Economic Quarterly—Volume 102, Number 4—Fourth Quarter 2016—Pages 261–279

Nonparametric Estimation of the Diamond-Dybvig Banking Model

Bruno Sultanum

The Diamond and Dybvig (1983) model has been extensively used to explain episodes of runs against financial institutions. In the model, depositors face uncertainty about whether they would prefer to consume in an early or late period. Because there are costs associated with an early liquidation of investments, depositors can benefit from an insurance contract with respect to their preference shock. The optimal insurance will transfer resources from those depositors who prefer to consume in the late period, and therefore get a better return in their investments, to those who prefer to consume in the early period. Such transfers, however, cannot be contingent on the depositor preference because these are not observed—the contract must be incentive compatible so they reveal their true preferences in equilibrium. As Diamond and Dybvig (1983) argue, an arrangement that has this property is a bank contract. The bank promises the efficient transfer in the early period to any depositor who claims the resources. In one equilibrium, only those who actually have preference for early consumption claim early payments. However, there is also an equilibrium where depositors fear that every other depositor, including those with preference for later consumption, will claim early payments. As a result, depositors fear no resources will be left at the bank for consumption late and all of them have incentives to claim payments in the early period, generating a self-fulfilling bank run.

I thank Huberto Ennis, Todd Keister, Neil Wallace, Thomas A. Lubik, Zhu Wang and Quang Vuong for helpful comments. The views expressed in this paper are those of the author and should not necessarily be interpreted as those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

Even though this argument is intuitive, whether or not the model generates runs under a well-designed bank contract depends on the particular specifications of the environment. For example, if there is no aggregate uncertainty, in the sense that the number of early consumers in the economy is known, a simple suspension scheme is able to prevent runs from happening. This observation was made by Diamond and Dybvig (1983), but they highlight that this is only true without aggregate uncertainty.[1] Later, Wallace (1988) pointed out the importance of *sequential service* in order to generate bank runs in the Diamond-Dybvig model. Sequential service is a constraint that payments must be done in sequence, as depositors arrive at the bank, so payments to one depositor cannot be made contingent on future withdrawal demand. However, as Green and Lin (2000, 2003) later point out, aggregate uncertainty and sequential service alone are not enough to generate bank runs in the Diamond-Dybvig model. As they pose it, the theory as it stood was incomplete. A huge theoretical literature has followed trying to understand what other ingredients are necessary for the existence of bank runs.

The approach to address this question and understand whether the Diamond-Dybvig model actually generates bank runs or not has been to build examples where bank runs do exist—see, for instance, Peck and Shell (2003), Ennis and Keister (2009b), and Sultanum (2014). The examples of bank runs are built with particular distributions of liquidity needs and other primitives, such as preferences. However, for the model to explain observed runs, we need its primitives to be consistent with empirical observations. In particular, the distribution of liquidity needs in the model should be consistent with the empirical one. Without this consistency between data and model, the model would still be "incomplete" as a theory to explain historical run episodes. Hence, developing tools to estimate the Diamond-Dybvig model is an important step in understanding how bank runs actually work and how to prevent them.

Moreover, the advantages of estimating the Diamond-Dybvig model go beyond the positive aspect of explaining empirical observations. There are also normative advantages. One way policymakers can use this tool is to estimate the model for different markets and institutions (possibly also making it state contingent) and then test for which markets and institutions a run equilibrium exists. That is, the model can guide policymakers to when and where runs are a possibility, allowing timely policy measures to be taken before a run ever happens.

[1] As pointed out later by Ennis and Keister (2009a), this result also relies heavily on the bank's ability to commit to a suspension scheme.

In this paper, I construct a structural estimator for the distribution of liquidity needs in the version of the Diamond-Dybvig model studied in Sultanum (2014). The data requirement for the estimator is that the econometrician observes the total amount withdrawn. This assumption serves two purposes. First, it is a very weak data requirement, which is always welcome since more detailed data may only be available to regulators (sometimes not even to them). Second, in the model, a depositor either withdraws his entire deposit or nothing. In practice, people can withdraw only part of their money or can withdraw money from multiple accounts at the same bank. So the "amount withdrawn" is a clear and well-defined measure both in theory and in the model. It is not clear how to match the observation of partial withdrawals or withdrawals from different accounts in the data to the model.

What makes this problem difficult is that aggregate payments are observed, but the preferences that define the liquidity needs are not observed. Therefore, in order to estimate the distribution of liquidity needs, one must establish a map between payments and preferences. However, because payments embed an insurance against the preference shock risk, how much is paid for any given realization of preference shocks is endogenous. In particular, it depends on the distribution of liquidity needs.

There is a large literature that studies whether past run episodes against financial institutions were due to coordination failure, as in Diamond and Dybvig (1983), or not. This literature focuses on indirect tests of theoretical frameworks. That is, it tests some of the implications of the theory rather than estimates a particular model and tests whether it generates runs or not. Two recent examples are Foley-Fisher et al. (2015) and Schmidt et al. (2016). Foley-Fisher et al. (2015) develop a model to study runs against extendible funding agreement-backed notes (XFABN) issued by life insurers. Their theory suggests an instrument for the strategic complementarity among investors that is (plausibly) exogenous to variations in fundamentals. So, to test whether self-fulling runs played a role in the withdrawals from XFABN or not, the authors test the correlation between the instrument implied by the theory and the observed withdrawals. Schmidt et al. (2016) study runs against money market mutual funds. The authors develop a model and test its different predictions. For example, they test whether outflows from sophisticated investors in reaction to worse fundamentals are greater than from unsophisticated ones, where sophistication is defined by the quality of the information the investor has access to.

I see the nonstructural approach that has been used as complementary to a fully structural one. It sheds light on whether self-fulfilling

bank runs exist or not. In fact, in my estimation procedure I assume that the econometrician knows whether past runs were due to coordination failure or not. However, without the estimation of a structural model, it is hard to test particular theories—such as Diamond-Dybvig. Once the theory is tested, then we can use it to make predictions and/or policy recommendations.

The econometric method I use in this paper builds on those developed for estimation of auctions. Specifically, it builds on Guerre et al.'s (2000) idea of using the equilibrium conditions of the model to map observable to unobservable variables. Since its publication, Guerre et al. (2000) has spurred a huge empirical and theoretical literature. Some of the more recent theoretical examples are Campo et al. (2011), which allows bidders to be risk-averse; Krasnokutskaya (2011), which considers bidders' unobserved heterogeneity; and Kastl (2011), which proposes an estimation method for auctions with discrete bids. On the empirical side we have, for example, Cassola et al. (2013), which uses the extension in Kastl (2011) to study liquidity demand from European banks during the 2007 financial crisis; and Hortaçsu and Kastl (2012), which quantifies the dealers' advantage from observing customers' orders using data on Canadian Treasury auctions.

Even though a lot can be done using, and improving on, the estimation procedure I discuss in this paper, as the literature on the estimation of auctions has shown, the goal of this paper is not to fully investigate all the properties and possible extensions of a particular estimator. The goal here is to provide an illustrative framework that future researchers can build on in order to estimate bank-run models in different settings. I believe that a full investigation is only worth it with a particular dataset and institutional framework in mind. For this reason, a lot of the discussion here is abstract and details on data and applications are left for future research.

The paper is organized as follows. Section 1 describes the model, the equilibrium concept, and provides a characterization of the solution. Section 2 describes the data requirement, discusses identification, and provides an nonparametric estimator and a numerical example of the procedure. Section 3 discusses how the model can be used to test for the existence of bank-run equilibria. Section 4 discusses practical difficulties and challenges associated with estimating the model. Section 5 concludes.

1. THE MODEL

The model builds on Sultanum (2014), which is an extentision of Peck and Shell (2003) with a continuum of agents. The advantage of this

setting is that the optimal bank contract can be easily characterized by a second-order differential equation, which will be used in the proposed estimation procedure.

Environment

There are three periods, zero, one, and two, and a unit measure of agents called depositors. In period zero, each depositor is endowed with one unit of wealth, which they can invest to consume in periods one and two (agents do not consume in period zero). The investment technology is as follows. Investments in period zero pay gross return 1, if liquidated in period one, and gross return $R > 1$, if liquidated in period two. Depositors are identical in period zero. In period one, each depositor receives a preference shock. The preference shock turns them into one of two types: *patient* or *impatient*. The utility of a type impatient depositor is $u(c_1)$, while that of a type patient depositor is $u(c_1 + c_2)$, where (c_1, c_2) is consumption in periods one and two, respectively. The utility function u is the constant relative risk-aversion (CRRA) utility function, $u(c) = \frac{c^{1-\gamma}-1}{1-\gamma}$, and I assume that the risk-aversion parameter γ is greater than 2.

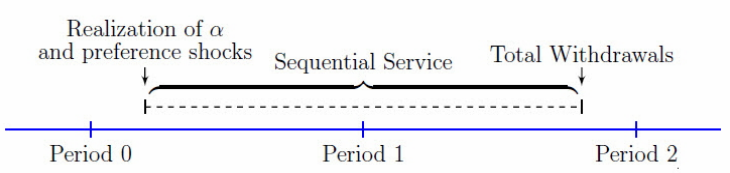
Let $\alpha \in [0, 1]$ denote the fraction of depositors of type impatient. The value of α is assumed to be a random variable with cumulative distribution function F and density function f, which satisfies $f(\alpha) > 0$ for all α in the support $[0, 1]$. The density f is also assumed to be continuous and differentiable in the support. Conditional on the realization of α, the event that a depositor is of type impatient is i.i.d. across agents and has Bernoulli distribution with parameter α. Throughout the text, I refer to α as the aggregate liquidity need in the economy.

One can think that the event where more than, say, 90 percent of depositors are type impatient has probability zero. This could be formalized by allowing the support for the aggregate liquidity need to differ from $[0, 1]$. That is, in general, F could have support $[\alpha_l, \alpha_h]$ with $0 \leq \alpha_l < \alpha_h \leq 1$. In this case, α_l and α_h would also have to be estimated. The approach I describe in this paper can be extended to address this case, which I believe to be of interest. However, as previously stated, going through all the details and extensions of the estimation procedure is beyond the goal of this paper.

Sequence of actions and bank contracts

Depositors face risk in the form of preference shocks (be patient or impatient). As a result, an insurance arrangement is desirable in order to improve depositors' *ex-ante* welfare. Following Peck and Shell

Figure 1 Sequence of Actions



(2003) and, more closely, Sultanum (2014), I focus on a form bank contract where resources are deposited in a bank and depositors can withdraw resources if they want to. That is, in period zero, all resources are deposited in the bank. In the beginning of period one, each depositor observes his own type (which is private information). No one observes the realization of α. Then, agents simultaneously decide whether to withdraw resources from the bank or not. The bank serves the withdrawal requests of the individuals in a random sequence, which the literature refers to as the sequential service constraint (see Wallace [1988] for details on sequential service). A depositor's position in the queue is uniformly distributed among depositors who decide to withdraw resources from the bank in period one.[2] After all withdrawal payments are made, what is left in the bank pays a gross return of R from period one to period two. In period two, the bank distributes the amount left to those who did not withdraw in period one. Figure 1, extracted from Sultanum (2014), depicts the sequence of actions.

A bank contract tells how much a depositor who withdraws in period one receives as a function of his queue position and how much a depositor who waits until period two to withdraw receives as a function of the number of withdrawals in period one. We can formalize it as a pair of continuous functions, $m = (c_1, c_2)$, where $c_1 : [0,1] \to \mathbb{R}_+$, and $c_2 : [0,1] \to \mathbb{R}_+$. The function $c_1(z)$ gives the payment to a depositor who withdraws in period one and has position z in the queue. The function $c_2(\bar{z})$ gives the payment to a depositor who waits until period two to withdraw when the fraction of people who withdrew in period one is $\bar{z} \in [0,1]$. The continuity on m is without loss of generality with

[2] Although depositors arrive at the bank in sequence, the rate of arrivals of depositors at the bank cannot be measured by the bank and, therefore, cannot be used as a factor to determine payments. This assumption can be rationalized by assuming that the time interval in which agents arrive varies proportionally to the number of agents visiting the bank.

respect to finding the constrained optimal outcome, which we define later.

Feasibility of a contract requires that payments must not be greater than the resources available. I impose that the total amount paid must exactly equal the resources available. This requirement is without loss of generality because utility functions are strictly increasing. The feasibility conditions can be written in terms of the functions c_1 and c_2 as

$$c_2(\bar{z}) = \frac{1 - \int_0^{\bar{z}} c_1(z)dz}{1 - \bar{z}} R \quad \text{for all} \quad \bar{z} \in (0,1), \quad \text{and} \quad \int_0^1 c_1(z)dz = 1. \tag{2.1}$$

A bank contract m and the sequence of actions induce a Bayesian game where each player has only two types, either *patient* or *impatient*, and two actions, either withdraw in period one or period two. A strategy profile is a function s that maps types $\theta \in \{patient, impatient\}$ into probability measures over the periods of withdrawal, {period one, period two}. I consider only symmetric Bayesian Nash equilibria of this game, where symmetric means that players of the same type use the same strategy.

It is important to note that the game is simultaneous. That is, when a depositor is deciding whether to withdraw or not, he does not observe the withdrawal decisions of other depositors. One could think that, in practice, people have at least some idea (or signal) of other depositors' actions. For instance, one could see whether or not there is a line in front of the bank, as beautifully illustrated in the Frank Capra movie *It's a Wonderful Life*. Of course, whether this signal is available or not depends on the setting. These days, when many withdrawal decisions are done online or by phone, such as in mutual funds and other shadow banks, it seems reasonable to assume that depositors do not have much information on other depositors' actions prior to their withdrawal decision. For simplicity, I do not allow depositors to observe any other depositors' actions or obtain any signal that is informative of such actions.

The optimal bank contract

The bank problem is to design a contract $m = (c_1, c_2)$ that maximizes ex-ante welfare of depositors. This assumption can be justified by an extension of the model where a competitive bank sector has banks competing to attract depositors from other banks. To keep the exposition simple, however, I follow the literature and directly assume that the goal of the bank is to maximize depositors' welfare.

The outcome that maximizes ex-ante welfare of depositors must be such that only impatient depositors consume in period one, while all the patient depositors consume in period two. This is the case because the return from period one to period two, R, is strictly greater than 1. Therefore, I am interested in bank contracts that have an equilibrium in which only impatient depositors withdraw in period one. I call such equilibrium a no-run equilibrium, and, when a bank contract has a no-run equilibrium, I call it an incentive-compatible bank contract.

When a depositor observes his type, he uses Bayes' rule to update his belief over the distribution of α. Let $f_p(\alpha) = (1-\alpha)f(\alpha)/\int_0^1(1-z)f(z)dz$ be the density of α conditional on the depositor being of type patient. Note that impatient depositors withdraw in period one because they derive no utility from period-two consumption. Therefore, in order to verify that a contract is incentive-compatible, we just need to verify that a patient depositor is better off withdrawing in period two when the other patient depositors are withdrawing in period two.

A feasible bank contract $m = (c_1, c_2)$ is incentive compatible if, and only if, it satisfies

$$\int_0^1 \int_0^\alpha \frac{u(c_1(z))}{\alpha} dz f_p(\alpha) d\alpha \leq \int_0^1 u\left(c_2(\alpha)\right) f_p(\alpha) d\alpha. \tag{2.2}$$

The left-hand side of the above inequality is the expected utility of a patient depositor if he withdraws in period one, and the right-hand side of the inequality is his expected utility if he withdraws in period two—all conditional on the other depositors withdrawing in period one only if they are impatient types.

When depositors are playing the no-run equilibrium, the *ex-ante* welfare associated with a bank contract $m = (c_1, c_2)$ is

$$W(m) = \int_0^1 \left[\int_0^\alpha u(c_1(z))dz + (1-\alpha)u\left(c_2(\alpha)\right) \right] f(\alpha) d\alpha. \tag{2.3}$$

A bank contract is said to be optimal if it achieves the maximum of $W(m)$ among all feasible and incentive compatible bank contracts $m = (c_1, c_2)$.

Let us assume for a moment that the incentive-compatibility constraint does not bind in this problem. Then, using the same approach as in Sultanum (2014), we can show that an optimal bank contract $m = (c_1, c_2)$ always exists and $w(\alpha) = \int_0^\alpha c_1(z)dz$ is the unique solution to the second-order differential equation

$$w''(\alpha)u''(w'(\alpha)) = h(\alpha)\left[u'(w'(\alpha)) - Ru'\left(\frac{1-w(\alpha)}{1-\alpha} R \right) \right] \tag{2.4}$$

with boundary conditions $w(0) = 0$ and $w(1) = 1$, where $h(\alpha) = \frac{f(\alpha)}{1-F(\alpha)}$.[3] Therefore, to solve the model it suffices to first solve the differential equation (2.4), then recover c_1 and c_2 using that $c_1(\alpha) = w'(\alpha)$ and $c_2(\alpha) = \frac{1-w(\alpha)}{1-\alpha}R$.

Equation (2.4) differs from the one in Sultanum (2014) because the Lagrange multiplier of the incentive-compatibility constraint shows up in their characterization, but it does not show up here. In the present setting, because the utility of the patient types is the same as the impatient, the incentive-compatibility constraint of agents does not bind. There are two steps to show this result. The first one is to note that, in any solution of equation (2.4), we must have $c_1(\alpha) = w'(\alpha) \leq c_2(\alpha) = \frac{1-w(\alpha)}{1-\alpha}R$ for all α. Otherwise, the boundary condition would not be satisfied. The second step is to show that this inequality in consumption implies that the period-two distribution of consumption stochastically dominates the period-one distribution of consumption when other patient types withdraw only in period two. That is, patient depositors are better off choosing period-two consumption when they believe that other patient depositors are also waiting to consume in period two. Therefore, we can conclude that the bank contract is incentive compatible.

2. ESTIMATION

The primitives of this economy are given by the risk-aversion parameter, γ, the return, R, and the distribution of the liquidity needs, F. In this section, we establish an estimator for the distribution of the liquidity needs under the assumption that we know γ and R or that they can be identified separately.

The assumption that the return, R, is known seems natural since one can observe market returns from bank balance sheets. The

[3] It is easier to solve the differential equation (2.4) in terms of a system of differential equations, where the marginal utility of period-one consumption, $m_1(\alpha) = u'(w'(\alpha))$, and period-two consumption, $c_2(\alpha) = \frac{1-w(\alpha)}{1-\alpha}R$, are the main variables. That is,

$$m_1'(\alpha) = h(\alpha)[m_1'(\alpha) - Rc_2(\alpha)^{-\gamma}]$$
$$c_2'(\alpha) = \frac{1}{1-\alpha}[c_2(\alpha) - Rm_1(\alpha)^{-1/\gamma}]$$

with boundary conditions $c_2(0) = R$ and $c_2(1) = R/m_1(1)^{1/\gamma}$. By picking the initial $c_1(0)$, one can target the final condition $c_2(1) = R/m_1(1)^{1/\gamma}$. That is, combining the fact that the solution is continuous in the initial condition and that the solutions cannot cross, one can use the intermediate value theorem to argue that an initial $c_1(0)$ such that the boundary condition is satisfied must exist. Moreover, one can show that in such a solution $c_1(1) = c_2(1) = 0$.

assumption that risk-aversion is known, however, deserves justification. This assumption is made for tractability since identifying risk-aversion and distributions together is challenging in this, and also other, settings. For example, Campo et al. (2011) establish that the risk-aversion parameter cannot be identified in first-price auctions together with the distribution of valuations. One could impose additional parametric assumptions in order to identify both the risk-aversion parameter and the distribution of liquidity needs. I consider such analysis interesting but leave it for future research.

There are also two structural assumptions that are necessary for our estimation procedure. Namely, that the bank contract is optimal, as described in the previous section, and that depositors play the no-run equilibrium, where only impatient types withdraw in period one. Alternatively, we could have assumed that we can separately identify the periods in which the no-run equilibrium is played. This is equivalent to saying that, at least ex post, we know whether a bank run happened or not.

In terms of observed data, we assume we have N independent instances of our economy, and in each one we observe only how many total early payments were made as a fraction of the total resources. That is, we can observe a sequence of realizations $\{w_n\}_n$ that are independent of each other. The sample can be interpreted either as a sample over time of the same bank or a sample with N identical banks. In either case, it is important that $w_n = w(\alpha_n)$, where $\{\alpha_n\}_n$ are independent and identically distributed according to F, and $w(\alpha)$ solves (2.4). In the next subsection I show that these data contain enough information to identify F.

