should be utilized as a last resort if the data to construct real indices of transactions do not exist and cannot be collected. It should be regarded as the exception, but not the norm.

The alternative we have proposed, as in Wang, Basu, and Fernald (2009), is in fact a return to traditional practice, at least in the BLS productivity group (see Royster, 2012). It is to construct indices of real service output based on counts of actual transactions—for example, checks cleared, ATM transactions processed, and mortgage applications screened. The individual indices of financial service output should then be aggregated using the nominal values as weights, as advocated and implemented in Wang (2003b) and Basu, Inklaar, and Wang (2011). The BLS has since adopted this aggregation method (Royster 2012), replacing its previous practice of using (rather dated) Functional Cost Analysis data to weight each output component by its share in total labor cost.

Two major difficulties, mostly practical, still confront our preferred approach to measuring the flow of financial services by directly counting transactions. One is to define the individual categories of services sufficiently finely so that each category covers homogeneous transactions. The other is to adjust the quantity and price indices for changes in service quality over time. Both difficulties arise primarily because of the dearth of direct data on financial transactions. Inklaar and Wang (2012) introduce some promising data sources that had not been utilized previously and estimate their implications for measured bank output. More generally, in our previous studies, we have repeatedly advocated collecting more and better data.

One important conceptual advantage of the BLS-type method is that it yields consistent measures of both implicitly and explicitly priced financial services. While our focus has been to derive the theory for decomposing nominal output of implicit service output into a price and a quantity component, exactly the same logic applies to decomposing explicit fees into the two components. In fact, our output measure makes even more intuitive sense when applied to services that generate explicit revenue; such services are more likely to be separate from any asset holding. With or without an associated asset balance, our method calls for measuring financial services just as we would any other service: define each type of transaction clearly, and obtain a quality adjusted quantity index of the transactions (such as the number of conforming residential mortgage loans screened). The combination of nominal and real output then implies a price deflator.

By comparison, when one follows the current implementation of the user-cost approach and measures implicit services output using the real balance of the associated financial instruments, one must also measure services charging explicit fees on an equal footing, if one is to obtain consistent aggregates. However, in the case of services charging explicit fees and not attached to any financial assets, it makes little sense to use some interest rate differential (which one?) as the implicit price deflator, and derive a quantity index on par with the real balance of some imaginary financial assets.

In this era of rapid technological progress and proliferation of new financial instruments (including innovative combinations of financial and service features), the greater conceptual consistency afforded by our output measure seems particularly desirable. That is, no matter how the composition of explicitly versus implicitly charged services changes both over time and across financial institutions (or even within an institution), our measure of real output should, in theory at least, generate consistent aggregates.

Take the commercial banking industry, for example, where such compositional changes abound. Ever since interest rate ceilings were removed for most types of deposit accounts, banks have been broadening the range of retail transactions carrying an explicit fee schedule, while raising the interest rates paid on deposits. Most banks also offer depositors a choice between paying a per-transaction fee or maintaining a higher balance. Such changes may well have affected the ratio between real deposit balances and the amount of transaction services both across banks and over time. Similarly, on the lending side, many banks now charge a fee for loan

applications and preapprovals, whether or not a loan is actually granted later. In addition, large banks increasingly engage in (aptly named) off-balance-sheet activities, such as underwriting derivatives contracts, that generate fees but produce no corresponding assets or liabilities on the balance sheet.[35]

Such developments have created difficulties for studying bank production technology because existing measures of output cannot generate a consistent aggregate for any bank engaged in these diverse activities. For example, Rogers (1998) measures the output of traditional bank lending using the balance-sheet value of loans, but measures off-balance-sheet activities using banks' explicit revenue. In contrast, the real-service-flow measure of output advocated here should in theory yield an output index comparable both across banks and over time.

Our results point to the need to reexamine the findings of a large literature analyzing the properties of banks' production technology. That literature features three approaches that differ only in what each defines as bank output.[36] Once a type of activity is taken to be output, all three approaches invariably measure it using the deflated book value of the corresponding financial assets or liabilities. Given the substantial changes in the scope and mode of operation in banking organizations, such book-value-based output measures may have led to biased estimates of banking technology parameters, and in turn to flawed policies.

IV. Conclusion

This paper demonstrates that unrealistically restrictive conditions are needed to obtain a fixed relationship between the quantity of a financial service and the volume of its associated financial instrument. It implies that a quantity index proportional to the real balance of financial assets is

[35] Many off-balance-sheet derivatives are defined based on a so-called notional value, which is reported in a special section of banks' regulatory filings. However, the notional value of a derivative contract (such as a swap or a forward) bears no definite relationship to the actual financial worth of the contract, let alone to the amount of financial services rendered.