I would like to emphasize that this is a very weak data requirement. Only total outflows from the financial institution being studied are necessary. One could try to improve upon the estimation procedure I discuss here by having additional data available, for example, by having microdata on individual depositors. Additional data would also allow for extensions of the model where more primitives of the economy could be identified. But, as I show, just data on outflows already provide a lot of information, allowing us to identify the distribution of liquidity needs.

Identification

A crucial problem in structural estimation is whether the observed data are enough to identify the primitives of the model. In the context of our model, the assumption is that we observe total withdrawals. Let the distribution of total withdrawals be denoted by G. So the question

is whether we can identify the distribution of liquidity needs, F, from the distribution of total withdrawals, G.

In order to answer this question, we use the solution condition of the model to relate G and F. If the map between these two distributions is unique, then the model is identified. So let us look at these conditions. First, the differential equation in (2.4) implies that w is strictly increasing and, therefore, the inverse of w exists. Moreover, because w takes value in the $[0,1]$ interval, G has also support $[0,1]$ and we have that $G(\tilde{w}) = \mathbb{P}[w(\alpha) \leq \tilde{w}] = F\big(w^{-1}(\tilde{w})\big)$. Because w and F are differentiable, we also know that G is differentiable (since it is the composition of differentiable functions) and it satisfies $g(w(\alpha))w'(\alpha) = f(\alpha)$. We can now use these conditions to rewrite the differential equation (2.4) in terms of G. We get that

$$w''(\alpha)u''(w'(\alpha)) = h_G(w(\alpha))w'(\alpha)\left[u'(w'(\alpha)) - Ru'\left(\frac{1-w(\alpha)}{1-\alpha}R\right)\right] \tag{3.1}$$

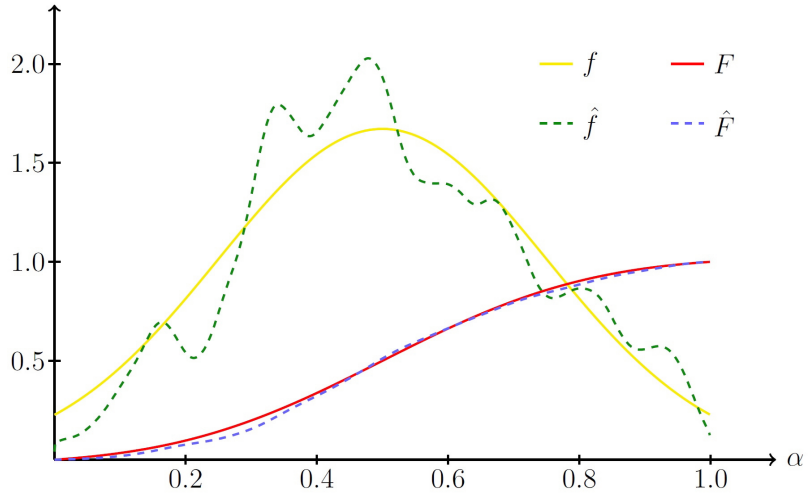
with boundary conditions $w(0) = 0$ and $w(1) = 1$, where $h_G(w) = \frac{g(w)}{1-G(w)}$.

Since we know u and R, for each G we can solve the differential equation (3.1) for w. Once we have w, we can recover F using $F(\alpha) = G(w(\alpha))$. Note that this procedure identifies F. To see this, assume that two distributions, F_1 and F_2, generate the same G. That is, $F_1(\alpha) = G(w_1(\alpha))$ and $F_2(\alpha) = G(w_2(\alpha))$, where w_1 and w_2 are solutions to the differential equation (2.4) associated with F_1 and F_2, respectively. If that is the case, then w_1 and w_2 would both have to solve (3.1). But one can show that equation (3.1) admits only one solution, which implies that $w_1 = w_2$ and $F_1 = F_2$. Therefore, we can conclude that the distribution of total withdrawals, G, combined with the first-order condition that characterizes the optimal bank contract, contains enough information to identify the distribution of aggregate liquidity needs F.

Estimation steps and numerical example

I propose an indirect nonparametric estimation of F. This estimation has three steps. First, we estimate the distribution of total withdrawals G. Call this estimator $\hat{G}$. Then we solve the differential equation (3.1) where G is replaced with $\hat{G}$. Call the solution to this differential equation $\hat{w}$. Finally, the estimator of the cumulative distribution of liquidity needs is $\hat{F}(\alpha) = \hat{G}(\hat{w}(\alpha))$, and its density estimator is $\hat{f}(\alpha) = \hat{g}(\hat{w}(\alpha))\hat{w}'(\alpha)$.

Figure 2 Distributions



The problem of estimating G is a standard nonparametric estimation problem for a continuous distribution over a compact support. One must chose a bandwidth $h > 0$ and a Kernel function $k : \mathbb{R} \to \mathbb{R}_+$, where $\int k(u)du = 1$. Then we have that

$$\hat{g}(w) = \frac{1}{Nh} \sum_n k\left(\frac{w - w_n}{h}\right) \quad \text{and} \quad \hat{G}(w) = \int_0^w \hat{g}(\tilde{w})d\tilde{w}. \tag{3.2}$$

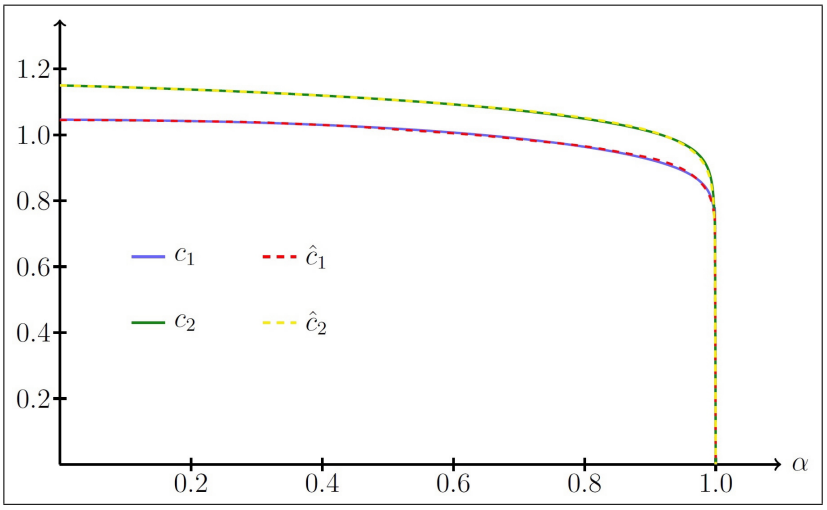
There are different ways of choosing the bandwidth h and the Kernel k. I refer the interested reader to Pagan and Ullah (1999) for a full discussion.

To illustrate how this estimator works in practice, below I simulate the model and use our procedure to estimate the underlying distribution F. I consider a specification of the model with $\gamma = 3.0$, $R = 1.15$, and F is a normal distribution with mean $\mu = 0.5$, standard deviation $\sigma = 0.25$, and truncated between 0 and 1.

The numerical exercise is performed with the following steps. I first solve the model for w using equation (2.4). Then I draw 500 observations α_n from F and generate the sample $\{w_n\}_n$ using $w_n = w(\alpha_n)$ for $n = 1, \ldots, 500$. For the nonparametric estimator of G, I choose the bandwidth $h = 1.06\hat{\sigma}_w N^{-1/5}$, where $\hat{\sigma}_w$ is the standard deviation of the sample $\{w_n\}_n$, and the Kernel function

$$k(u) = 1(|u| \leq 1)\frac{35}{32}(1 - u^2)^3.$$

Figure 3 Bank Contract



After estimating $\hat{G}$, I obtain $\hat{w}$ using equation (3.1). Then I compute $\hat{F}$ and $\hat{f}$.

This procedure allows us to estimate two interesting objects. The first one is the distribution of liquidity needs, which is characterized by its cumulative distribution $\hat{F}$ and the associated density $\hat{f}$. Another interesting outcome of this procedure is the estimation of the bank contract itself. Hence, we can see how the bank contract $m = (c_1, c_2)$ compares to the estimated one $\hat{m} = (\hat{c}_1, \hat{c}_2)$.

Figures 2 and 3 depict the estimation outcomes. Figure 2 depicts the distribution of aggregate liquidity needs. We can see that the estimates, $\hat{F}$ and $\hat{f}$, provide a good approximation of the true cumulative distribution F and its density f. Figure 3 shows the graph with the true bank contract $m = (c_1, c_2)$ and the estimated one $\hat{m} = (\hat{c}_1, \hat{c}_2)$ for comparison. The contracts are extremely close.

3. TESTING FOR THE EXISTENCE OF BANK RUNS

A bank-run equilibrium is defined as an equilibrium where depositors withdraw not because they have liquidity needs but because they believe all other depositors are withdrawing. When estimating a bank-run model, an important question the econometrician could have in mind is whether a bank-run equilibrium exists or not. In this section we discuss possible econometric tests to address this question.

Consider a more general formulation of our model where the support of F is any interval $[\alpha_l, \alpha_h] \subset [0, 1]$. I focus on this generalization because, in the previous formulation of the model, a bank-run equilibrium always exists if $\alpha_h = 1$, so the question of whether a bank run exists would be uninteresting. In the context of this generalized model, the condition for the existence of a bank-run equilibrium is that

$$\int_0^{\alpha_h} u(c_1(\alpha))d\alpha \geq u(c_2(\alpha_h)). \tag{4.1}$$

The left-hand side of the inequality is the expected utility of a *patient* depositor if he decides to withdraw in period one, while the right-hand side of the inequality is his expected utility if he decides to withdraw in period two—all conditional on every other depositor withdrawing in the first period.

Define the propensity to run as

$$P = \int_0^{\alpha_h} u(c_1(\alpha))d\alpha - u(c_2(\alpha_h)). \tag{4.2}$$

For an econometrician who has the prior that the model is capable of explaining observed bank runs, the null hypothesis is that $P \geq 0$. The advantage of formulating the problem in this way is that this is a hypothesis that can be empirically tested. That is, define the statistic $\hat{P}$ as

$$\hat{P} = \int_0^{\hat{\alpha}_h} u(\hat{c}_1(\alpha))d\alpha - u(\hat{c}_2(\hat{\alpha}_h)). \tag{4.3}$$

Traditional econometric methods can be applied to derive the distribution of $\hat{P}$, build confidence intervals, and, ultimately, test the hypothesis that $P \geq 0$. That is, the properties of $\hat{P}$ can be used to test whether the model generates bank runs or not.

The test I propose here is essentially different from what is tested in most empirical literature on bank runs. While the focus of the existing empirical literature is to test whether or not past episodes of bank runs were due to coordination failure by estimating the model, the test I am proposing can reveal whether bank runs can happen or not in equilibrium.

Testing for the possibility of a bank run in the model is helpful in two ways. First, this allows us to test the theory itself. Second, if the theory is successful in explaining past run episodes, it can be used to also inform policymakers of which markets and institutions are vulnerable to runs prior to a run happening, when measures can still be taken to prevent them.

The model can also extend to be state contingent, so the propensity to run can be a function $P(\theta)$, where θ contains relevant information

such as economic growth and unemployment. This would allow us to predict under which conditions runs are more likely to happen.

4. CHALLENGES AND PRACTICAL PROBLEMS

Many assumptions are necessary to use the estimation procedure discussed above. Below I discuss some of the difficulties an econometrician would face when taking the model to a particular dataset.

The data requirement for the estimator is very weak, an econometrician only has to observe the early withdrawals in the Diamond-Dybvig model. One issue that arises, however, is that it is not clear how to match early withdrawals in the model with the data. Each application requires the econometrician to define what in Diamond-Dybvig is labeled as *early* versus *late* dates. For the XFABN studied by Foley-Fisher et al. (2015), for example, the answer seems natural. These notes feature specific dates when investors have the option to extend their notes. In other settings, however, the answer may be more challenging.

Before taking the model to the data, an econometrician also has to decide what exactly is the unit of observation the Diamond-Dybvig model represents. Does it represent the entire financial sector? Or does it represent particular financial institutions? If financial institutions have access to a complete set of liquidity contracts, then liquidity demand that is idiosyncratic to one of them does not matter because they would insure against using the available liquidity contracts. In this case, only liquidity demand in the banking sector as a whole matters for allocations. However, if financial institutions do not have access to a complete set of liquidity contracts, then each one should be considered in isolation as a unit of observation.

Another difficulty the model suggests is that, similar to estimation of auctions, combined identification of risk aversion and distribution is challenging. In the estimation procedure I propose here, I assume that the risk-aversion parameter of depositors' utility is known (or it could be separately identified). Once the risk-aversion parameter is known, the econometrician can use the second-order differential equation that characterizes the optimal contract to pin down the distribution of liquidity needs in the economy from the distribution of total withdrawals. However, the map between the two is only unique because the risk-aversion parameter is known. That is, just information on the distribution of total withdrawals would not be enough for the econometrician to identify risk aversion and distribution of aggregate liquidity needs.

A problem, similar to the one created by having to identify the risk-aversion parameter, would also rise if the econometrician has to

identify the Lagrange multiplier of the incentive-compatibility constraint. In the version of the Diamond-Dybvig model I study, the incentive-compatibility constraint of the patient depositors does not bind because the utility function of patient and impatient depositors is the same. In the original Diamond-Dybvig model, however, the utility of a patient depositor is ρ times the utility of an impatient depositor. Hence, the preferences I use are a particular case of Diamond and Dybvig (1983) where ρ equals one. Under the general formulation used in Diamond and Dybvig (1983), the incentive compatibility can bind and the solution to a second-order differential equation that characterizes the optimal contract would depend on the Lagrange multiplier associated with this constraint. Diamond and Dybvig (1983) assume that ρR is greater than one. If we assume the same in our model, the incentive compatibility does not bind for the same reason it does not bind when ρ equals one. However, the econometrician would still have to identify the preference parameter ρ.

Another crucial assumption I make is that the econometrician can identify periods when depositors played the run and no-run equilibrium. This can be challenging in practice for many reasons. In particular, there seems to be a lot of disagreement among economists, after an episode of high demand for liquidity, over whether such episode was caused by fundamental liquidity demand or by a self-fulfilling run. If it is the former, the econometrician should keep this observation in the sample; if it is the latter, he should exclude it. However, if the econometrician eliminates observations with high liquidity demand because he mistakenly identifies those as runs, he would create a sample selection problem and bias the estimator. That is because his sample would be $w_n = (\alpha_n)$, but the α_n would not be drawn from F because he is excluding with some probability observations of high α.

An econometrician would have a similar problem if depositors are more likely to run when the realization of the aggregate liquidity demand is high. Imagine a situation where depositors use as a coordination device a "sunspot" variable x that is correlated with the aggregate liquidity need α. In this case, if the econometrician excludes observations where there is a run, his sample would suffer selection issues and the estimator would be biased. The issue is the same as before, the α_n would not be drawn from F because he is excluding with some probability observations α_n that correlate with the realization of the sunspot variable that leads depositors to run.

5. CONCLUSION

Green and Lin (2000) have called for a complete theory of bank runs—a theory that explains why bank runs happen and also why society is unable to design mechanisms to prevent such bad outcomes. However, a scientific theory is only complete once it is consistent with empirical observations. Thus, particular examples that generate bank runs, as the literature has provided, are important steps toward a complete theory of bank runs, but they are not the final step.

In order to move closer to this final goal, in this paper I attempt to illustrate how the theory can be taken to the data by providing an approach to estimate the version of the Diamond-Dybvig model proposed by Peck and Shell (2003) and extended by Sultanum (2014). The estimator builds on the literature that studies the estimation of auctions. In particular, it builds on the indirect nonparametric approach to estimate first-price auctions proposed by Guerre et al. (2000). I believe this exercise can provide us with a laboratory to think about issues relating to the estimation of the model.

The exercise highlights many challenges we have to handle in order to successfully take this model to the data. However, I believe the main message of this paper is very positive. Many of these challenges have been faced by economists in different fields, and we can borrow many of the tools they have developed.

REFERENCES

Campo, Sandra, Emmanuel Guerre, Isabelle Perrigne, and Quang Vuong. 2011. "Semiparametric Estimation of First-Price Auctions with Risk-Averse Bidders." *Review of Economic Studies* 78 (January): 112–47.

Cassola, Nuno, Ali Hortaçsu, and Jakub Kastl. 2013. "The 2007 Subprime Market Crisis Through the Lens of European Central Bank Auctions for Short-Term Funds." *Econometrica* 81 (July): 1309–45.

Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 91 (June): 401–19.

Ennis, Huberto M., and Todd Keister. 2009a. "Bank Runs and Institutions: The Perils of Intervention." *American Economic Review* 99 (September): 1588–1607.

Ennis, Huberto M., and Todd Keister. 2009b. "Run Equilibria in the Green-Lin Model of Financial Intermediation." *Journal of Economic Theory* 144 (September): 1996–2020.

Foley-Fisher, Nathan C., Borghan Narajabad, and Stephane H. Verani. 2015. "Self-fulfilling Runs: Evidence from the U.S. Life Insurance Industry." Board of Governors of the Federal Reserve System Finance and Economics Discussion Series 2015-032 (March).

Green, Edward J., and Ping Lin. 2000. "Diamond and Dybvig's Classic Theory of Financial Intermediation: What's Missing?" Federal Reserve Bank of Minneapolis *Quarterly Review* 24 (Winter): 3–13.

Green, Edward J., and Ping Lin. 2003. "Implementing Efficient Allocations in a Model of Financial Intermediation." *Journal of Economic Theory* 109 (March): 1–23.

Guerre, Emmanuel, Isabelle Perrigne, and Quang Vuong. 2000. "Optimal Nonparametric Estimation of First-Price Auctions." *Econometrica* 68 (May): 525–74.

Hortaçsu, Ali, and Jakub Kastl. 2012. "Valuing Dealers' Informational Advantage: A Study of Canadian Treasury Auctions." *Econometrica* 80 (November): 2511–42.

Kastl, Jakub. 2011. "Discrete Bids and Empirical Inference in Divisible Good Auctions." *Review of Economic Studies* 78 (July): 974–1014.

Krasnokutskaya, Elena. 2011. "Identification and Estimation of Auction Models with Unobserved Heterogeneity." *Review of Economic Studies* 78 (January): 293–327.

Pagan, Adrian, and Aman Ullah. 1999. *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Peck, James, and Karl Shell. 2003. "Equilibrium Bank Runs." *Journal of Political Economy* 111 (February): 103–23.

Schmidt, Lawrence, Allan Timmermann, and Russ Wermers. 2016. "Runs on Money Market Mutual Funds." *American Economic Review* 106 (September): 2625–57.

Sultanum, Bruno. 2014. "Optimal Diamond-Dybvig Mechanism in Large Economies with Aggregate Uncertainty." *Journal of Economic Dynamics and Control* 40 (March): 95–102.

Wallace, Neil. 1988. "Another Attempt to Explain an Illiquid Banking System: The Diamond and Dybvig Model with Sequential Service Taken Seriously." Federal Reserve Bank of Minneapolis *Quarterly Review* 12 (Fall): 3–16.

Economic Quarterly—Volume 102, Number 4—Fourth Quarter 2016—Pages 281–320

The Rise and Fall of the Quantity Theory in Nineteenth Century Britain: Implications for Early Fed Thinking

Robert L. Hetzel

Since Friedman and Schwartz (1963), monetary historians have been critical of the performance of the pre-World War II Federal Reserve System (see also Hetzel 2014; Meltzer 2003). That poor performance raises the question addressed here. In the first half of the nineteenth century, there was a flowering of the quantity theory. The Bank of England was a solid institution in Great Britain with a rich tradition. Why then did not the founders of the Fed learn from British experience in the nineteenth century?

The objective of the Bank of England prior to World War I was to maintain the gold standard. Maintaining the convertibility of the paper pound into gold made London the center of the world market for financing international trade. That central position complemented Britain's position as the center of a world empire. The Bank's success in maintaining the gold standard required solving two related problems. First, it had to become a "central bank" in the sense that management of its discount rate moved predictably the complex of interest rates in money markets. Second, it had to figure out how to be a lender of last resort in bank panics while also maintaining a gold reserve sufficient to

The author acknowledges helpful comments from Marvin Goodfriend and William Roberds. The author is indebted to the referees on his editorial committee: Borys Grochulski, Daniel Schwam, John Weinberg, and Alexander Wolman. The author is a senior economist and research advisor at the Federal Reserve Bank of Richmond. The views in this paper are those of the author and not those of the Federal Reserve Bank of Richmond or of the Federal Reserve System.

maintain convertibility. Solving this latter problem required limiting the moral hazard encouraged by access to the discount window through limiting the risk-taking of commercial banks and discount houses.

Solving these problems was an extraordinary achievement. Moreover, the Bank was successful in the sense of surviving independently of the British treasury [Exchequer]. What is relevant here, however, is that the Bank of England solved these problems pragmatically without the need for recourse to the analytical framework of the quantity theory. At the same time, the gold standard became monetary orthodoxy. The quantity theory would have been essential if the intellectual and policymaking environment had been receptive to consideration of the alternative monetary standard of fiat money. It was not, and the quantity theory withered away. Despite a revival in the 1920s, its ideas were still largely unknown in the Great Depression. Moreover, because of the association with paper money, in policymaking circles, it was considered subversive of the established social order (Hetzel 1985).

This paper starts with a review of quantity-theoretic thought in nineteenth century Britain. It continues with an overview of the development of Bank of England orthodoxy as the linchpin of the international gold standard. With this background, the paper then explains the reasons why quantity-theoretic thought had largely disappeared by the last part of the nineteenth century. The overview makes the point that the Bank's understanding of the world developed as a pragmatic response to the need to solve the problems mentioned above. By the time of the founding of the Fed in 1913, there was no analytical framework in current use that would have allowed the founders of the Fed to understand the ramifications for the control of prices of its creation.

Real bills filled the vacuum left by the absence of the quantity theory. Real bills was the school of thought that banks should only discount bills arising in the course of commercial transactions. (A real bill was an IOU promising to pay a given amount on a specified date typically in London. Discounting it, that is, paying an amount less than the face value by a discount house or by a bank, provided the financing for goods in transit between producers and consumers.) The over-issue of bank notes that could threaten the ability of a bank to maintain gold convertibility was possible if the bank lent for speculative purposes. The problem the Bank of England had to solve of how to manage the moral hazard from committing to meet the liquidity demands of banks during a bank panic made real bills into an obvious operating principle for its discount window. The principle in itself was a useful part of risk management for a bank and for management by the Bank of England of its discount window lending. Taken by itself as a principle for central banking, without the nominal anchor of the gold standard, it turned

into a disaster for the early Fed. This paper concludes with thoughts about why the early Federal Reserve learned very little in the way of useful knowledge from British monetary experience.

1. DEVELOPMENT OF THE EARLY QUANTITY THEORY

This section highlights through citations the major contributions of early nineteenth century economists to the quantity theory.[1] The British philosophers John Locke and David Hume formulated rudimentary versions of the quantity theory, which in brief is the hypothesis that the institutional arrangements of a country for determining money also determine the behavior of prices (Humphrey and Keleher 1982, Ch. 3). Their work also illustrates the importance of the way in which episodes of monetary disturbances occur for testing the usefulness of the quantity theory. Tests of the validity of the quantity theory require episodes that make evident the causal nature of changes in nominal money.

Locke formulated the key analytical distinction of the quantity theory—the distinction between nominal and real—in his criticism of the plan of the British government to make uniform the silver content of its coinage. Due to wear and clipping, old coins had lost silver content and were worth less in exchange than full-weight coins. The government proposed to equalize the exchange value of coins by increasing the nominal value of the full-weight coins and by reducing the silver content of newly minted coins. Locke (1695 [1968], 43 and 9) protested that the recoinage would impose losses on existing debt holders. "People who are to receive Money upon Contracts already made, will be defrauded of 20 per Cent. of their due. ... Men in their bargains contract not for denominations ..., but for the intrinsick value." (See Mazumder and Wood 2012; Eltis 1995). Locke could then theorize about "the value of money," that is, the price level. "[T]he value of money in any one country, is the present quantity of the current money in that country, in proportion to the present trade. ..." (Locke 1823 [1963], 49, cited in Leigh [1974]).

Hume drew on the discovery of the silver mines in the Americas in order to test the hypothesis that the price level is a monetary phenomenon. Hume (1752 [1955]) gave the classic statement of the short-run nonneutrality of money and its long-run neutrality:

[1] An Appendix ("A Brief Overview of the Quantity Theory") provides a framework for understanding the excerpts cited below.

> Though the high price of commodities be a necessary consequence of the increase in gold and silver, yet it follows not immediately upon that increase. ... At first, no alteration is perceived; by degrees the price rises, first of one commodity, then of another; till the whole at last reaches a just proportion with the new quantity of specie. ... [I]t is only in this interval or intermediate situation, between the acquisition of money and rise of prices, that the increasing quantity of gold and silver is favourable to industry. ... From the whole of this reasoning we may conclude, that it is of no manner of consequence, with regard to the domestic happiness of a state, whether money be in a greater or less quantity.

Hume formulated the price-specie-flow mechanism. He did so in a challenge to the mercantilists, who believed that a country should increase its wealth through restrictions on imports that would increase its gold by producing a positive balance in its international trade. Using a counterfactual in which money declined exogenously, Hume (1752 [1987]) wrote:

> Suppose four-fifths of all the money in Great Britain to be annihilated in one night, and the nation reduced to the same condition, with regard to specie, as in the reigns of the Harry's and Edwards, what would be the consequence? Must not the price of all labour and commodities sink in proportion, and every thing be sold as cheap as they were in those ages? What nation could then dispute with us in any foreign market, or pretend to navigate or to sell manufactures at the same price, which to us would afford sufficient profit? In how little time, therefore, must this bring back the money which we had lost, and raise us to the level of all the neighbouring nations? Where, after we have arrived, we immediately lose the advantage of the cheapness of labour and commodities; and the farther flowing in of money is stopped by our fulness and repletion.

However, a problem with empirical verification and acceptance of the quantity theory was lack of data. There were no time-series data on the price level. Also, the law forbade the melting of coin and the export of bullion. Because of its illegality, there were then no data on the export of bullion that could test a monetary theory of the equilibration of the balance of international payments. Critics could argue that the theory was irrelevant to real world practice. As a result, after the restriction in 1797 in which the Bank of England suspended convertibility, there was no general acceptance of a monetary theory that would have led to the conclusion that the depreciation of the paper pound on the exchanges served the role formerly played by external gold outflows in adjusting to an excess emission of money (banknotes).

The Napoleonic Wars and their fallout produced the monetary disturbances that spurred development of the quantity theory (Fetter 1965, 20-21). In 1797, Britain was the main adversary left of Napoleon. In that year, rumors of an invasion force landing on British shores caused bank runs. The gold reserve of the Bank of England had already been stressed. The abandonment of the paper currency (assignats) in France beginning in 1775 and France's subsequent return to the gold standard probably caused an external drain of gold from the Bank (Hawtrey 1950, 276-7). A drain of gold to Ireland had occurred in 1795 and 1796. In May 1797, Parliament passed the Bank Restriction Act, which suspended the legal requirement that the Bank of England make its bank notes convertible into gold (see Laidler 2000; History of Economic Thought).

Starting in 1800, the pound price of gold bullion rose to a value 10 percent in excess of the old mint price under convertibility. The depreciation initiated a debate over the consequences of restriction.[2] In the absence of index numbers measuring inflation, debate centered on the value of the paper pound on the foreign exchanges where it traded for currencies on a commodity standard. (Using currently available numbers, Figures 1 and 2, respectively, show the level of the consumer price index from 1791 through 1819 and from 1820 through 1910.) The resulting debate led to the development of the quantity theory and at least for a brief period the modern concept of a central bank.[3]

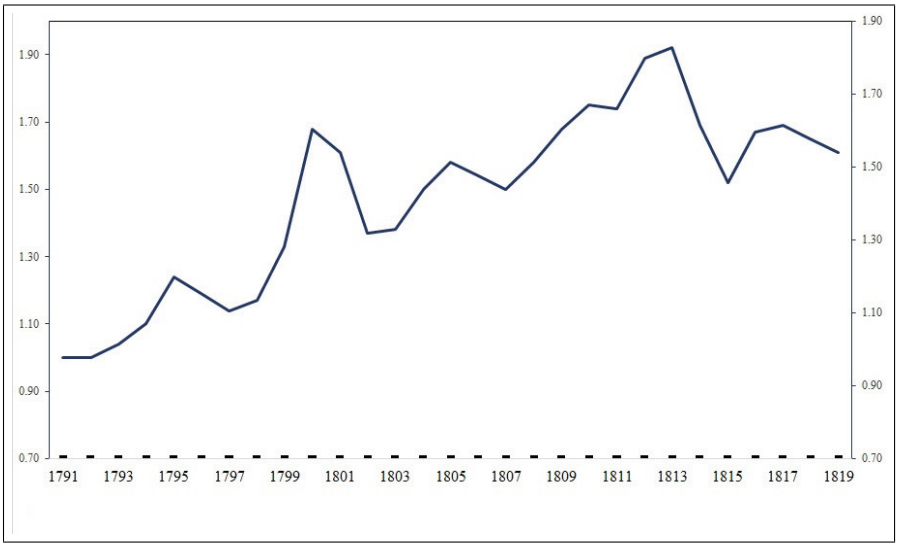
After a revival of inflation toward the end of the first decade of the 1800s, Parliament commissioned the "Bullion Committee" to investigate the causes of the inflation and depreciation of the pound on the foreign exchanges. The committee included most notably Henry Thornton, who more than anyone articulated the idea of a central bank (Hetzel 1987). He did so in his book, *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain* (1802) and in *Two Speeches* (1811). The members of the Bullion Committee (the bullionists) argued

---

[2] Later, two British economists, Joseph Lowe (1823 [1967]) and G. Poulett Scrope (1833 [1968]) advanced the idea of price indices and the indexation of contracts in order to protect against price changes. William Stanley Jevons (1876) revived their ideas. See Humphrey (1974). Joseph Schumpeter (1954, 701) wrote that almost all British economists "up to and including J. S. Mill, distrusted it [the method of index numbers] or even did not grasp its possibilities." Fetter (1965, 139) noted that in 1832 Charles Jones "advocated a policy of price stabilization by a national bank of issue through open market operation, buying public debt when a twenty-commodity price index fell, and selling public debt when the price index rose." One has to wait until Irving Fisher (1935, 97), cited in Humphrey (1990a), for a revival of this proposal.

[3] For an overview of the bullionist/antibullionist debate, see Fetter (1965), Flandreau (2008), Hetzel (1987), Laidler (2000), Mints (1945), Viner (1937), and Wood (2002), as well as various publications by Humphrey (1982a, 1982b, 1983, 1986, 1990b, 1999).

Figure 1 Consumer Price Index in the United Kingdom: 1791 to 1819



Notes: Annual observations. Data from the Bank of England "Three Centuries of Data—version 2.3," via the St. Louis Fed's FRED database.

that the depreciation of the pound on the foreign exchanges and the rise in the price level resulted from the over-issue by the Bank of England of its banknotes. They recommended a return to convertibility. The directors of the Bank of England, who became known as antibullionists, responded with the logic of the real bills argument. That is, over-issue could not occur as long as they issued banknotes only by discounting real bills.

In a piecemeal way, the bullionists developed the analytical insights that would underpin the quantity theory of money. In a rudimentary way, they thought of the real rate of interest as a price with a well-defined (natural) value. Over-issue, that is, an increase in money that caused changes in prices, arose from a Bank of England discount rate held below the natural rate of interest. Bullionists developed the idea of the price level as a monetary phenomenon. Over-issue of the currency caused the price level to rise and the paper pound to depreciate on the foreign exchanges.

Although their opponents, the antibullionists, never developed an analytical framework, they had plausible explanations for the pound's

depreciation and for inflation. The payments made by Britain to support its continental army and allies in the war against France produced the depreciation and the rise in the price of imported commodities. That is, the payments produced an adverse movement in the terms of trade. Wartime demands caused an increase in the price of commodities. That explanation also fit well with the deflation after 1812, the year of the defeat of Napoleon in his Russian campaign. As Fetter (1965, 16) pointed out, the absence of historical precedence for such large remittances rendered any resolution of the monetary versus nonmonetary depreciation of the paper pound on the exchanges problematic.

The directors who conducted the business of the Bank of England (the antibullionists) understood their responsibilities from the perspective of financial intermediation. According to their real bills view, the role of the central bank was to proportion credit to the needs of trade (short-term financing of production) by providing credit for productive purposes as opposed to speculative purposes. In hearings before the Bullion Committee of the British Parliament in 1810, the Bank of England, represented by Gov. Whitmore, defended the criterion of real bills as an adequate safeguard against over-issue of its banknotes.[4] That is, the Bank of England would regulate the issue of banknotes associated with the discounting of bills of exchange by restricting discounting to productive as opposed to speculative purposes.

Asked to state "[W]hat is the criterion which enables the Bank ... to guard the circulation of this country against the possibility of any excess," Whitmore replied, "[T]he criterion by which I judge of the exact proportion to be maintained is by avoiding as much as possible to discount what does not appear to be legitimate mercantile paper.... We never discount without the circumstances being considered; namely ... the appearance of its [the bill of exchange] being used for commercial purposes...." Moreover, the Bank of England insisted that the real bills criterion was sufficient to guard against over-issue of banknotes regardless of the level at which it set the discount rate. Members of the Bullion Committee asked Whitmore, "Is it your opinion that the same security would exist against any excess in the issues of the Bank if the rate of discount were reduced from five to four percent?" Whitmore replied, "I conceive there would be no difference, if our practice remained the same as now, of not forcing a note into circulation."

---

[4] The quotations in the next paragraph come from Wood (2005, 14-18). See also Hetzel (1987) and Viner (1937).

Citing testimony of Whitmore before the Bullion Committee in 1810, W.T.C. King (1936, 73), who was an editor of the *Economist*, wrote:

> The central theory which governed the Bank's credit policy was a conviction that it was impossible to over-issue notes so long as the issues were made by commercial discounts, and for legitimate business only. The working rule was, "let the public act upon the circulation": the public would apply for notes when they were required, and would always repay its loans when less notes were required, to the extent of the excess. The Bank never "forced" its circulation, said the Governor in 1810, with apparent pride. It thus professed to follow a purely passive policy, in the placid faith that it could safely comply with all the demands upon it, provided that it was satisfied that these came from "solid" merchants for bona-fide and not speculative transactions.

For the bullionists, the identification of shocks as monetary in origin started with an assumption that the price system works well in the sense that it produces well-defined values of real variables, in particular, a natural real interest rate. Money creation and destruction exert an independent influence when they emerge in response to a wedge between the market interest rate set by the central bank and the natural rate. Bullionists gave predictive content to this principle through the argument that central banks should operate constrained by a rule that provides for a nominal anchor (gives money a determinate value) and allows market forces to determine the real interest rate.

In this spirit, Henry Thornton (1802 [1939], 259) wrote:

> To limit the total amount of paper issued, and to resort for this purpose, whenever the temptation to borrow is strong, to some effectual principle of restriction; in no case, however, materially to diminish the sum in circulation, but to let it vibrate only within certain limits; to afford a slow and cautious extension of it, as the general trade of the country enlarges itself; to allow of some special, though temporary, encrease in the event of any extraordinary alarm or difficulty ... this seems to be the true policy ... of the Bank of England.

In the *Bullion Report* (Great Britain 1810), Thornton distinguished between credit and money in criticizing the belief that real bills criteria are sufficient in order to limit the quantity of money: "The fallacy upon which it is founded lies in not distinguishing between the advance of

capital to Merchants and an additional supply of currency to the general mass of circulating medium."[5]

As long as the Bank of England's discount rate lay below the rate of interest obtainable in capital markets, the Bank would extend credit and create banknotes (Thornton 1802 [1939], 227 and 253-4):

> Every additional loan obtained from the Bank ... implies an encreased issue of paper. In order to ascertain how far the desire of obtaining loans at the Bank may be expected at any time to be carried, we must enquire into the subject of the quantum of profit likely to be derived from borrowing thereunder the existing circumstances. This is to be judged of by considering two points: the amount, first, of interest to be paid on the sum borrowed; and, secondly, of the mercantile or other gain to be obtained by the employment of the borrowed capital. ... Any supposition that it would be safe to permit the Bank paper to limit itself, because this would be to take the more natural course, is, therefore, altogether erroneous.

A key premise is the nonmonetary character of the interest rate determined in the market for real capital. A Bank rate arbitrarily set differently from the rate on real capital will lead to unlimited changes in the money supply (Thornton 1802 [1939], 255-6):

> [C]apital ... cannot be suddenly and materially encreased by any emission of paper. That the rate of mercantile profits depends on the quantity of this bona fide capital and not on the amount of the nominal value which an encreased emission of paper may give to it, is a circumstance which it will now be easy to point out. ... It seems clear that when the augmented quantity of paper ... shall have produced its full effect in raising the price of goods, the temptation to borrow at five percent. will be exactly the same as before; for the existing paper will then bear only the same proportion to the existing quantity of goods, when sold at the existing prices, which the former paper bore to the former quantity of goods, when sold at the former prices; the power of purchasing will, therefore, be the same; the terms of lending and borrowing must be presumed to be the same; the amount of circulating medium alone will have altered, and it will have simply caused the same goods to pass for a larger quantity of paper. ... [T]here can be no reason to believe that even the most liberal extension of bank loans will have the smallest tendency to produce a permanent diminution of the applications to the Bank for discount.

---

[5] The authors of the *Report* were Henry Thornton, Francis Horner, and William Huskisson. Thornton and Horner very likely wrote this sentence. The excerpt is cited in Wood (2005, 19).

In an explanation of why inflation had increased the incentive for banks to discount at the Bank of England's fixed discount rate, Thornton (1811 [1939], 335) distinguished between the real and nominal rate of interest. Thornton argued that if one borrowed £1000 at 5 percent in 1800 and repaid it in 1810, he

> would have paid an interest of £50 per annum for the use of the money; but, if from this interest were deducted the £25 or £30 per annum which he had gained by the fall in the value of the money, he would find that he had borrowed at 2 or 3 per cent., and not at the 5 per cent. as he had appeared to do.

Because Bank of England banknotes circulated as money along with gold coin, Thornton (1802 [1939], 288; Bordo 1990; Capie and Wood 2007) argued, the Bank was unique among banks in its lender of last resort responsibility. Similarly, it was unique among banks in that its note issue controlled the note circulation of the entire banking system not only among London banks, but also the country banks. The antibullionists argued in opposition that Bank of England banknotes and country bank notes were substitutes in the banking system's production of money balances. A change in Bank of England notes would be counteracted, they argued, by an offsetting change in country bank notes. Thornton (1802 [1939], 225) countered "that the restriction of the paper of the Bank of England is the means both of maintaining its own value, and of maintaining the value, as well as of limiting the quantity, of all the paper in the country."

Thornton's argument made use of Hume's price-specie flow adjustment mechanism in an internal context. The note circulation of the Bank of England determined the price level in the area of London. Given the real terms of trade between London and the country, the price level was then determined for the country. This price level, in turn, determined the quantity of notes that the country banks could circulate. Any attempt by the country banks to issue an amount of notes beyond this given quantity would produce a trade deficit with London, which would produce a reserve outflow to the London banks and counter the initial excess note issue (Thornton 1802 [1939], 208):

> [L]et it be admitted, for a moment, that a country bank has issued a very extraordinary quantity of notes. We must assume these to be employed by the holders of them in making purchases in the place in which alone the country bank paper passes, namely, in the surrounding district. The effect of such purchases, according to the principles established in this Chapter, must be a great local rise in the price of articles. But to suppose a great and merely local rise, is

> to suppose that which can never happen or which, at least, cannot long continue to exist; for every purchaser will discover that he can buy commodities elsewhere at a cheaper rate. ... [H]e will, therefore, require to have his country bank note turned into a Bank of England note.

Thornton also used the price-specie flow mechanism in order to explain how the operation of the international gold standard would cause excess money creation to lead to reserve outflows. With the suspension of note convertibility into gold, money creation instead produced depreciation in the value of the pound on the foreign exchanges. He argued that a concern for this depreciation, in practice, had led the directors of the Bank of England to limit their note issue, despite their professed adherence to the real bills principle (Thornton 1802 [1939], 225 and 249):

> Let the manner in which an extravagant issue of notes operates ... be recollected. It raises ... the cost of British goods. It thus obstructs the export of them, unless a compensation for the high price is afforded to the foreign buyer in the rate of exchange; and the variation in our exchange produces a low valuation of our coin, compared with that of bullion. The variations in the value of bullion, as compared with that of the circulating medium, serve, therefore, to detect and restrain that too great emission of notes to which all countries would otherwise be prone.