[36] Specifically, the three approaches are distinguished by the treatment of deposits—as an output or an input. The asset approach views deposits as an input to the making of loans, which, together with market securities, constitute the output. The value-added approach views every financial product whose creation requires labor and capital as inputs, and it therefore records deposits as an output. The user-cost approach, which is also the foundation for the National Income and Product Accounts' measure, classifies input and output endogenously: given a reference rate, financial assets (liabilities) whose realized rates of return are greater (less) than the reference rate are defined to be output, and other financial assets as input. So transaction deposits are typically found to be outputs in data. See Berger and Humphrey (1997) for a survey.

unlikely to be a robust proxy for the true real output of actual financial services. This conclusion is general, even though features of the transaction technology in the model are, for the sake of intuition, chosen to resemble payment services provided by real-world banks to their depositors and by mutual funds to their shareholders.

The focus of this study is the measurement of real output of financial services and the corresponding price deflators. It is a natural continuation of our earlier work developing a new measure of the nominal output of financial services. In those studies (particularly Wang 2003a and Wang, Basu, and Fernald 2009), we argue that the user-cost approach can serve as the theoretical basis for measuring nominal output of implicitly priced financial services, once it is extended to take account of the (systematic) risk in the associated financial instruments. That is, there is no single reference rate, and each specific rate depends on the risk of the relevant financial securities.

In this study, we emphasize that, independent of how to deal with risk in measuring nominal output, real financial services are unlikely to be demanded in fixed proportion to the real balance of any specific category of financial instruments. In fact, this argument can be made most forcefully in the case where there is no risk, as in the framework of this paper. Without risk, all financial assets must offer an identical rate of return, and so the quantity of any individual class of assets is indeterminate, unless additional constraints are imposed. In contrast, the quantity of each type of transaction services is determined by its production technology.

Therefore, indices that directly measure real services produced by financial institutions are robust to the types of technological and institutional changes that we observe. Thus, even though these indices are typically more difficult to construct because the detailed data needed to construct the right quality-adjusted individual quantity indices are unavailable, they are strongly preferred conceptually. Furthermore, an index based on direct measures of service output yields an aggregate measure of financial services that is conceptually meaningful, whether the services are implicitly or explicitly charged for. This seems a particular advantage in an era of rapid innovation and increasing diversity in financial institutions' modes of operation.

The measurement community has dealt successfully with many challenging tasks—for example, constructing quality adjusted price indices for durables and for medical services. Now that the conceptual foundations for measuring real and nominal financial sector output are falling

into place, we are confident that patient, persistent effort on both the theoretical and empirical fronts will soon bear fruit in this area as well.

References

Barnett, William A. 1978. "The User Cost of Money." *Economic Letters* 1(2):145–149.

Barnett, William A. 1980. "Economic Monetary Aggregates: An Application of Index Number and Aggregation Theory." *Journal of Econometrics* 14(1):11–48.

Basu, Susanto, Robert Inklaar, and J. Christina Wang. 2011. "The Value of Risk: Measuring the Services of U.S. Commercial Banks." *Economic Inquiry* 49(1):226–245.

Baumol, William J. 1952. "The Transactions Demand for Cash: An Inventory Theoretic Approach." *Quarterly Journal of Economics* 66:545–556.

Berger, Allen N., and David B. Humphrey. 1997. "Efficiency of Financial Institutions: International Survey and Directions for Future Research." *European Journal of Operational Research* 98(2):175–212.

Donovan, Donald J. 1978. "Modeling the demand for liquid assets: An application to Canada. *IMF Staff Papers* 25 (December): 676–704.

Feenstra, Robert C. 1986. "Functional Equivalence between Liquidity Costs and the Utility of Money." *Journal of Monetary Economics* 17(2):271–291.

Fischer, Stanley. 1983. "A Framework for Monetary and Banking Analysis." *Economic Journal* Supplement 0(0):1–16.

Fixler, Dennis J. and Kimberly D. Zieschang. 1992. "User Costs, Shadow Prices, and the Real Output of Banks." In *Studies in Income and Wealth*, vol. 56: 219–243, ed. Z. Griliches. Cambridge, MA: NBER.

Fixler, Dennis J. and Marshall B. Reinsdorf. 2006. "Computing Real Bank Services." NBER summer institute. Available at http://users.nber.org/~confer/2006/si2006/prcr/reinsdorf.pdf.

Hancock, Diana. 1985. "The Financial Firm: Production with Monetary and Nonmonetary Goods." *Journal of Political Economy* 93(5):859–880.

Inklaar, Robert, and J. Christina Wang. 2012. "Measuring Real Bank Output: Considerations and Comparisons." *Monthly Labor Review* (July):18–27.

Jovanovic, Boyan. 1982. "Inflation and Welfare in the Steady State." *Journal of Political Economy* 90(3):561–577.

Rogers, Kevin E. 1998. "Nontraditional Activities and the Efficiency of U.S. Commercial Banks." *Journal of Banking and Finance* 22(4):467–482.

Romer, David. 1986. “A Simple General Equilibrium Version of the Baumol-Tobin Model.” *Quarterly Journal of Economics* 101(4):663–685.