Along with Thornton, David Ricardo articulated the quantity theory.[6] Ricardo (1824 [1951], 276) wrote on the distinction between money creation and financial intermediation:

---

[6] It is interesting that Wicksell (1935 [1978], 178), who independently formulated a bank-rate/natural-rate model, initially criticized Ricardo for not explaining how "the banks could succeed in putting a larger amount of their stocks of money or notes into circulation" in a way that made money into the causal factor to which prices had to adjust. However, Ricardo (1821, 364) had written:

> The applications to the Bank for money, then, depend on the comparison between the rate of profits that may be made by the employment of it, and the rate at which they [the Bank of England directors] are willing to lend it. If they charge less than the market rate of interest, there is no amount of money which they might not lend—if they charge more than that rate, none but spendthrifts and prodigals would be found to borrow of them.

When this passage was pointed out to him, Wicksell (1935 [1978], 200) commented that Ricardo's model "is very much on the same lines as the theory I have developed."

> The Bank of England performs two operations of banking, which are quite distinct, and have no necessary connection with each other; it issues a paper currency as a substitute for a metallic one; and it advances money in the way of loan, to merchants and others. That these two operations of banking have no necessary connection, will appear obvious from this,–that they might be carried on by two separate bodies, without the slightest loss of advantage, either to the country, or to the merchants who receive accommodation from such loans.

Although the bullionist/antibullionist controversy dealt with the inflation that followed the suspension of convertibility by Britain in 1797, the bullionists were concerned with both inflation and deflation. Ricardo (1810 [1951], 94, cited in Laidler [2000, 21]) argued for a gradual reduction in paper money in order to lessen the economic disruption of resumption (return to the gold standard at the original parity by making the purchasing power of a paper banknote equal to that of a nominally equivalent gold coin):[7]

> The remedy which I propose for all the evils in our currency is that the Bank should gradually decrease the amount of their notes in circulation until they have rendered the remainder of equal value with the coins which they represent ... or, in other words, until the [pound] prices of gold and silver bullion shall be brought down to their money [parity] price. I am well aware ... that even its sudden limitation would occasion so much ruin and distress that it would be highly inexpedient to have recourse to it as the means of restoring our currency to its just and equitable value. ... If gradually done, little inconvenience would be felt.

In work beginning in the early 1820s, Thomas Joplin expanded upon how a divergence between the natural rate of interest and the market rate of interest (in this case determined by banks allowing their reserves to vary) would make money creation causal with respect to prices. (On Joplin, see Humphrey [1986] and Link [1959].) Joplin considered markets for the quantity of money, goods, and loans. He used the loanable funds framework in which the supply of debt derives from the demand for investment and the demand for debt derives from the supply of saving. Joplin termed the interest rate that equates the supply of saving and the demand for investment the "natural" or "true" rate. Banks can cause the loan rate to diverge from the natural rate. The money supply then changes to the extent that this

[7] Parliament voted for resumption in 1819, and actual resumption of gold convertibility occurred in 1821.

divergence produces a difference in the saving and investment planned by the public.

> When the supply of capital is greater than the demand, it has the effect of compressing it [money supply]; when the demand is greater than the supply, it has the effect of expanding it [Joplin (1832), 101]. Money comes into the market ... from the banks ... in consequence not of a demand for currency, but of a demand for capital, determined by the interest which the banks charge proportioned to the market [natural] rate. And in all cases the influx of money into the market ... is not the effect, but the cause of high prices [Joplin (1823) [1970], 258-9].

This quantity-theoretic way of identifying the causality of money with respect to prices survived in John Stuart Mill's *Principles* (1848 [1909], Book III, Ch. XXIII, 2-3, 15-16, and 22):

> The rate of interest will be such as to equalize the demand for loans with the supply of them. Nevertheless, there must be, as in other cases of value, some rate which (in the language of Adam Smith and Ricardo) may be called the natural rate; some rate about which the market rate oscillates, and to which it always tends to return. ... The rate of interest bears no necessary relation to the quantity or value of money in circulation. The permanent amount of the circulating medium, whether great or small, affects only prices; not the rate of interest. ... But though the greater or less quantity of money makes in itself no difference in the rate of interest, a change from a less quantity to a greater, or from a greater to a less, may and does make a difference in it.
>
> The rate of interest, then, depends essentially and permanently on the comparative amount of real capital offered and demanded in the way of loan; but is subject to temporary disturbances of various sorts from increase and diminution of the circulating medium. ... All these distinctions are veiled over and confounded by the unfortunate misapplication of language which designates the rate of interest by a phrase ("the value of money") which properly expresses the purchasing power of the circulating medium. The public, even mercantile, habitually fancies that ease in the money market, that is, facility of borrowing at low interest, is proportional to the quantity of money in circulation.

At the same time, Mill's unwillingness to apply his framework to alternative monetary standards rendered the quantity theory an irrelevancy. Schumpeter (1954, 715) noted Mill's dismissal of a paper money standard because its "power 'to depreciate the currency without limit' is an 'intolerable evil.'" He commented that through "Mill's refusal to

consider the theory of managed money ... he impoverished monetary analysis."[8]

2. HOW THE BANK OF ENGLAND BECAME THE CENTER OF THE INTERNATIONAL GOLD STANDARD

The Bank of England developed pragmatically the procedures required in order to assure without any doubt gold convertibility and thus to make London the center of the world gold market and the center for the financing of international trade. It was just as important what problem the Bank of England was not trying to solve. It was not trying to stabilize the economy and unemployment.

Horsley Palmer was a director of the Bank of England from 1811 until 1857 and governor from 1830 to 1833. He admitted that the gold standard, which transmitted shocks from around the world to the domestic British financial system, periodically destabilized the economy. In 1848, he testified to the Commons Committee. In reply to Thomas Baring, he said, "[T]he raising of the rate of interest ... stopped very largely the mercantile transactions of the country—exports as well as imports." An exchange with James Spooner, a Birmingham banker, followed on the consequences of raising the discount rate in response to gold outflows (citations from Hawtrey 1938, 28):

> Palmer: It destroys the labour of the country; at the present moment in the neighbourhood of London and in the manufacturing districts you can hardly move in any direction without hearing universal complaints of the want of employment of the labourers of the country.
>
> Spooner: That you ascribe to the measures which it was necessary to adopt in order to preserve the convertibility of the note?
>
> Palmer: I think that the present depressed state of labour is entirely owing to that circumstance.

Thomas Attwood (1832, cited in Fetter [1965, 115]), a banker from Birmingham, argued for a paper currency under the control of the government in order to avoid the periodic contractionary episodes required in order to maintain the gold standard (also see Humphrey 1977). If the Bank of England had wanted to reinvent the monetary standard in order to stabilize the domestic economy and unemployment, it would

[8] Hawtrey (1950, 335) wrote, "Experience of the continental currency and the assignats engendered a deep-seated suspicion of all paper money not directly convertible into metallic currency."

have found these ideas useful. However, it did not need them in order to achieve its objective of maintaining the gold standard.

During the restriction period, there was widespread resentment against "the inequity of the Bank of England monopoly" (Fetter 1965, 111). In May 1819, Lord Livermore expressed these views when he addressed Parliament (cited in Kynaston [1995, 20]):

> No body of men was ever entrusted with so much power as the Bank of England. ... [W]ould Parliament consent to commit to their hands what they certainly would refuse to the sovereign on the throne, controlled by Parliament itself—the power of making money, without any other check or influence to direct them, than their own notions of profit and interest?

As reflected in the disparaging reference to the Bank as "a company of merchants" by Ricardo (1822, 9 and 8), the Bullionists advocated a return to the gold standard as a rule that would curtail the Bank's discretion:[9]

> Whoever ... possessed the power of regulating the quantity of money could always govern its value. ... [T]he currency ... was left entirely under the management and control of a company of merchants—individuals, he [Ricardo] was most ready to admit, of the best character, and actuated by the best intentions; but who, nevertheless ... did not acknowledge the true principles of the currency, and who, in fact, in his [Ricardo's] opinion, did not know anything about it.

It was ironic that the bullionists got the gold standard they desired but that that standard created an environment that caused the quantity theory to become an apparently irrelevant historical artifact. Moreover, for policymakers to understand the applicability of the quantity theory, they must solve the "identification" problem in a particular way. That is, they must understand that the way in which money is created and destroyed (the arrangements of a country for controlling money) is the primal force that drives inflation and significant cyclical fluctuations. However, during the gold standard, nonmonetary explanations explained these phenomena equally well, and they did not require the sophisticated analytical apparatus of the quantity theory. Only the outside "exogenous" event of a change in the monetary standard (clearly evident to all) could offer convincing information about causation. However, even such an event was not "ceteris paribus" in that other factors were always at play. As already noted, the course

[9] On the early history of monetary rules, see Flandreau (2008).

of the Napoleonic Wars could explain the internal and external value of the pound. The British orthodoxy that maintained the gold standard from 1821 until 1914 prevented the kinds of experiments required in order to demonstrate the superior predictive power of the quantity theory, say, in explaining both the behavior of the internal (price level) and the external (exchange rate) value of the currency.

As a way of elucidating how the Bank of England solved the two related problems required in order to maintain the gold standard (the control of market interest rates and the moral hazard arising from lending in a bank panic), the remainder of this section reviews British monetary experience over the time period 1825 through 1907. Much of the focus is on recessions and panics because they were the times when the gold standard was tested. (The definition of recession used is a decline in annual real GDP. See Figure 3).

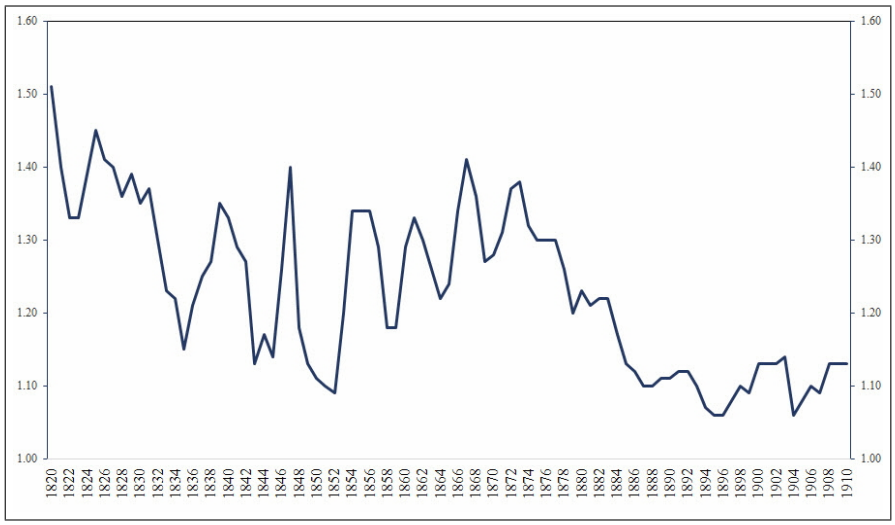
The Bank of England had accumulated large reserves associated with resumption in 1821. In an effort to "employ" those reserves, it lowered the discount rate and lengthened the maturity of bills eligible for discount. The foreign exchanges then turned against the pound, and an outflow of gold, which began in late 1824, continued through the fall of 1825. Bank depositors, fearful of another restriction, initiated an internal drain in the fall of 1825 (King 1936, 35-8). The Bank of England continued discounting but limited its discounting even for high-grade paper. In December 1825, a bank with many country correspondent banks failed and a full-scale panic commenced. Real output declined in 1826 (Figure 3).

The crisis of 1825 marked the first time the Bank of England accepted a lender of last resort responsibility. That is, not until after resumption did the Bank respond to a panic that highlighted the special role of its banknotes as money.[10] In December 1825, the Bank began to discount freely and raised the discount rate from 4 percent to the legally allowed ceiling of 5 percent. In the words of one of its directors, "We lent ... by every possible means, and in modes that we never had adopted before. ... And we were not on some occasions over-nice; seeing the dreadful state in which the public were, we rendered every assistance in our power" (Fetter 1965, 1112-4).

Deflation lasted from 1825 through 1835 (Figure 2). The cause of the 1832 recession is unclear. The Bank rate stayed at 4 percent from 1828 until July 1836. As evidenced by an increase in the market rate

[10] In 1793, following the outbreak of war with France, a severe panic and commercial crisis erupted. However, the government dealt with it through the loan of Exchequer (Treasury) bills to meet the demand for currency while the Bank of England played no special role (see introduction by Hayek in Thornton [1802 (1939)] and Fetter [1965, 12-14].)

Figure 2 Consumer Price Index in the United Kingdom: 1820 to 1910



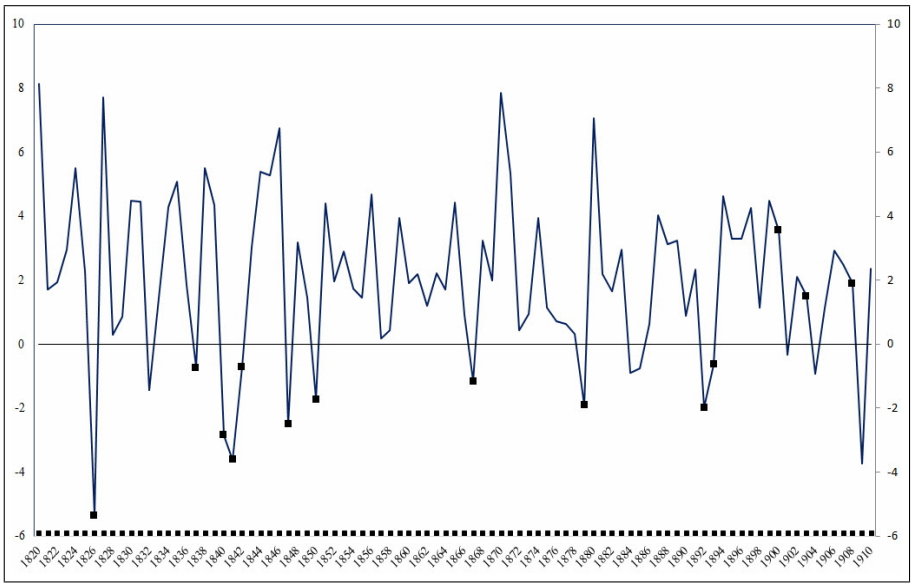
Notes: Annual observations. Data from Bank of England "Three Centuries of Data—version 2.3," via the St. Louis Fed's FRED database.

from 2.8 percent in 1831 to 3.7 percent in 1832, there was financial market stringency (King 1936, 80). An increase in the Bank rate in July 1836 to 4.5 percent and in September 1836 to 5 percent preceded the 1837 recession.[11] According to Temin (1969, 174-5), the Bank of England engineered the 1936 Bank rate increases and refused to discount bills of exchange financing the United States trade in order to offset gold flows to the United States.

The years 1840, 1841, and 1842 all experienced declines in real GDP (Figure 3). Britain again began to lose gold at the end of 1838 because of a financial crisis in France and Belgium and because of payments for imports of corn due to a poor domestic harvest (Hawtrey 1938, 19). In May 1839, the Bank of England raised the Bank rate from 4 percent to 5 percent, in June to 5.5 percent, and to 6 percent in August. In January 1840, the Bank of England lowered the Bank rate back to 5 percent, where it remained until being lowered to 4 percent in April

[11] Figures for the Bank rate are from the St. Louis Fed's FRED database, "Bank of England Policy Rate in the United Kingdom," which in turn are from the Bank of England, "Three Centuries of Data—version 2.3."

Figure 3 Annual Growth Rates of Real GDP in the United Kingdom: 1820 to 1910



Notes: Annual rates of growth of real GDP at market prices. Squares indicate financial crises: 1825, 1839, 1847, 1857, and 1866. Recession years are 1826, 1837, 1840-42, 1847, 1850, 1867, 1879, 1892-93, 1900, 1903, 1908. Data from Bank of England, "Three Centuries of Data—version 2.3" via the St. Louis Fed's FRED database.

1842. However, the Bank limited discounts in 1840 in an "exceptionally stringent" way "as an alternative to an increase in the rate" (King 1936, 82). The price level again declined from 1839 to 1843.

In the 1830s, under governor Horsley Palmer, the Bank moved toward its end-of-the-century practice of keeping the Bank rate above the market rate and serving as a source of discounts primarily at times of financial panic (King 1936, 78). However, that movement was interrupted by the Bank Charter Act of 1844 (Peel's Act). The act divided the Bank into two departments: a banking department and an issue department. The banking department accepted deposits from the public (the London banks, discount houses, and some private individuals). It also discounted bills of exchange. The issue department issued and redeemed banknotes for gold one-to-one. Specifically, above a fixed fiduciary issue of £14,000,000, banknote issues had to be backed 100 percent by gold.

The Bank's directors believed that with the issue department protecting the gold reserve, the banking department could compete as a regular commercial bank, and it did so aggressively. After passage of the act, the directors lowered the Bank rate from 4 percent to 2.5 percent. With the Bank discounting paper at the market rate, it gave up control of the quantity of its discounts. As its discounts increased, its reserves declined (King 1936, 130).

Late in 1846, the Bank of England also began to lose gold due to a deficient wheat harvest and a failure of the potato crop and the resulting heavy imports of food. In response, the Bank raised the discount rate to 3.5 percent in January 1847, but gold outflows continued. The Bank's bullion reserve continued its decline, and in April the Bank raised the Bank rate to 5 percent. It placed a cap on the amount of bills it would discount and drained reserves by engaging in reverse repurchase agreements (RPs) using Consols. Panic then set in and "a virtual paralysis ... of trade resulted." An internal drain of gold further depleted both the banking reserve and the bullion reserve (King 1936, 138-9).[12] Real output declined in 1847 (Figure 3).

On October 1, 1847, a further fall in the reserve caused the Bank also to refuse to make any advances on Exchequer bills. The first bank failed in mid-October. On October 25, the government wrote a letter to the Bank promising indemnity if it exceeded its legal note issue, but the Bank had to charge at least 8 percent on its discounts. Once reassured that the Bank would discount, the panic ended and the reserve recovered (King 1936, 144-7; Hawtrey 1938, 21-23). By allowing the Bank to discount bills and create banknotes above the £14,000,000 limit, temporary suspension of the act gave the Bank of England a war chest for supplying the currency demanded in a panic (Flandreau and Ugolini 2013).

Contemporary observers laid the blame for the 1847 panic on the collapse of speculative excess. Critics of the Bank blamed it for encouraging the speculative lending that had preceded the panic (King 1936, 149). Despite these real bills interpretations, it is clear that contractionary monetary policy preceded the crisis. The Bank did not realize that its deposits also served as money in that commercial banks used them as reserves. Also, it had no understanding of how it could create

[12] The banking reserve was the difference between the legally allowed maximum unbacked note issue of £14,000,000 and the actual note issue. The bullion reserve was the Bank's gold reserve.

destabilizing monetary emissions through driving a wedge between the market rate and a natural rate set in a world financial market.[13]

The next financial crisis, which occurred in 1857, shows how difficult it is to have "clean" experiments that separate monetary and credit disturbances. Hawtrey (1938, 32) characterized the period prior to 1857 as reflecting "the expansive tendency during the active phase of a trade cycle. ..." King (1936, 195-6) described how with the outbreak of the Crimean War in March 1854 raw materials prices and freight rates for shipping rose. Their collapse led to "a wave of failures." According to King, the large number of startups of new cotton mills illustrated the existence of "sustained speculative activity." King's only explanation for why the presumed speculative bubble did not burst sooner was the "continuing Australian gold arrivals." The discovery of gold deposits in California and Australia had produced an inflow of gold into Great Britain and by 1853 a rise in the price level, which was accompanied by labor unrest and strikes (Figure 2).

However, in the fall of 1857, gold flowed out in response to the Indian Mutiny in India and financial panic in the United States.[14] The Bank lent until its reserve (additional legally allowable note issue) was nearly exhausted. When news of the American crisis reached Britain in August 1857, banks with American connections failed. The government again suspended the 1844 act. By the end of November, the crisis began to subside and gold flowed in from the interior and from abroad (King 1936, 193 and 197-9; Hawtrey 1938, 25-26.)

At the same time, the interval of an elevated Bank rate was quite short. It went from 5.5 percent in September 1857 to 10 percent in November 1857 and then declined sharply to 3 percent in February 1858. Growth slowed in 1857 and 1858 but did not turn negative. Given the short-lived incidence of monetary stringency, the absence of a recession despite a slowing of growth suggests that sustained contractionary monetary policy is required for a serious recession. That implication is counter to the conventional real bills view that periods of prosperity lead to "over-trading," the inevitable collapse of which causes a period of purging in the form of recession.

A financial crisis arose again in 1866, and real output declined in 1867. In response to a drain of gold to Ireland that produced a decline

[13] Figure 3 shows a recession in 1850. Over this year, the Bank kept the Bank rate at 2.5 percent and there was no financial panic. Presumably, the recession arose from nonmonetary factors.

[14] Calomiris and Schweikart (1991) attributed the financial panic in the United States to a decline in western land values and the failures of railroads invested in the West. Eastern financial institutions with loans in the West and in railroads suffered runs prompted by fears for their solvency.

in the Bank of England's reserve between June and October 1865, the Bank of England raised its discount rate from 3 percent to 7 percent. In January 1866, the Bank raised the Bank rate to 8 percent "in consequence partly of a bullion drain to the East" (King 1936, 240-1; Hawtrey 1938, 82).

On May 10, 1866, Overend, Gurney & Co. suspended payments. King (1936, 243) cited the contemporaneous response as recorded in the *Bankers' Magazine*:

> [A]s the shock of an earthquake. It is impossible to describe the terror and anxiety which took possession of men's minds for the remainder of that and the whole of the succeeding day. No man felt safe. A run immediately commenced upon all the banks, the magnitude of which can hardly be conceived.

Macleod (1866, 194-5) wrote:

> [O]n the afternoon of Thursday, May 10, the terrible news spread through London that the great establishment of Overend, Gurney & Co. had stopped payment, with liabilities exceeding £10,000,000—the most stupendous failure that had ever taken place in the City.

The Bank of England raised its discount rate to 10 percent, discounted "legitimate" bills, and the government suspended the Bank Charter Act. Panic then began to subside and "mercantile failures were surprisingly few" (King 1936, 240 and 244).

Overend, Gurney occupied an unchallenged position at the top of Britain's credit structure. In the mid-nineteenth century, it expanded beyond bill broking and became a repository for the deposits of country and London banks (King 1936, 120). King (1936, 117) quoted the *Times* that Overend, Gurney could "rightly claim to be the greatest instrument of credit in the Kingdom." However, in 1865, Overend, Gurney became a limited-liability company. That act capped an expansion of its activities beyond bill broking to equity investments in railways and shipping. Failure of those enterprises brought it down. Given the essential position of Overend, Gurney in the credit markets of Britain, it is surprising that real GDP did not decline in 1866 and declined only modestly in 1867.

In 1871, Germany went on the gold standard. As other countries joined, the demand for monetary gold grew relative to the available gold stock. From 1872 until 1896, the British CPI declined by about 23 percent or about 1 percent a year (Figure 2). Over the period 1872 through 1879, annual real GDP growth averaged less than 0.8 percent.

(In the period 1820 through 1871, in contrast, annual real GDP growth averaged 2.4 percent.) As described by Hawtrey (1938, 65):

> [T]he Bank was repeatedly led by a decline in its reserve to raise its rate to 5 per cent. or more, at times when, far from there being any danger of excessive expansion, the vicious circle of contraction was already at work. ... In the periods of depression ... when there were long spells of cheap money with no serious interruption, revival would begin at an early stage.

In response to exports of gold to the United States, the Bank began raising the Bank rate from 2 percent in February 1878 to 6 percent in October 1878. Output declined in 1879. "The depression had been greatly aggravated by the crisis and dear money of 1878. ... It was the stringency and crisis of 1878 that at last brought British industry to a sufficient state of prostration to free the Bank of England from anxiety in regard to the reserve" (Hawtrey 1938, 98-102). Output again declined in 1884 and 1885. Due both to an internal and external drain of gold, the reserve fell from June to November 1884. The Bank rate rose from 2 percent in September 1884 to 5 percent in November 1884 and then gradually declined back to 2 percent in May 1885. "Dear money had again been applied at a time of growing depression" (Hawtrey 1938, 103-4).

The causes of the 1892-93 recession are unclear. The height of the free-silver agitation in the United States occurred in the years 1890 through 1892, and the 1892 election appeared at first to be "an unequivocal victory for the cheaper-money free-silver forces" (Timberlake 1993, 170). As Sayers (1976, 9) noted, "[W]orld gold flows were distorted by the repercussion of American coinage controversies." In 1891, the Bank rate fluctuated widely, varying between 2.5 percent and 5 percent, but those moves did not appear to translate into stringency in the money market.

The 1900 recession was accompanied by contractionary monetary policy. In October 1899, the Boer War broke out. When the reserve declined sharply, the Bank raised Bank rate to 6 percent. The mild 1903 recession appears to be nonmonetary in character. The 1908 recession followed on the gold outflows produced by the 1907 crisis in the United States. "In September, 1906 ... there arose an intense demand for gold for exportation to the United States. ... In four weeks the reserve fell from £24,762,000 to £18,290,000 (12th September to 10th October). ..." Knickerbocker Trust in New York failed on October 22, 1907. "Enormous exports of gold from England to the United States followed. ... Bank rate was put up to 7 ... on the 7th November" (Hawtrey 1938, 116).

After the 1847 financial crisis, the Bank of England began mainly to keep the Bank rate moderately above the market rate by following the market rate. Inflows of gold, for example, caused market rates to fall and the Bank followed them down. With these procedures, the Bank could claim to be "following" the market. However, they created confusion about the Bank's intentions. By 1875, the Bank had developed the practice of maintaining the Bank rate well above the market rate and moving it based on its reserve position (King 1936, 163-66 and 286-7; Hawtrey 1938, 23).

As noted by Hawtrey (1938, 63), as the classical gold standard developed, in setting the Bank rate, "the Bank of England was guided not by evidence of the state of business but by the state of the reserve." King (1936, 167-68 and 317) made the same point:

> [A]t all costs it [the Bank] must preserve an adequate reserve. ... From these considerations there was evolved the practice of regulating Bank rate almost solely according to movements in the Bank's reserve. ... [T]he authorities, once they had realized the dangers of attempting to resist the consequences of foreign-imposed influences, had no practicable alternative but to pursue a so-called automatic policy, regulating Bank rate almost mechanically by gold movements and the trend of the leading exchanges.

Paul Warburg (1910, 16) explained:

> The government bank's discount rate ... is, as a rule, so much higher than that of the general banks, and the restrictions as to the character of the paper which the government bank can take directly are so much more rigid than the requirements of the commercial banks, that in normal times the bulk of the business is done by the general banks and the bankers. Only when the demand for money increases does the rate of the general banks begin to approach that of the government bank, but when this happens the government bank, as a rule, raises its rate, so as to maintain its margin over that of the general banks.[15]

With the Bank rate above the market rate, the Bank of England had to enforce its Bank rate in financial markets. As the directors pointed out in their argument that they were not responsible for the

[15] Warburg (1910, 17) continued to explain that in response to internal drains of currency understood as transitory, such as drains associated with the seasonal movement of crops, the Bank of England left its discount rate unchanged and encouraged discounts and an increase in circulating currency. In response to a persistent drain, however, the Bank of England would "raise the rate in order to protect the reserve and to force liquidation. ..."

speculation in financial markets prior to the 1847 crisis, the Bank had become a small player in the credit markets and could not influence interest rates in a direct way (King 1936, 150). The Bank did control money (bank reserves), however. If market rates failed to increase in response to a drain of gold, the Bank would drain reserves through sales of Consols combined with an agreement to repurchase them, "borrowing on Consols," or reverse RPs in modern terminology. City banks followed the Bank rate knowing that if the market rate fell much below the Bank rate, the Bank of England would raise the market rate through draining reserves (Sayers 1976, 37-38). Also, Hawtrey 1938, 68) wrote:

> When the Bank lost gold or the active circulation of notes increased, there was an equivalent decline in its [commercial bank] deposits. The money available for the commercial banks was diminished, and the gap was made good by the sale of bills to the Bank of England. Thus the loss of gold or the increase in the active circulation itself made Bank rate effective.

Banks acted in anticipation of movements in the Bank rate. King (1936, 320-21) wrote:

> [T]he banks and discount houses began to watch closely the trend of official policy, anxiously scanning the Bank returns, the trend of money rates abroad, and, above all, the exchanges, for any clue as to what the action of Threadneedle Street would be. ... And at such times a helpful hint from the authorities that there might be breakers ahead came to be almost as effective in Lombard Street as even the most direct disciplinary actions had been in the past (King 1936, 320-21).

As the nineteenth century progressed, the Bank of England also developed its lender-of-last-resort responsibility (Capie and Wood 2007 and 2015; Humphrey 1975, 1989, 2010; and Humphrey and Keleher 1984). Bagehot (1873) formalized the concept of lender of last resort based on the bank panic of 1866. In order to prevent moral hazard, the Bank of England enforced a real bills policy on the discount houses. As described in King (1936, 215), by the end of the nineteenth century, the Bank imposed qualitative controls on discounts in a number of ways. It would exercise that control "by exercising a rigorous discrimination against speculative bills"; "rediscounts should be confined where possible to bills of shorter currency than the Bank itself held"; and "the total accommodation to be afforded ... would always be considered strictly in relation to the capital and private resources of each applicant."

3. WHY DID THE QUANTITY THEORY WITHER AS A USEFUL ANALYTICAL TOOL?

As the gold standard became orthodoxy, the quantity theory withered away. The readily "intuitive" demonstration of the quantity theory occurs in a monetary regime in which money creation, especially, to finance government deficits, precedes inflation. In the gold standard organized around the Bank of England's manipulation of the discount rate, in contrast, the norm was for money not to be created but to be supplied on demand through the import of gold. The exceptions to the norm could only be inferred from a sophisticated analytical apparatus appealing to the construct of a natural rate of interest. Consider in this respect the operation of the gold standard.

With the gold standard, the price level for a country comprises a baseline of the average price level in countries on the gold standard. That baseline is determined in the long run by the marginal cost of producing gold, but it is influenced in the shorter run consisting perhaps of decades by the monetary and nonmonetary demand for gold. The price level of a country relative to other countries (relative to the baseline) then adjusts in order to produce the equilibrium real terms of trade that provides for balance of payments equilibrium. In the gold standard, a country's price level is then determined through the working of the marketplace. With this given price level, gold flows are the equilibrating variable in that they adjust through the balance of payments in order to give the public its desired amount of money.

In contrast to the gold standard norm described above, money is "created" by the central bank rather than "supplied" by the market when the central bank creates a divergence between the natural rate of interest and its policy rate. However, the natural rate of interest is not observed. One must infer it from an understanding of the role of the interest rate in the price system *and* from an assumption that the price system works well to determine well-defined values of relative prices. Stated alternatively, the power of the central bank both for good and mischief derives from an ability to create a wedge between the market rate and the natural rate through money creation (destruction). The Achilles' heel of the quantity theory as an intuitive analytical framework for bringing coherence to a monetary regime in which nominal money is demand-determined is the invisibility of the natural rate of interest.

One must infer its existence indirectly. In order to maintain convertibility, the Bank had to enforce the internationally determined interest rate (the natural or market-determined world interest rate) on the British banking system and on the discount houses financing world trade. In a quantity-theoretic spirit, that power derived from its ability

to move the market rate in a way that produced a wedge between the market rate and the natural rate. A market rate above the natural rate led to the destruction of paper money, an increase in market rates, and ultimately a decrease in the price level. Conversely, a market rate below the natural rate led to the creation of paper money, a decrease in market rates, and ultimately an increase in the price level.

The Bank of England's power to enforce the international gold standard had to derive from its unique role as a creator of money. As the nineteenth century evolved, the Bank of England became a negligible player in credit markets. Moreover, it could not run a commodity-price stabilization scheme in order to peg the price of gold because its gold holdings amounted only to between a third and a sixth of the gold in circulation in the United Kingdom (Hawtrey 1938, 41, using figures of William Stanley Jevons). Despite these last two facts, the Bank became the linchpin of the world financial system as the guarantor of the convertibility of the pound sterling. The assurance of the ability and the ease of converting Exchequer bills into gold and vice versa made London into the world's money market. The continental central banks like the Reichsbank, Banque de France, and the Riksbank, which all originally maintained a significant presence in credit markets in the nineteenth century, eventually came to emulate the Bank of England (Sayers 1976, 2).

At the same time, however, a test of the quantity theory that would be readily observable would be a change in the monetary arrangements of a country that caused money to be the forcing variable and prices the equilibrating variable. That is, a test of the quantity theory would have required the unambiguous "experiment" provided by a change in the monetary regime to a paper money standard, especially, one driven by the requirement that the central bank finance government deficits. There were, however, no more departures from the gold standard and returns to it that forced economists and policymakers to adopt a conceptual framework for understanding the resulting behavior of inflation.

There was an additional problem in demonstrating the relevance of the quantity theory beyond the uniqueness of the suspension period. During the Napoleonic Wars, Britain had a usury ceiling of 5 percent. The stimulus to credit provided by wartime demands at times caused the ceiling to bind. There was then a natural counterpart to the Bank rate-natural rate construct used by the bullionists that did not require an understanding of the role of the interest rate in the operation of the price system. That construct disappeared with the end of exceptional wartime demands. It disappeared definitively after 1833 when the Bank Charter Act exempted bills up to three months from the ceiling with an extension to all maturities a few years later.

In the gold standard, the natural experiments relevant to the way in which the behavior of money explains the behavior of prices are the exogenously determined gold flows that originate abroad in the world gold market. For the bullionists, they separate purely monetary shocks from the impairment to credit flows.[16] Thornton (1802 [1939], 271 and 307, cited in Humphrey [2010]) identified monetary excess and instability as the source of inflation and cyclical volatility:

> [I]t is by the amount not of the loans of the Bank of England, but of its paper ... that we are to estimate the influence on the cost of commodities. It is not the limitation of Discounts or Loans, but ... the limitation of Bank Notes or the Means of Circulation that produces the Mischiefs [of lost output and employment].

However, absent the analytical framework of the quantity theory, for contemporary observers in the nineteenth century, the readily "intuitive" explanation of recession and deflation started with the cyclically low interest rates that existed in periods prior to economic expansions and continued with the association of expansions with the optimism about the future that encouraged investment. This "look-out-the-window" story (correlation implies causation) of the business cycle was then that "low" interest rates encouraged speculative excess, and the collapse of that excess caused recessions. For example, in his summary of the debate over the 1847 crisis, King (1936, 149) highlighted the different views on the responsibility assigned to the Bank of England for creating "low" interest rates. However, there was agreement in blaming "the commercial world, for its reckless overtrading, its foolish speculations and its irrational exuberance." King (1936, 149) drew the following conclusion from his review of a century of British experience:

> [T]he commercial world and the general public of all ages since capitalism began have been prone to overreach themselves by an irrational and cumulative optimism which must ultimately bring its own corrective in the shape of a more or less sharp recession, both of confidence and economic activity.

Although the last financial panic in Britain occurred in 1866, recessions continued. All the same, heightened uncertainty about the future and increased risk of default remained characteristic of reces-

[16] Humphrey (2010, 342) expressed the distinction, "Money does what credit cannot do, namely serve as the economy's unit of account and means of [achieving finality of] payment."

sions. It was then difficult to identify natural experiments that would have allowed an unambiguous distinction between monetary disorder and financial disorder as sources of the disruption to trade. What participants observed directly was the cessation of the ready availability of credit. Henry Sidgwick (1883, 265) described how outflows and inflows of gold, a clearly monetary phenomenon, confounded the source of shocks by working through financial markets:

> [I]t should be observed that those who confound the two meanings of "value of money" are not wrong in supposing that the value of the use of money tends to be lowered by an unusual influx of metallic money or bullion, and raised by an efflux: they are only wrong in overlooking the transitoriness of these effects. An increased supply of gold, not accompanied by a corresponding increase in the work that coin has to do ... tends ultimately to lower the purchasing power of money. ... [I]n the first stage of the process that leads to this result, the increment of coin ... must pass through the hands of bankers. ... Hence the price paid for the use of money will tend to fall, and this fall to cause increased borrowing, and consequent extended use of the medium of exchange; and then through the resulting rise in prices generally, the greater part of the new coin ... will gradually pass into ordinary circulation. ... In the same way, when gold has to leave a country ... it will generally be taken chiefly from the reserves of banks; and the need of filling up the gap thus created will make it expedient for bankers to restrict their loans, and so tend to raise the rate of discount. This effect will generally be greater, the smaller the reserve of metal kept by the aggregate of banks, compared with the amount of the medium of exchange that they supply.

Perhaps the most important reason for the fading of the quantity theory as a framework for analysis was that the intellectual environment became hostile because of the theory's association with purposeful money creation by a central bank. Representatives of the agricultural sector and of the industrial interests of Birmingham, such as Thomas Attwood, cited above, criticized the gold standard and the deflation that lasted from 1815 until 1835 (Schumpeter 1954, 405 and 715; Fetter 1965, 99). In response to their arguments for paper money or bimetallism, conservatives rallied around the gold standard and around the Bank of England as the protector of the established order.

After the Bank Act of 1844, investigations into the behavior of money and prices practically disappeared. It was the adherents of the currency school who advocated the act and maintained the quantity theory tradition who should have possessed a natural interest in monetary phenomena. They believed, however, that the provisions of the 1844 act took discretionary control of money away from the Bank of

England. The behavior of the money supply was then believed to depend only upon gold flows in accordance with the discipline of the international gold standard. The automatic operation of the gold standard militated against any need for an analytical understanding of its working.

The commitment to the gold standard in the pre-World War I period came from identification of that standard with the British Empire. Just as London was the center of the empire, London was the center of the world market for gold (Bordo 1999). The absence of any kind of controls on the export and import of gold was the foundation for the central role of London in the finance of world trade.[17] The conservatism of the established status quo with its belief that any alternative to the gold standard was synonymous with social disorder rendered disreputable the development of the quantity theory. King (1936, 317) noted, "The only available alternative [to the gold standard], the substitution of a pure managed currency for an international standard, was unthinkable before the War."

Fetter (1965, 141) commented, "Among economists from the 1820's on the gold standard was a matter of economic theology rather than economic analysis." Fetter (1965, 141) quoted an 1822 speech of Charles Callis Western, member of Parliament and opponent of resumption:

> A degree of something like superstitious veneration has been created for what they [the bullionists] called a SOUND METALLIC currency at the ANCIENT standard of value; a sort of priesthood is exercised by the learned on this subject, by which, as in the case of religious superstition, unassuming patient men are induced to believe that there are mysteries beyond the reach of common sense, and in like manner, give up the use of their own understanding, thus undergoing the fate of all honest dupes. [capitals in original]

The intellectual heirs to the bullionists in the nineteenth century also came to associate the gold standard with laissez faire. Schumpeter (1954, 405) expressed their view: "An 'automatic' gold standard is part and parcel of a laissez-faire and free-trade economy."

---

[17] Eichengreen (1987) discussed the leadership role of the Bank of England in the international monetary system of the nineteenth century. Roberds (2016) highlighted the difficulties other European countries experienced in their attempts to replicate the example of the Bank of England. The creation of a government bank holding interest-bearing government debt financed by currency (noninterest-bearing debt) provided seigniorage revenues to the government. The ability of government to use money creation to finance its spending too often ended in inflation and monetary instability. See also Tullock (1957) for the case of China. As a result, a fiat money standard and the quantity theory came to be associated with runaway inflation and instability.

4. WHAT COULD THE EARLY FED HAVE LEARNED FROM THE BANK OF ENGLAND?

The answer to the question of what could the early Fed have learned from the experience of the Bank of England is "very little." The Bank of England learned pragmatically how to manage the gold standard without understanding or articulating the analytical basis for its operation. What transmitted then in Bank of England practice was not a quantity-theoretic understanding of maintenance of the gold standard but rather the real bills practices followed in its discounting of bills as a way of limiting moral hazard. As a result, when the gold standard broke down with World War I, knowledge of the quantity theory required reinvention.

However, given the atrophied state of the quantity theory and the prejudice against "managed" money, the task proved impossible for the early Fed. Its founders were in no position to make the intellectual leap required for management of a monetary regime with a nominal anchor aimed at the domestic price level (Hetzel 1985). Ultimately, policymakers were overwhelmed by the Depression and the popular belief that it arose from the collapse of speculation. Real bills nostrums filled the intellectual vacuum (Hetzel 2012 and 2014).