Rotemberg, Julio J. 1984. “A Monetary Equilibrium Model with Transactions Costs.” *Journal of Political Economy* 92(1):40–58.

Royster, Sara, E. 2012. “Improved Measures of Commercial Banking Output and Productivity.” *Monthly Labor Review* (July):3–17.

Sidrauski, Miguel. 1967. “Rational Choice and Patterns of Growth in a Monetary Economy.” *American Economic Review* 57(2):534–544.

Tobin, James. 1956. “The Interest Elasticity of Transactions Demand for Cash.” *Review of Economics and Statistics* 38:241–247.

Triplett, Jack E., and Barry P. Bosworth. 2004. *Productivity in the U.S. Services Sector: New Sources of Economic Growth.* Washington DC: Brookings Institution Press.

Wang, J. Christina. 2003a. “Loanable Funds, Risk, and Bank Service Output.” *Federal Reserve Bank of Boston, Working Paper Series*, 03–4.
Available at http://www.bos.frb.org/economic/wp/wp2003/wp034.htm.

Wang, J. Christina. 2003b. “Service Output of Bank Holding Companies in the 1990s and the Role of Risk.” *Federal Reserve Bank of Boston, Working Paper Series*, 03–6.
Available at http://www.bos.frb.org/economic/wp/wp2003/wp036.htm.

Wang, J. Christina, Susanto Basu, and John Fernald. 2009. “A General-Equilibrium Asset-Pricing Approach to the Measurement of Nominal and Real Bank Output.” In *Price Index Concepts and Measurement*, NBER Studies in Income and Wealth 70, eds. W. E. Diewert, J. S. Greenlees, and C. R. Hulten: 273–328. Chicago: Chicago University Press.

Appendix: Model Solution in Steady State

This appendix derives the solutions of the endogenous choice variables in the steady state. Since all variables are in steady state, time subscripts are omitted. First, we can express K for each sector as a function of sectoral L based on equation (46):

$$K^x = kL^x, \text{ where } x = G, B \text{ and } Z, \text{ and } k \equiv \left(\frac{\gamma G}{\rho + \delta}\right)^{1/(1-\gamma)}. \qquad \text{(A.1)}$$

Combined with the definition of total productive capital and labor, this means that $K = k$.

Second, equation (46) combined with (48) and (49) implies that

$$K^B = K^Z \text{ and } L^B = L^Z.$$

This is consistent with consumers' optimal demand for the number of transfers relative to banks' payment services so that they spend the same amount on mutual fund and bank services to minimize total expenditures on both types of transaction services. This then implies that

$$L^G = 1 - 2L^B.$$

Plugging these results, along with (48), into the resource constraint (47) yields the following equation of L^B only:

$$B\left(kL^B\right)^{\gamma}\left(L^B\right)^{1-\gamma} + \delta k = G\left[k\left(1-2L^B\right)\right]^{\gamma}\left(1-2L^B\right)^{1-\gamma}. \qquad \text{(A.2)}$$

Hence, we can solve for L^B, which equals L^Z:

$$L^B = \frac{Gk^{\gamma} - \delta k}{\left(B + 2G\right)k^{\gamma}} = L^Z.$$

This immediately implies that

$$L^G = 1 - 2\frac{Gk^{\gamma} - \delta k}{\left(B + 2G\right)k^{\gamma}} = \frac{Bk^{\gamma} + 2\delta k}{\left(B + 2G\right)k^{\gamma}}.$$

Solutions of labor input then yield solutions for capital input $K^x = kL^x$, where $x = G$, B, and Z. With all the inputs of production derived, we use (48) to solve for consumption C and bank payment services S. Both depend on the technologies of both the goods-producing industry and the banking industry because of our setup where consumption can only be realized with the help of equal units of bank payment services:

$$S = C = B\frac{Gk^{\gamma} - \delta k}{B + 2G}. \qquad \text{(A.3)}$$

We can then easily derive the steady-state output of mutual funds' transfer services, since they employ the same amount of factor input:

$$N = Z\frac{Gk^{\gamma} - \delta k}{B + 2G}. \tag{A.4}$$

First-order condition (24) gives the price for mutual funds' transfer services:

$$P^{N} = \frac{\gamma^{\gamma}\left(1-\gamma\right)^{1-\gamma} G}{Z}. \tag{A.5}$$

With *C* and *N* derived, we easily obtain the average balance of deposits, which simply depends on the relative technology between the mutual fund and the bank sector:

$$\bar{D} = \frac{C}{2N} = \frac{B}{2Z}. \tag{A.6}$$

Condition (35) solves for the interest rate on deposits:

$$r^{D} = \rho - \frac{2ZG\gamma^{\gamma}\left(1-\gamma\right)^{1-\gamma}\left(Gk^{\gamma} - \delta k\right)}{B\left(B + 2G\right)}. \tag{A.7}$$

First-order condition (12) yields the wage rate: $W = \left(1-\gamma\right)Gk^{\gamma}$.