Moreover, the early Fed operated in a different environment than the Bank of England. Without understanding its environment, the early Fed did in fact create a regime of "managed money." The founders of the Fed were adamant that they were not creating a central bank, which was then understood as the Bank of England (Hetzel 2014; Lowenstein 2015). With the creation of the Federal Reserve, the United States went off the gold standard in that gold flows (external and internal inflows and outflows) no longer determined bank reserves and money in a systematic way. That situation persisted after the end of World War I and into the 1920s. However, because the United States maintained legal convertibility, there was little understanding of how the monetary regime had changed.

Early Fed policymakers assumed the United States was on a gold standard because it had maintained convertibility throughout World War I and the 1920s. However, this simulacrum of a gold standard in no way constrained money creation the way it did for the Bank of England. London was the center of the world gold market, and the Bank of England routinely set its Bank rate based on the way in which gold flows affected its reserve. In contrast, the individual Fed Regional Banks watched their gold cover (the required gold backing of

their note issue and deposits of member banks).[18] Because the United States ended World War I with a large fraction of the world's monetary gold, only on three occasions did the gold cover bind.[19] As a result, early Fed policymakers experimented with operating procedures for controlling the cost of funds to member banks but understood them in the context of their real bills views (Hetzel 2012 and 2014).

The critique that the bullionists applied to the Bank of England also applied to the early Fed. In the same way that the directors of the Bank of England failed to understand the need to replace their gold peg during restriction with a new nominal anchor, early Fed policymakers failed to realize that gold convertibility in itself in the absence of a discount rate tied to gold flows did not provide a nominal anchor that endowed money with a well-defined value in terms of goods.[20] In the absence of that understanding, what came through were real bills principles. In the *Report on the High Price of Bullion* (1810), quoted in Cannan (1919 [1969], 48-49), the bullionists argued:

> So long as the paper of the Bank was convertible into specie at the will of the holder, it was enough, both for the safety of the Bank and for the public interest in what regarded its circulating medium, that the Directors attended only to the character and quality of the Bills discounted, as real ones payable at fixed and short periods. They could not much exceed the proper bounds in respect of the quantity and amount of Bills discounted, so as thereby to produce an excess of their paper in circulation, without quickly finding that the surplus returned upon themselves in demand for specie. ... It was hardly to be expected of the Directors of the Bank, that they

[18] Morys (2010) documented that the peripheral countries of the gold standard made discount rate decisions based on their domestic gold cover rather than on the gold points, which are the ranges of tolerance of their paper currencies with their par gold value that set off exports or imports of gold.

[19] In the fall of 1919, the fall of 1931, and early 1933, the Regional Banks raised their discount rates in response to gold outflows.

[20] Under the gold standard, the real (goods) price of a gold coin containing a specified amount of gold alloy was determined by the market value of gold. As long as the Bank of England guaranteed convertibility at a fixed par value between its bank notes marked one pound and gold coins marked one pound, arbitrage equated the goods value of the Bank of England paper pound and the equivalent of a one-pound gold coin. If, say, the gold coin was more valuable in exchange, individuals could take paper pounds to the Bank of England and demand gold coins. That would reduce Bank of England gold reserves. The Bank of England would then raise its discount rate. As businesses contracted lending in response, the money stock declined and the price level fell. The fall in the price level increased the real (goods) value of the paper pound by increasing its purchasing power. Deflation continued until the paper pound had the same value in exchange as the gold coin minted as one pound. In this way, the price level (the goods price of the paper pound) was determined. Britain then had a "nominal anchor" that gave the paper pound a well-defined value in terms of its purchasing power (a well-defined price level) although not a stable one because the market price of gold varied.

> should be fully aware of the consequences that might result from their pursuing, after the suspension of cash payments, the same system which they had found safe before. ... [W]hile the convertibility into specie no longer exists as a check to the over issue of paper, the Bank Directors have not perceived that the removal of that check rendered it possible that such an excess might be issued by the discount of perfectly good bills.

In the United States, real bills views mixed with American populism. The unit banking system resulted in the pyramiding of reserves in which the reserves of country banks ended up held by correspondent banks in reserve and central reserve cities, especially New York City. The New York correspondent banks lent funds in the call money market, which financed the purchase of stocks on margin. The founders of the Fed believed that such lending financed speculation. In the last quarter of the nineteenth century in Britain, in contrast, branch banking replaced unit banking and, as a result, the need for a bill market to allocate funding from surplus to deficit regions disappeared. Bills of exchange became a primary instrument for financing world trade (King 1936, 268). There was then no association of the speculation assumed to produce boom-bust cycles with the concentration of bank reserves in London and with the presumed need to prevent them from spilling over into stock market speculation.

The reformers who created the Federal Reserve desired an "elastic" currency, that is, a currency that would expand *and* contract with the "needs of trade" so that excess credit would not spill over into speculation. In the United States, prior to the establishment of the Fed and under the National Banking Act of 1863, the issue of banknotes was inelastic because they had to be collateralized by government bonds. As a consequence, it was assumed that, when the demand for credit was cyclically low, an excess supply of credit would spill over into the speculative extension of credit. Extending credit based on real bills would proportion the supply of credit to the demand for productive uses and thus prevent the speculation, the collapse of which would lead to recession. (See Carter Glass, cited in Hetzel [2014, 175].)

Moreover, until March 1933, the Regional Fed Banks employed procedures in which member banks had to borrow from the discount window in order to obtain the marginal reserves they required and thus could be monitored for speculative lending (Hetzel 2012, Ch. 4). Because Fed policymakers interpreted the Great Depression as resulting from the collapse of a speculative bubble in land and equities, they maintained a contractionary monetary policy until forced to relinquish control of monetary policy to the Treasury in March 1933. Nothing in the early thinking of the Fed recognized its responsibility for the price

level or for setting an interest rate compatible with the required money creation (Hetzel 2012, Ch. 4).

In his commentary on the foreign exchanges, Hawtrey (1938, 34) commented that no one "would deny the fact of a great decline in manufacturing activity and consequent unemployment" following elevation of the Bank rate. However, he also noted that markets expected that the period of an elevated Bank rate would be short. In the case of Britain, a one-time deflation would depress the real terms of trade and restore balance of payments equilibria thereby arresting the external gold drain. Under the influence of real bills views, the Fed's founders believed that once past the credit liquidation required in order to eliminate a speculative lending mania, the economy would recover and grow again. Nothing in their experience prepared them for the Great Depression.

Of course, one cannot undo the mistakes that led to the Great Depression. What one can do is to ask that central banks employ a systematic procedure for learning from past mistakes. For that to happen, they need to operate with an analytical framework like the quantity theory that yields counterfactuals of different monetary policies. Asking that central banks use their models to learn from the past in a way that disciplines the present is a natural extension of transparency and accountability.

APPENDIX: A BRIEF OVERVIEW OF THE QUANTITY THEORY

The foundation of the quantity theory is the distinction between real variables and nominal variables. Real variables are the relative prices of goods and services (rates of exchange between them) and physical quantities. Nominal variables are quantities denominated in money (dollars or pounds), especially the stock of money and the price level (the money price of goods). Real variables are determined within the price system by tastes, technologies, and endowments. In contrast, in a world of fiat money, which possesses no intrinsic value, some constraint imposed from outside the price system (a nominal anchor) is required in order to give nominal variables well-defined values. The quantity theory explains the fundamentally different behavior of real and nominal variables, especially, the difference in the determination of relative prices and the price level.

Giving empirical content to the quantity theory requires distinguishing between financial intermediation and money creation. It also requires the assumption that the price system "works" in the sense that, despite changes in the price level, over time markets will separate changes in the price level from the determination of relative prices. One consequence is that markets will determine a real rate of interest—the natural rate of interest—based on real determinants. If the central bank sets its policy rate differently from the natural rate, in a way analogous to price fixing in an individual market, the central bank will engender money creation (destruction) that will destabilize the price level. Although destabilizing to real variables, as long as central bank interference with the price system is temporary, market-clearing relative prices will reemerge.

Finally, giving empirical content to the quantity theory entails understanding how the nature of the nominal anchor determines the equilibrating role of money and the price level. In a gold standard with no changes in the world supply and demand for gold that force changes in the commodity price of gold and with no changes for a country in its terms of trade, money becomes the equilibrating variable. Through gold flows arising out of the balance of payments, the money supply (gold) varies in order to accommodate money demand (Humphrey and Keleher 1982). In a pure fiat money standard or in a gold standard, if the central bank creates a discrepancy between its policy rate and the natural rate of interest, changes in money force changes in the price level.

REFERENCES

Attwood, Thomas. 1832. Memorandum in the Peel Papers in the British Museum, Add. MS. 40384.

Bagehot, Walter. [1873] 1962. *Lombard Street: A Description of the Money Market*. Reprinted Homewood, Ill.: Richard D. Irwin.

Bank of England. 2016. "Three Centuries of Macroeconomic Data." Version 2.3 [June 30].

Bordo, Michael D. 1990. "The Lender of Last Resort: Alternative Views and Historical Experience." Federal Reserve Bank of Richmond *Economic Review* 76 (January/February): 18–29.

Bordo, Michael D. 1999. *The Gold Standard and Related Regimes: Collected Essays*. Cambridge: Cambridge University Press.

Calomiris, Charles W., and Larry Schweikart. 1991. "The Panic of 1857: Origins, Transmission, and Containment." *Journal of Economic History* 51 (December): 807–34.

Cannan, Edwin. [1919] 1969. *The Paper Pound of 1797-1821*. London: P.S. King & Son.; reprinted New York: Augustus M. Kelley.

Capie, Forrest, and Geoffrey E. Wood. *The Lender of Last Resort*. New York: Routledge, 2007.

Capie, Forrest, and Geoffrey E. Wood. 2015. "The Development of the Bank of England's Objectives: Evolution, Instruction, or Reaction?" December 5.

Eichengreen, Barry. 1987. "Conducting the International Orchestra: Bank of England Leadership under the Classical Gold Standard." *Journal of International Money and Finance* 6 (March): 5–29.

Eltis, Walter. 1995. "John Locke, the Quantity Theory of Money and the Establishment of a Sound Currency." In *The Quantity Theory of Money: From Locke to Keynes and Friedman,* edited by Mark Blaug. Aldershot, U.K.: Edward Elgar, 4–26.

Fetter, Frank W. 1965. *Development of British Monetary Orthodoxy, 1797-1875*. Cambridge, Mass.: Harvard University Press.

Fisher, Irving. 1935. *100% Money*. New York: Adelphi.

Flandreau, Marc. 2008. "Pillars of Globalisation: A History of Monetary Policy Targets, 1797-1997." In *The Role of Money: Money and Monetary Policy in the 21st Century*, edited by Andreas Beyer and Lucrezia Reichlin. Frankfurt-am-Main, Germany: European Central Bank, 208–43.

Flandreau, Marc, and Stefano Ugolini. 2013. "Where It All Began: Lending of Last Resort at the Bank of England Monitoring During the Overend-Gurney Panic of 1866." In *The Origins, History, and Future of the Federal Reserve*, edited by Michael D. Bordo and William Roberds. New York: Cambridge University Press, 113–61.

Friedman, Milton and Anna J. Schwartz. 1963. *A Monetary History of the United States, 1867-1960*. Princeton: Princeton University Press.

Great Britain. 1810. *Report from the Select Committee on the High Price of Gold Bullion*. Parliamentary Papers, Commons. (349), III.

Hawtrey, Ralph G. 1938. *A Century of Bank Rate*. London: Longmans, Green and Co.

Hawtrey, Ralph G. 1950. *Currency and Credit*, 4th ed. London: Longmans, Green and Co.

Hetzel, Robert L. 1985. "The Rules versus Discretion Debate over Monetary Policy in the 1920s." Federal Reserve Bank of Richmond *Economic Review* 71 (November/December): 3–14.

Hetzel, Robert L. 1987. "Henry Thornton: Seminal Monetary Theorist and Father of the Modern Central Bank." Federal Reserve Bank of Richmond *Economic Review* 73 (July/August): 3–16.

Hetzel, Robert L. 2012. *The Great Recession: Market Failure or Policy Failure?* Cambridge: Cambridge University Press.

Hetzel, Robert L. 2014. "The Real Bills Views of the Founders of the Fed." Federal Reserve Bank of Richmond *Economic Quarterly* 100 (Second Quarter): 159–81.

History of Economic Thought. *The Bullionist Controversy*. www.hetwebsite.net/het/schools/bullion.htm.

Hume, David. [1752] 1987. "Of the Balance of Trade." In *Essays Moral, Political, Literary*, edited by Eugene F. Miller. Indianapolis, Ind.: Liberty Fund.

Hume, David. [1752] 1955. "Of Money." In *Writings on Economics*, edited by Eugene Rotwein. Freeport, N.Y.: Books for Libraries Press.

Humphrey, Thomas M. 1974. "The Concept of Indexation in the History of Economic Thought." Federal Reserve Bank of Richmond *Economic Review* 60 (November/December): 3–16.

Humphrey, Thomas M. 1975. "The Classical Concept of the Lender of Last Resort." Federal Reserve Bank of Richmond *Economic Review* 61 (January/February): 2–9.

Humphrey, Thomas M. 1977. "Two Views of Monetary Policy: The Attwood-Mill Debate Revisited." Federal Reserve Bank of Richmond *Economic Review* 63 (September/October): 14–22.

Humphrey, Thomas M. 1982a. "The Real Bills Doctrine." Federal Reserve Bank of Richmond *Economic Review* 68 (September/October): 3–13.

Humphrey, Thomas M. 1982b. "Of Hume, Thornton, the Quantity Theory, and the Phillips Curve." Federal Reserve Bank of Richmond *Economic Review* 68 (November-December): 13–18.

Humphrey, Thomas M. 1983. "Can the Central Bank Peg Real Interest Rates? A Survey of Classical and Neoclassical Opinion." Federal Reserve Bank of Richmond *Economic Review* 69 (September/October): 12–21.

Humphrey, Thomas M. 1986. "Cumulative Process Models from Thornton to Wicksell." Federal Reserve Bank of Richmond *Economic Review* 72 (May/June): 18–25.

Humphrey, Thomas M. 1989. "Lender of Last Resort: The Concept in History." Federal Reserve Bank of Richmond *Economic Review* 75 (March/April): 8–16.

Humphrey, Thomas M. 1990a. "Fisherian and Wicksellian Price-Stabilization Models in the History of Monetary Thought." Federal Reserve Bank of Richmond *Economic Review* 76 (May/June): 3–12.

Humphrey, Thomas M. 1990b. "Ricardo versus Thornton on the Appropriate Monetary Response to Supply Shocks." Federal Reserve Bank of Richmond *Economic Review* 76 (November/December): 18–24.

Humphrey, Thomas M. 1991. "Nonneutrality of Money in Classical Monetary Thought." Federal Reserve Bank of Richmond *Economic Review* 77 (March/April): 3–15.

Humphrey, Thomas M. 1999. "Mercantilists and Classicals: Insights from Doctrinal History." Federal Reserve Bank of Richmond *Economic Quarterly* 85 (Spring): 55–82.

Humphrey, Thomas M. 2010. "Lender of Last Resort: What It Is, Whence It Came, and Why the Fed Isn't It." *Cato Journal* 30 (Spring/Summer): 333–64.

Humphrey, Thomas M., and Robert E. Keleher. 1982. *The Monetary Approach to the Balance of Payments, Exchange Rates, and World Inflation.* New York: Praeger Publishers.

Humphrey, Thomas M., and Robert E. Keleher. 1984. "The Lender of Last Resort: A Historical Perspective." *Cato Journal* 4 (Spring/Summer): 275–321.

Jevons, William Stanley. 1876. *Money and the Mechanism of Exchange*. London: D. Appleton.

Joplin, Thomas. [1823] 1970. *Outlines of a System of Political Economy*. Reprinted New York: Augustus M. Kelley.

Joplin, Thomas. 1832. *An Analysis and History of the Currency Question*. London: James Ridgway.

King, W.T.C. 1936. *History of the London Discount Market*. London: Routledge.

Kynaston, David. 1995. "The Bank of England and the Government." In *The Bank of England: Money, Power and Influence 1694-1994*, edited by Richard Roberts and David Kynaston. Oxford: Clarendon Press, 19–55.

Laidler, David. 2000. "Highlights of the Bullionist Controversy." University of Western Ontario Department of Economics Research Report No. 2000-2 (March).

Leigh, Arthur H. 1974. "John Locke and the Quantity Theory of Money." *History of Political Economy* 6 (Summer): 200–19.

Link, Robert G. 1959. *English Theories of Economic Fluctuations, 1815-1848*. New York: Columbia University Press.

Locke, John. [1695] 1968. "Further Considerations Concerning Raising the Value of Money." In *Several Papers Relating to Money, Interest, and Trade*. Reprinted New York: Augustus M. Kelley.

Locke, John. [1823] 1963. *The Works of John Locke,* vol. 5. London: Th. Tegg; reprinted Aalen: Scientia.

Lowe, Joseph. [1823] 1967. *The Present State of England*, 2nd ed. Reprinted New York: Augustus M. Kelley.

Lowenstein, Roger. 2015. *America's Bank: The Epic Struggle to Create the Federal Reserve*. New York: Penguin Press.

Macleod, Henry Dunning. 1866. *The Theory and Practice of Banking*, vol. II. London: Longmans, Green, Reader, and Dye.

Mazumder, Sandeep, and John H. Wood. 2012. "The Quantity Theory and the Gold Standard." Manuscript, Wake Forest University.

Meltzer, Allan H. 2003. *A History of the Federal Reserve,* vol. 1, 1913-1951. Chicago: University of Chicago Press.

Mill, John Stuart. [1848] 1909. *Principles of Political Economy with some of their Applications to Social Philosophy*. Edited by William J. Ashley. London; Longmans, Green and Co. http://www.econlib.org/library/Mill/mlP.html.

Mints, Lloyd W. 1945. *A History of Banking Theory in Great Britain and the United States*. Chicago: University of Chicago Press.

Morys, Matthias. 2010. "Monetary Policy under the Classical Gold Standard (1870s–1914)." Manuscript, University of York.

Ricardo, David. [1810, 1821, 1824] 1951. "The High Price of Bullion (1810)"; "On the Principles of Political Economy and Taxation (1821)"; "Plan for the Establishment of a National Bank (1824)." In *Works and Correspondence of David Ricardo*, edited by Piero Sraffa. Cambridge: Cambridge University Press for the Royal Economic Society.

Ricardo, David. 1822. "Mr. Ricardo's Speech on Mr. Western's Motion for a Committee to Consider of the Effects Produced by the Resumption of Cash Payments, delivered the 12th of June, 1822." London: G. Harvey.

Roberds, William. 2016. "Review of *Making Money: Coin, Currency, and the Coming of Capitalism* by Christine Desan." *Journal of Economic Literature* 54 (September): 906–21.

Sayers, R.S. 1976. *The Bank of England 1891-1944*. Cambridge: Cambridge University Press.

Scrope, G. Poulett. [1833] 1968. *Principles of Political Economy*. Reprinted New York: Augustus M. Kelley.

Schumpeter, Joseph A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Sidgwick, Henry. 1883. *The Principles of Political Economy*. London: Macmillan and Co.

Temin, Peter. 1969. *The Jacksonian Economy*. New York: W.W. Norton and Co.

Thornton, Henry. [1802, 1811] 1939. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain (1802) and Two Speeches (1811)*, edited with an Introduction by F A. v. Hayek. New York: Rinehart and Co.

Timberlake, Richard H. 1993. *Monetary Policy in the United States: An Intellectual and Institutional History*. Chicago: University of Chicago Press.

Tullock, Gordon. 1957. "Paper Money – A Cycle in Cathay." *Economic History Review* 9 (April): 393–407.

Viner, Jacob. 1937. *Studies in the Theory of International Trade*. New York: Harper & Brothers Publishers.

Warburg, Paul M. 1910. "The Discount System in Europe." Washington, D.C.: National Monetary Commission/Washington Government Printing Office. Available at https://fraser.stlouisfed.org/docs/historical/nmc/nmc_402_1910.pdf.

Wicksell, Knut. [1935] 1978. *Lectures on Political Economy*, vol. 2. Fairfield, N.J.: Augustus M. Kelley.

Wood, John H. 2002. "Notes for Money and Banking." Manuscript, Wake Forest University.

Wood, John H. 2005. *A History of Central Banking in Great Britain and the United States*. New York: Cambridge University Press.

Economic Quarterly—Volume 102, Number 4—Fourth Quarter 2016—Pages 321–351

Private Efforts for Affordable Mortgage Lending Before Fannie and Freddie

David A. Price and John Walter

A number of federal government initiatives in the United States have sought both to make home mortgages more broadly available and to increase the availability of features rendering those mortgages more affordable to borrowers, such as lower interest rates, long-term fixed rates, and lower down payments. Most notable among these initiatives have been the government-sponsored enterprises (GSEs) Fannie Mae, created in 1938 in response to the Great Depression, and Freddie Mac, established in 1970.[1]

In the period prior to the advent of Fannie Mae, private activities played an important role in improving the affordability of U.S. mortgage markets, likely lowering interest rates as well as producing more favorable noninterest terms. Two examples of such activities are mortgage-backed securities (MBS) that arose in the late nineteenth century and the building and loan associations that first appeared in the early nineteenth century. Both of these financial arrangements were modeled after similar ones that appeared previously in Europe. In addition, large life insurance companies competed with other institutions for home mortgage lending business and grew to become important nationwide mortgage lenders between the 1880s and 1920s.

The authors thank Jackson Evert, Arantxa Jarque, Bruno Sultanum, and John Weinberg for helpful comments. The views in this paper are those of the authors and not necessarily those of the Federal Reserve Bank of Richmond or of the Federal Reserve System.

[1] Such public efforts in the United States have been numerous, varied, and ongoing for a century. Edson (2011), p. 3.

These historical activities of the private sector are of interest in two respects. First, they reflect a range of responses to the classic tension in mortgage lending: seeking the benefits of portfolio diversification and efficiencies of scale by pooling risks across regions, on one hand, versus seeking the benefits of local market knowledge and more effective oversight of agents by lending on a local scale, on the other. The GSEs have sought to manage this tension by combining national-level portfolios with measures such as imposing standardized underwriting requirements and demanding representations and warranties[2] from mortgage originators. In addition, the implicit public guarantee of the GSEs may have helped them paper over the tension to some extent until the crisis of 2007–08. Historical private-sector responses to the tension prior to the GSEs—and prior to the emergence of sophisticated information technology that has facilitated national-level mortgage lending and securitization—may be instructive. As will be seen, these private-sector institutions were not wholly successful in addressing the tensions either.

Second, the emergence and subsequent record of MBS issuers, building and loan associations, and life insurer mortgage operations of the nineteenth and early twentieth centuries could be suggestive of the types of institutions that would develop if GSEs were to become a less significant part of the mortgage landscape—through the operation of public policy or otherwise—and could shed light on the likely strengths and weaknesses of those emergent institutions.

1. PRIVATE MORTGAGE-BACKED SECURITIES AND EQUIVALENTS

MBS allow investors to achieve geographical diversification and open a broader pool of funds for mortgage borrowers. MBS in the United States have a long but checkered history extending back to the 1870s. The earliest MBS were loosely modeled after European mortgage banks, which issued the equivalent of today's *covered bonds*, backed by mortgages.[3] These European institutions were often created by the government, in some cases were granted monopoly power by the government,

[2] Fannie Mae has explained, "Representations and warranties are a lender's assurance to the GSE that the GSE can rely on certain facts and circumstances concerning the lender and the mortgage loans it is selling. ... Violation of any representation and warranty is a breach of the Lender Contract, entitling Fannie Mae to pursue certain remedies, including a loan repurchase request." Fannie Mae, "Selling Guide Announcement SEL-2012-08," September 11, 2012, p. 1.

[3] In modern parlance, covered bonds are backed by a pool of assets (often mortgages) on the balance sheet of the bond issuer. The bonds are "covered" in that they are collateralized (i.e., covered) by the pool of assets.

and operated under strict government rules. The European structures became important and long-lasting entities in the housing and building finance industries likely, in part, because of government aid that propped them up during episodes of financial distress. In the United States, issuers of MBS in the late nineteenth and early twentieth centuries suffered several episodes of wide-scale default but, in contrast to the earlier European experience, were not rescued by the public sector.

Early European Mortgage Banks

One of the earliest examples of MBS arose in Europe when, in 1770, Frederick the Great, king of Prussia, called for the creation of Landschaften mortgage-lending institutions.[4] This first Landschaft was formed in 1770, soon after the end of the Seven Years War in 1763, and was located in the Prussian province of Silesia. Landschaften arose as a means of providing credit for agricultural production—for example, for the purchase of seed, horses, and cattle. The Seven Years War, and government credit policy actions following the war, had interrupted traditional credit channels.[5] Landschaften were compulsory corporations including all land-owning nobles of a region.[6] Landowners submitted their land as collateral, borrowed from the corporation, and the corporation sold bonds to investors to fund the loans. The Landschaften bonds were the liabilities of the corporation, but members were also jointly responsible for repayment of the bonds.[7]

A number of other Prussian provinces soon followed suit in setting up their own Landschaften.[8] This arrangement shared two important features with later mortgage institutions in Europe: (1) government support for the formation and for the risk-limiting characteristics of the association (such as loan-to-value limits and restrictions on activities that might diminish the value of collateral) and (2) the creation of liabilities that were backed by a large portfolio of mortgages, producing a diversified source of income to support bond payments. Initially, between 1770 and 1830, Landschaften mortgage agreements provided only

[4] Tcherkinsky (1922), pp. 13, 14, 22; Snowden (1995b), p. 270.

[5] Wandschneider (2015), p. 794.

[6] Tcherkinsky (1922), p. 14. Wandschneider (2014), pp. 312–13, argues that the requirement that all landowners participate reduced the risk of an adverse selection problem limiting the attractiveness of Landschaft bonds. Wandschneider notes that "Adverse selection is an ex ante informational problem where under certain conditions only borrowers that are a poor credit risk will be attracted into a market. In response, lenders will not be willing to supply capital to this pool of 'lemons.'"

[7] Tcherkinsky (1922), p. 13.

[8] Tcherkinsky (1922), p. 15.

for annual interest payments with no clearly specified repayment date.[9] When repayment occurred it was either by repurchase of Landschaften bonds or with cash payments. But in the 1830s, the Landschaften introduced amortizing mortgage loans, whereby borrowers paid interest plus principal repayments that extinguished (amortized) their debt over time.[10]

The Landschaften were closely aligned with the government, likely encouraging the view that support would be forthcoming if they experienced financial trouble. Early Landschaften were begun using government-provided capital, and the president of the organization was chosen by the king.[11] Some of the employees of Landschaften were chosen by local government assemblies, were sworn in and faced government discipline, and had the standing of state employees, including facing reduced taxes like other state employees.[12] In at least one case the government borrowed from a regional Landschaft (East Prussia), against government-owned lands, to cover war-related expenses.[13] The expectation of government support likely explains, in part, the low interest rates paid on Landschaften bonds and paid by the borrowers funded by Landschaften mortgages.

Indeed, Landschaften investors' belief that the government would protect their bond holdings seems to have been confirmed when between 1820 and 1830 the government came to the aid of troubled Landschaften in East Prussia and West Prussia.[14] Therefore, it seems possible that without government aid, the Landschaft experience would have been similar to the later experience of U.S. mortgage companies and MBS issuers in the 1890s and during the Great Depression (discussed in the next section) that suffered pervasive failures but did not receive government aid.

These organizations survived widespread economic turmoil during the Napoleonic Wars (1803–15) and agricultural crisis during the 1820s and were operating until the end of World War II.[15] Even fairly early in their history, they were significant lenders. For example, about one-third of land-owning estates in East Prussia had outstanding loans

[9] Wandschneider (2015), p. 317.

[10] Tcherkinsky (1922), pp. 33-34.

[11] Wandschneider (2015), pp. 794–95.

[12] Tcherkinsky (1922), p. 26.

[13] Wandschneider (2015), p. 800.

[14] Tcherkinsky (1922), pp. 43–44; Wandschneider (2015), pp. 800–01.

[15] Wandschneider (2015), p. 815; Wandschneider (2014), p. 307. Tcherkinsky (1922), pp. 22–23, notes that at the time of his writing, there were "in Germany ... 21 credit institutions of the Landschaft type." Tcherkinsky also provides a chronological table listing the location and year of formation (from 1770 through 1895) of rural Landschafts.

from Landschaften as of 1823.[16] The amount of borrowing from Landschaften increased significantly in the latter half of the nineteenth century.[17] Wandschneider (2015) argues that Landschaften were responsible for lowering the cost of credit for agricultural estates and increasing the value of the estates that could borrow from Landschaften. Rates on Landschaften bonds were similar to rates on government bonds and they were popular investments in Prussia and internationally, thus providing an extensive source of funding for Prussian mortgage borrowers.

U.S. Mortgage Companies and Private MBS

MBS-issuing institutions in the United States arose in the 1870s and filled a niche for a nationally diversified source of funds for home and farm mortgages. Specialized mortgage lenders, such as United States Mortgage Company, provided mortgages and issued MBS, but mortgage insurance companies also established trusts that purchased mortgages and issued MBS, employing the model that was later adopted by Fannie Mae and Freddie Mac.

While in the eighteenth century Landschaften provided an internationally derived source of funds for Prussian borrowers, even in the late 1800s lending by institutions accounted for far less than half of U.S. mortgage lending, so the United States seemed ripe for growth of institutions that could provide these diversification and funding-source-widening benefits. (In some parts of the country, lending institutions did play an important role—especially in New England and the Pacific states.)[18] An apparent difference between the U.S. and the European experience is that in Europe the development of regional or nationwide mortgage markets had been encouraged, regulated, and subsidized, while in the United States such markets had, to a degree, been discouraged by legislation that limited the range of banks and some other potential lenders.

Because some commercial banks, savings banks, mutual savings banks, and life insurance companies were, in many cases, prohibited from mortgage lending on an interstate basis, other entities not subject to these prohibitions filled the gap to provide a means of diversified (nationwide) mortgage lending along with local credit analysis. These entities were mortgage companies or mortgage trusts and began being formed in the early 1870s.[19]

---

[16] Wandschneider (2015), p. 805.
[17] Wandschneider (2015), p. 318.
[18] Snowden (1995a), p. 220.
[19] Brewer (1976), pp. 358–61.

National banks (those banks chartered by the federal government rather than state government) were prohibited from investing in mortgages by the National Bank Act of 1864. This prohibition remained in place until 1913.[20] New York-headquartered life insurance companies, which held 50 percent of all U.S. life insurance assets, could invest only in mortgages on properties within that state or within fifty miles of New York City until at least the late 1870s.[21] Mutual savings banks, significant members of the banking community in the northeastern portion of the United States, were typically limited to making mortgages on properties in their home states in the late nineteenth century.[22] Given limited interstate communication and transportation technology in the nineteenth century, such restrictions were likely viewed by supervisors as reasonable limitations for safety and soundness purposes or perhaps as a way to ensure that deposits gathered locally were also invested locally, but the restrictions probably significantly limited competition for mortgage loans outside of the Northeast.

One of the mortgage entities that arose in this environment that restricted bank and insurance company mortgage lending was United States Mortgage Company, which is described in detail by Brewer (1976).[23] The company was chartered by legislation passed by the state of New York in 1871. The company lent to mortgage borrowers—both residential and farm—in the United States and issued bonds equal to its mortgage holdings.[24] It offered borrowers the option of paying off their loans in installments (i.e., an amortizing loan) or paying in full at maturity.[25] Bonds issued by United States Mortgage Company had maturities of five to fifty years.[26] Given that the company's securities—or bond issues—were backed by mortgage loans, these issues amounted to nineteenth-century MBS. These MBS were the liabilities of the United States Mortgage Company, much as today's Fannie Mae- and Freddie

[20] Davis (1965), p. 358.

[21] Davis (1965), p. 383; Brewer (1976), p. 358 and footnote 11. Brewer (1976), p. 358, notes that mortgages accounted for 54 percent of life insurance company assets in 1875.

[22] Brewer (1976), pp. 358–59.

[23] See Brewer (1976), pp. 362–72. United States Mortgage Company was ultimately absorbed by Chemical Bank, which merged with Chase Manhattan Bank in 1996, keeping the Chase name. Chase and J.P. Morgan merged in 2000 to form today's JPMorgan Chase. See Brewer (1976), p. 363, footnote 25; Hansell (1995); and JPMorgan (2017).

[24] Brewer (1976), p. 363.

[25] Brewer (1976), p. 363.

[26] Brewer (1976), pp. 364–65.

Mac-issued MBS are the liabilities of Fannie Mae and Freddie Mac.[27] United States Mortgage Company MBS were held by investors in the United States and in Europe.

Mortgage bonds were attractive investments at the time because of the limited number of competitor securities. For example, U.S. Treasury securities were paying an unusually low rate of interest at the time because banks were required by law to hold them to back their currency issues, thus creating heavy demand for Treasuries and driving down the interest rates they paid. The other main competitor bonds were railroad bonds, and those were disfavored by investors in the mid-1870s due to widespread bankruptcies by railroad companies.[28] United States Mortgage Company created earnings by lending at an interest rate that exceeded the interest rate it paid on its bonds.[29]

Lending by United States Mortgage focused heavily on western mortgages with lending boards created in Chicago and St. Louis. Bonds were sold in Europe, through a Paris office, and were also listed on the New York Stock Exchange beginning in 1874. The company established local lending "boards" to handle mortgage loan origination, pricing, and credit quality.[30]

But United States Mortgage was not alone. According to Snowden (1995b), seventy-four western mortgage companies were selling mortgage backed securities (mostly based on farm mortgages) in Massachusetts and New York between 1890 and 1897, and their issues amounted to $800 million at a time when total mortgage debt outstanding was about $6 billion.[31] The issuers of these bonds also guaranteed them against default risk. Still, by 1897 most of these entities had failed (many by defaulting on their securities issues) due to a decline in western land values.

MBS, created by insurance companies, and with structures almost identical to those used by Fannie Mae and Freddie Mac today (whereby securities representing a proportional cash flow of underlying mortgages are sold to investors, with the seller providing default insurance

---

[27] Before 2010, MBS issued by Fannie Mae and Freddie Mac were guaranteed (in terms of principal and interest payments) by Fannie Mae and Freddie Mac but not shown on their balance sheets as their liabilities. Following an accounting rule change that took effect in 2010, these MBS are now shown as Fannie Mae and Freddie Mac liabilities. One difference between Fannie Mae and Freddie Mac and United States Mortgage Company is that the latter actually made the mortgage loans itself, while Fannie Mae and Freddie Mac buy mortgages from outside lenders.

[28] Brewer (1976), pp. 359–60.

[29] Brewer (1976), p. 360.

[30] Brewer (1976), p. 364.

[31] See Snowden (1995b), p. 278; Snowden (1995a), p. 220.

for the securities), arose and grew large in New York in the 1920s.[32] These MBS structures developed over a period of forty years starting first in the late 1880s when several New York-headquartered companies formed for the purpose of guaranteeing mortgage payments due to mortgage investors and providing title insurance. This mortgage-guarantee business was small until after World War I, when a boom in construction caused a rapid increase. In 1921, New York firms guaranteed $500 million worth of loans and by 1932, $2.8 billion.[33] The latter figure compares to $24.9 billion in outstanding residential mortgages in 1932.[34]

At first, the mortgage payment and title insurance companies simply provided default insurance on mortgage payments. But in 1906 companies began selling participation certificates in guaranteed (in terms of principal and interest payments) mortgage pools—the same MBS structure employed by Fannie Mae and Freddie Mac today. By 1933, the outstanding amount of mortgage-participation certificates was $810 million, which had been sold to 213,000 separate investors.[35]

During the Great Depression, rapidly declining house prices and homeowner incomes meant that most of the guarantee companies failed and many participation certificate investors suffered proportionally large losses on their investments. Ultimately, following the Depression-era failures of these structures, many as a result of weak underwriting standards of the lenders, federal and state laws were passed that prohibited mortgage insurance, a fundamental feature of the structures, for the next two decades.[36]

2. BUILDING AND LOAN ASSOCIATIONS

Another private effort to lower the cost of housing prior to the GSEs was a form of thrift institution known as building and loan associations. They were based on notions of mutual self-help, that is, self-reliance combined with mutual aid: individuals held shares in the institutions and, in return, had borrowing privileges as well as the right to dividends. Broadly speaking, while operating plans varied, members committed to make regular payments into the association and took turns taking out mortgages with which to buy homes; the determination of the next borrower was often decided by an auction among the

[32] Snowden (1995b), p. 283–88.
[33] Alger and Cook (1934), pp. 7–9.
[34] Grebler, Blank, and Winnick (1956), p. 443.
[35] Alger (1934), p. 3.
[36] Snowden (1995b), pp. 283, 285–86.

membership. From their advent in the 1830s until their demise during the Great Depression, building and loan associations were generally small and local. At the peak of their numbers in 1927, some 12,804 of the associations were in operation with 11.3 million members—at a time when the entire U.S. population was only 119 million—and $7.2 billion in assets.[37] In addition, a rival group of "national" building and loans was a significant force from the 1880s until the late 1890s.

Because the primary purpose of a building and loan was to make home mortgages accessible to its members, they developed loan products with payment terms that were more attractive to typical homebuyers. Where mortgages from commercial banks during the 1920s had an average length of three years and were nonamortized, those from buildings and loans averaged eleven years and 95 percent were self-amortizing.[38]

Early Development and Diffusion

American building and loan associations had their roots in British building societies, which appear to have originated in Birmingham, England, in the 1770s or 1780s.[39] At least a dozen of the societies were founded in Birmingham in the last quarter of that century.[40] These increased to sixty-nine societies by 1825 and then proliferated rapidly to 2,050 by 1851.[41] In general, members bought shares and paid for them over time and rotated receiving home loans—until all the members had taken a turn, at which point a society terminated.[42]

The British working class at the time already had a longtime tradition of "friendly" societies, cooperatives of mutual self-help to which members would make regular payments and from which they could receive a loan in the event of certain hardships, such as fire, job loss, or sickness.[43] Conceptually, it was perhaps a short distance from the institution of the friendly society to that of the building society. Britain in the nineteenth century may also have been fertile soil for building

[37] Bodfish (1931), p. 136. At that time, the total residential mortgage debt held by all lenders was approximately $24.4 billion. Grebler, Blank, and Winnick (1956), p. 466. While 1927 was the peak year for the number of associations, the number of members and total assets continued to increase briefly.

[38] White (2014), p. 136. Although longer loan terms likely made the loans more costly in the aggregate given the greater interest expense, they were desirable and more affordable in the sense that they resulted in lower monthly payments.

[39] Mason (2004), p. 14; Bodfish (1931), p. 11; Price (1958), p. 20.

[40] Price (1958), p. 21.

[41] Mason (2004), p. 15.

[42] Mason (2004), p. 14.

[43] Mason (2004), p. 13.

societies because ideas of mutual self-help were in the air more generally in other settings. Mutual-improvement societies, for example, were groups of working-class men who combined money to buy reading material that they shared for discussion at meetings.[44]

The conditions that apparently drove the application of these ideas to homebuying were created by the Industrial Revolution. The rise of factory work meant, for many, regular wage incomes. Higher-skilled workers with relatively greater incomes might wish to purchase a home to avoid tenement-like conditions and to gain the accumulation of equity possible through mortgaging rather than leasing. (In addition, homeownership brought with it the right to vote for one's representative in Parliament.) But those workers were stymied by the conventional mortgage offerings of the time with their high down payment requirements and short loan terms.[45] The British building society enabled some to overcome these obstacles.

The building society model appears to have been transmitted from Britain to the United States by British immigrants. The first building and loan association, Oxford Provident Building Association, was founded in Frankford, Pennsylvania, (now part of Philadelphia) in 1831 by two factory owners who were natives of England.[46] The model spread from there to the northeast and mid-Atlantic, with associations established in Connecticut, Maryland, New Jersey, and New York by 1850, along with additional associations in Pennsylvania.[47] (In addition, several associations were established in Charleston, South Carolina, during this period, at least one of them founded by an English immigrant.[48]) Associations were established in the majority of other states during the 1860s and 1870s. Illinois, California, and Texas leapfrogged other states outside the East Coast, with associations established in 1851, 1865, and 1866, respectively, a pattern that may have been the result of westward migration of individuals who were familiar with the model.[49]

As in Britain, the growth of building and loan associations in the United States was likely aided by the factory system and the swelling of a wage-earning class—combined with a dearth of affordable financing sources for homebuyers.[50] As noted earlier, under the National Bank Act of 1864, national banks were not permitted to make loans

[44] Griffin (2013), pp. 174, 177.
[45] Mason (2004), pp. 13-14; Price (1958), pp. 130-31; Foulke (1941), pp. 146-47.
[46] Haveman and Rao (1997), p. 1608; Foulke (1941), p. 147.
[47] Foulke (1941), p. 182; Bodfish (1931), pp. 76-83.
[48] Bodfish (1931), pp. 562-64.
[49] Mason (2004), p. 29; Bodfish (1931), pp. 81, 84.
[50] Foulke (1941), p. 146.

secured by real estate.[51] Mortgages from state commercial banks required large down payments, up to 60 percent of the home's value, and the loans were short-term (typically five years or less) and nonamortized. Mutual savings banks—which, notwithstanding the name, were not cooperatively owned—offered longer loan terms than commercial banks, but their mortgages still involved high down payments. Insurance companies—another source of mortgage finance in the nineteenth century, as discussed more fully below—also required high down payments.[52]

In the early decades of American building and loan associations, during the first half of the nineteenth century, they closely followed the form of operation of the British building societies. This model came to be known as the "terminating plan," so named because an association's existence was required to be wound up at a predefined point—when all of its loans had been repaid, or more precisely, when the shares of stock that members purchased over time in connection with membership had matured.[53]

An illustration of how the terminating plan worked, taken from that of the Oxford Provident association, is the following.[54] The building and loan would be formed by a group of individuals (members), each of whom paid a membership fee of $5 at the time of formation. Each member also subscribed to a number of shares of stock—between one and five shares—with a predetermined maturity value or par value of, say, $500. Then each member was required to pay in $3 per month per share until the amount paid in per share equaled the shares' maturity value. In general, no other members were allowed to join unless they paid, up front, an amount equal to that already paid in by the founding members. Once members' payments reached the maturity value of the shares, the association was terminated and members were repaid.

While the association was operating, members could pledge their stock and thereby take out home mortgage loans equal to as much as the matured value of all their shares of stock (though at the time of the loan, the member might have paid in much less than this amount). For example, if a member had subscribed to five shares, each with a maturity value of $500, the member could borrow as much as $2,500. (The borrower pledged his or her stock when taking out a mortgage, then continued paying for the stock on an installment plan until the

[51] Behrens (1952), p. 15.
[52] Mason (2004), pp. 16-17.
[53] Bodfish (1931), pp. 85-86.
[54] Byers (1927), p. 20; Bodfish (1931), pp. 35-36.

stock was paid for, which had the effect of cancelling the loan.[55]) In the rotation of home loans, members who wished to receive the next loan bid against one another; the bidding determined the premium that the winner would pay to secure that place in the rotation. Most commonly, the amount of the premium would be deducted from the loan when it was disbursed.[56]

The relative simplicity of the terminating plan made it an attractive framework for the associations during the first decades of the movement. A difficulty of the terminating plan, however, is that it was burdensome for members to join once an association was underway: as noted, all shares were issued at the same time, so members who joined later were required to pay a lump sum on entry to cover the payments they had missed. (In modern terms, a terminating plan was "closed end" in the sense that it generally issued shares only at its inception.) Moreover, in the waning period of the association's life, an association with idle money to lend and no borrower to take it might require a member (chosen by lot) to accept a loan whether he wanted it or not. Finally, the automatic termination of an association was perceived by some as wasteful given the efforts involved in organizing it and its potential usefulness if it were a continuing concern.[57]

The 1850s saw the emergence of a variation on the terminating plan that partially addressed these shortcomings. An association organized under the "serial plan" issued multiple series of shares over the course of its existence. In effect, a serial-plan association was like a collection of terminating-plan groups, each with its own onset and termination dates, under one organizational umbrella. New series were commonly offered on a quarterly or semiannual schedule. Thus, someone who had not been a member at the association's birth could join when the association later issued a new series of shares without the obstacle of making a back payment. Because the association was periodically adding member-borrowers to its rolls, there was no need to require someone to take an unwanted loan. Finally, the association as a whole had no defined termination date.[58]

A third form of organization, the permanent plan, arose in the 1870s. It did away with the concept of series of shares and instead issued shares to each member that were independent of the shares of other members; consequently, members could join and leave at the

[55] Dexter (1889), pp. 316-17.

[56] Wrigley (1869), pp. 29, 71.

[57] Bodfish (1931), p. 86; Mason (2004), pp. 18-19.

[58] Bodfish (1931), p. 87; Mason (2004), p. 19; Mason (2012), p. 382; Snowden (1997), p. 231; Foulke (1941), pp. 182-83

times of their own choosing.[59] As noted by Haveman and Rao (1997), the structural evolution from the terminating plan to serial and then permanent plans enabled building and loans to serve a sometimes transient homebuying population with less burdensome, more flexible arrangements.[60]

Still another plan, specific to the city of Philadelphia, had a separate track of development. There are conflicting accounts of when it originated, but a majority of sources point to the first half of the nineteenth century.[61] Under the Philadelphia plan, the homebuyer making a 20 percent down payment financed the other 80 percent by taking out a first mortgage for 50 percent of the purchase price from a bank, insurance company, or other lender, together with a second mortgage for 30 percent of the purchase price from a building and loan association. The result of this arrangement was low monthly payments: on the first mortgage—which typically had a three- to five-year term but could readily be renewed—the purchaser made interest-only payments. On the second mortgage, the purchaser made full self-amortizing payments, but the loan term was longer, typically eleven years.[62]

The Philadelphia plan was the predominant method of home finance within that city. It saw little adoption elsewhere, however, perhaps in part because most states did not allow a building and loan association to hold a second mortgage on a property for which it did not also hold the first mortgage; in those states, evidently, the second-mortgage business was considered too risky for building and loans.[63] In any event, the Philadelphia plan represented a distinctive and successful model of affordable lending, apparently contributing to the city's high rate of homeownership.[64]

With the further increase in U.S. urbanization in the 1880s, building and loan associations experienced a major wave of growth; thousands of local associations were founded.[65] Associations spread into every state during this decade (except Oklahoma, which saw its first building and loan in 1890).[66] By 1893, according to a survey taken by the U.S. commissioner of labor, there were 5,598 local associations

---

[59] Bodfish (1931), pp. 93-94; Foulke (1941), p. 183.

[60] Haveman and Rao (1997), p. 1638.

[61] Loucks (1929), pp. 7-8.

[62] Loucks (1929), pp. 1, 6.

[63] Loucks (1929), pp. 1-2, 6.

[64] Among the eleven U.S. cities with a population above 500,000, Philadelphia in 1920 ranked second in its percentage of owner-occupied homes (39.5 percent). Loucks (1929), p. 39.

[65] Snowden (1997), p. 228.

[66] Bodfish (1931), p. 81.

with a total of 1,349,437 members and $473.1 million in assets.[67] The same survey indicated that the associations' memberships drew heavily from the working class; among the associations that reported their members' occupations, over 59 percent of members were "laborers and factory workers," "housewives and housekeepers," or "artisans and mechanics."[68]

While the serial, permanent, and terminating plans continued to dominate, a new form of organization emerged during this period. The Dayton plan, first used in Dayton, Ohio, in the early or mid-1880s, permitted some members to participate only as savers without borrowing, somewhat reducing the centrality of mutual self-help in those institutions.[69] In addition, it allowed borrowers to determine their own payment amounts, with higher payments reducing their total interest, a feature that partially anticipated the structure of a typical modern mortgage allowing early prepayment without penalty.

The National Associations: A Cul-de-Sac

Beginning in the mid-1880s, a class of national building and loan associations emerged. Unlike the local associations, the national associations operated across city and state lines by opening branches. The term "national" referred to the nonlocal scale of the associations rather than any federal-level regulation or charter. (The term was somewhat of a misnomer since the associations could not operate on a truly nationwide basis; some large states adopted laws effectively barring "foreign"—that is, out-of-state—associations from doing business within their borders by requiring them to put up prohibitively high bonds with the state.[70]) From their starting point of two institutions in Minneapolis, Minnesota, the national associations had grown by 1893 to some 240 national associations, with at least one established in every state.[71]

According to economic theory, national associations could have brought about more efficient allocation of capital compared to local associations, all other things equal: their larger geographic scope meant they could receive deposits (sell shares) in markets where loanable funds were abundant and make home loans in markets with high

[67] Bodfish (1931), pp. 134-36.

[68] Mason (2004), p. 29.

[69] Snowden (1997), p. 233; Mason (2004), p. 20; Haveman and Rao (1997), pp. 1617-19. Bodfish (1931) puts the introduction of the Dayton plan by the Mutual Home and Savings Association of Dayton, Ohio, at 1880, while Mason (2004) puts it at the mid-1880s.

[70] Bodfish (1931), p. 113; Haveman and Rao (1997), p. 1639.

[71] Bodfish (1931), p. 104.

demand. The national associations cited this advantage in one of their publications in 1889, stating that they were "selling stock in vicinities where money is plenty and loaning it where money is scarce, which locals cannot depend upon doing." In addition, the national associations contended, they were able to "supply loans of larger dimensions than the local societies could fill" and "supply money to towns and villages which are not large enough to support a local association."[72] Their larger scope also brought benefits of greater diversification in their loan portfolios as well as efficiencies of scale.

The financial structure of the national associations had roots in the permanent-plan form of the local associations. But there were significant differences between the two. Where all of a member's payments into a local building and loan went into paying down his or her shares, payments into a national association went in part to an "expense fund" that served to boost the organizers' profits. The portion allocated to the expense fund varied from one association to another; a range of 5 percent to 7 percent appears to have been common.[73] Local associations did, of course, spend a portion of their funds on operating expenses, but the amounts involved were much lower at 1 percent to 2 percent of revenues.[74] Moreover, if a member of a national association failed to keep up his payments, he would forfeit the payments he had already made even if he had not yet taken a loan.[75] (Additionally, as with any mortgage, those who had taken a loan were subject to foreclosure of their houses.) Countervailing these disadvantages, from the point of view of prospective members, were the high rates of return that the national associations advertised: the dividend yields they promised were several times those available from banks, local associations, or government bonds.[76]

The local associations responded to the new entrants in part by forming statewide trade groups that fought the nationals through public education—that is, vituperative criticism—and restrictive legislation. (In some states, trade groups for local building and loan associations were already in place before the emergence of the nationals.[77]) These organizing efforts within the industry culminated in 1893 in the formation of a nationwide body of the state trade groups, the U.S. League of Local Building and Loan Associations; its first

[72] Bodfish (1931), p. 106.
[73] Bodfish (1931), pp. 109, 111.
[74] Mason (2004), p. 33.
[75] Bodfish (1931), p. 101; Haveman and Rao (1997), p. 1639.
[76] Mason (2004), p. 33; Bodfish (1931), pp. 102–03.
[77] Mason (2004), p. 38.

convention took place that year in Chicago in conjunction with the World's Columbian Exposition.[78] In addition to opposing the national associations, the state groups and their national body were concerned with promoting homeownership and the local associations.[79]

In their criticisms of the new entrants, the groups representing the local associations held that the nationals were cooperatives in theory but proprietary for-profits in fact. A U.S. League publication argued, "The only object in organizing or carrying on the [national] association is to create and gobble up this expense fund. Their name should be changed."[80] Seymour Dexter, founder and first president of the U.S. League, told the league's second convention in 1894, "Whenever so fine a field of operations presents itself to the scheming and dishonest as the present system of the National Building and Loan Association, we may rest assured that the scheming and dishonest will enter it and pluck their victims until restrained by proper legal restrictions."[81]

Whatever the share of national associations with "scheming and dishonest" organizers, a weakness of their business model was the difficulty of monitoring—of assessing properties and real estate market conditions in branch areas. This difficulty reflected the informational disadvantage of a centralized lending operation; the information technology that would eventually help lenders overcome the disadvantages of distance in home mortgage lending was, of course, not yet in place. Consequently, in contrast with the local associations and their locally based operations, national associations ran a higher risk of lending on the basis of inflated appraisals or lending to poor-quality borrowers.[82]

The downfall of the national associations was put in motion by a major real estate downturn associated with the Depression of 1893. In the first few years of the downturn, the assets of the nationals actually grew as they were perceived as a low-risk investment, but they would come to be hard hit.[83] While mortgage lenders in general suffered, national building and loans were particularly vulnerable on account of the lower average quality of their loans. In addition, as economic conditions reduced the number of new members, the national associations lost a source of new expense-fund contributions and other fees, which some institutions relied on to meet their obligations.[84] The knockout blow for the national associations was the failure in 1897 of the largest of

[78] Bodfish (1931), pp. 140–43.
[79] Mason (2004), p. 38.
[80] Bodfish (1931), p. 108.
[81] Bodfish (1931), p. 107.
[82] Mason (2004), p. 34.
[83] Mason (2012), p. 386–87.
[84] Haveman and Rao (1997), pp. 1639–40; Mason (2004), p. 36.

them, the Southern Building and Loan Association of Knoxville, Tennessee, which gravely damaged confidence in the remaining nationals; virtually all of those institutions ceased operation within a few years.[85]

Final Wave of Growth in the 1920s and Demise

During and after the collapse of the national building and loan associations, some in the local building and loan movement expressed concern that the record of the nationals would leave a long-term stigma on the local associations. An article in the official newsletter of the Building Association League of Illinois and Missouri, for example, noted in 1896 that in many "smaller cities and towns" hundreds of savers had trusted their money to a national association only to lose it all. "It will be years," the newsletter held, "before it will be possible to establish a genuine building and loan association in such a community, after the name of building association has been besmirched and prostituted, and brought into grave disrepute through the actions of the schemers who have run these bogus concerns."[86]

Although the membership and assets of local building and loans did remain essentially flat during the first few years of the 1900s, perhaps as a result of the stigma left by the failed national associations, they resumed their growth afterward: from about 1.5 million members and $571 million in assets in 1900 to about 2.2 million members and $932 million in assets in 1910. Even more rapid growth was still to come: by 1920, membership had more than doubled to nearly 5 million and assets had grown more than 2 1/2-fold to $2.5 billion. (The number of associations also rose, but less dramatically, reflecting an increase in the average institution size: from 5,356 in 1900 to 5,869 in 1910 and 8,633 in 1920.) In 1930, despite the financial crisis the preceding year, membership was up to 12.3 million and assets totaled $8.8 billion.[87]

Several developments aided the growth of the local associations and of their model of affordable mortgage lending during this period. One is that the locals became more promotion-minded and more sophisticated about promotion. While hard data on their promotional efforts are scarce, it appears that the locals during this time increasingly supplemented their primary means of acquiring new members—word of mouth—with the use of newspaper advertisements and window

[85] Mason (2004), p. 37; Bodfish (1931), pp. 114–15.
[86] "Downfall of the 'Nationals.'" (1896).
[87] Bodfish (1931), p. 136 (table 1).

displays.[88] This shift appears to have been partly the result of encouragement and guidance from the U.S. League[89] but is also consistent with the increasing scale of the local associations, which could better support such efforts.

Another development that boosted local associations during this time was the real estate boom in California and other western states, together with the embrace of building and loan associations there as a form of affordable housing finance. The assets of building and loans in the West grew from 1920 to 1930 at an average annual rate of 47.1 percent, compared with 25.1 percent for the nation as a whole.[90]

Additionally, the 1920s saw a trend of developers and builders establishing, in effect, captive associations that they dominated to support the sale of their houses. While developers, builders, and brokers had long been involved in local building and loan associations, there is evidence that they went further during this period in coopting the building and loan model, possibly boosting the numbers of building and loans.[91]

Recessions were frequent during this period, even before the Great Depression—eight recessions occurred from 1900 to 1928, an average of one every three and a half years[92]—but these did not appear to interfere with the growth of building and loans. In general, building and loans tended to be more stable than banks during periods of market stress, such as the panic of 1907, because their savers were member-owners rather than creditors. While bank depositors could, by definition, demand the immediate return of demand deposits, not all building and loan plans allowed for withdrawal before a prescribed maturity date, and under those plans that did, the association had a significant period (commonly thirty or sixty days) to carry out a member's withdrawal request.[93] Thus, building and loans were not exposed to the extent

[88] Mason (2004), p. 46.

[89] Mason (2004), pp. 40, 47.

[90] Mason (2012), p. 388 (table 2).

[91] Snowden (2010), p. 9.

[92] National Bureau of Economic Affairs, "U.S. Business Cycle Expansions and Contractions." n.d.

[93] Mason (2004), p. 53; Mason (2012), p. 390; Rose (2014), p. 250. The withdrawal process is accurately represented in the 1946 film *It's a Wonderful Life*, which involved the fictional Bailey Bros. Building and Loan.

TOM: I got two hundred and forty-two dollars in here, and two hundred and forty-two dollars isn't going to break anybody.

GEORGE (handing him a slip): Okay, Tom. All right. Here you are. You sign this. You'll get your money in sixty days.

TOM: Sixty days?

GEORGE: Well, now that's what you agreed to when you bought your shares.

that banks were to a risky mismatch between long-term assets and short-term liabilities.[94]

Following the crash of 1929 and the ensuing Great Depression, a large number of building and loans did close; the number of associations dropped from 12,342 in 1929 to 8,006 a decade later.[95] These closures did not result from depositor runs but from the effects of the Depression on the banking sector: as many building and loans required short-term lending from banks (given that their assets were mainly longer-term mortgages), the widespread extent of bank failures led to a short-term credit crunch for the associations. In addition, in the early years of the Depression, building and loan failures were concentrated in Pennsylvania, where building and loans members taking out second mortgages under the Philadelphia plan were unable to roll over their short-term first mortgages (made by a bank or another conventional lender) as the mechanism of the Philadelphia plan assumed.[96] It is reasonable to assume, also, that the sharp drop in nominal real estate prices[97] contributed to building and loan closures. During the roughly one hundred years in which local building and loans thrived, however, they played a significant role in extending homeownership through affordable mortgage lending.

3. INSURANCE COMPANIES AS MORTGAGE LENDERS

Beyond the business of creating MBS from mortgages and guaranteeing these MBS, insurance companies were important providers of mortgage loans themselves, making mortgages and holding them as investments. Insurance companies accounted for about 7 percent of all mortgages outstanding as of the early 1890s, meaning about $400 million of $6 billion outstanding mortgages at that time.[98] And while, as noted earlier, regulatory prohibitions had limited the ability of some of the largest insurance companies (those located in the state of New York) until the mid-1880s, by the 1920s the major insurance companies were

[94] An 1869 tract exhorting working-class Americans to participate in building and loans cited freedom from the risk of runs as an advantage of the associations over depository institutions. Wrigley (1869), pp. 4-5, 47.

[95] Mason (2012), p. 390; Bodfish (1931), p. 136.

[96] Mason (2012), pp. 390-91.

[97] An index of single-family house prices for twenty-two U.S. cities indicates an average drop of 30.4 percent from the pre-Depression peak in 1925 to 1933. Grebler, Blank, and Winnick (1956), p. 347. An index for Manhattan that includes multifamily dwellings finds a more pronounced 67 percent drop from the third quarter of 1929 to late 1932. Nicholas and Scherbina (2013).

[98] Snowden (1995a), p. 220.

important nationwide mortgage lenders.[99] Indeed, by 1929 insurance companies had grown in importance as mortgage lenders, so that they were holding 16 percent of the $47 billion in all types of mortgage debt then outstanding and were responsible for about 15 percent of all residential mortgages outstanding in 1930.[100]

Because life insurance company liabilities—death benefit payments on life policies and payments on annuity contracts—tend to be long-term and predictable, it is natural that life companies would also tend to hold long-term assets.[101] Banks, with a high percentage of their funding coming from short-term deposits, wish to limit their exposure to long-term assets for fear that depositors would suddenly demand repayment of deposits, causing a run on the bank. Additionally, given the mismatch in their short-term liabilities and long-term assets, banks face interest rate risk.[102] Life insurance companies face less of both of these risks, so they tend to have an advantage over banks in holding extremely long-term assets such as mortgages, and they will, as a result, find investments in mortgages to be attractive.[103]

---

[99] Saulnier (1950), p. 39. Snowden (1995a), pp. 230–42 provides a thorough and fascinating discussion of the means by which life insurance companies handled the delegate-monitoring problem that they faced when lending to mortgagors distant from insurance company headquarters. Much of his discussion focuses on life insurance farm mortgage lending, but it covers residential lending as well.

[100] Saulnier (1950), p. 2. Saulnier (1950), p. 4, notes that as of 1938, the earliest year for which he provides a breakdown by type of mortgage borrower, insurance companies were more important lenders in the commercial mortgage market than in the home-mortgage market. In the home-mortgage market (1-4 family), they held about 8 percent of all outstanding home mortgages ($17.1 billion), while in commercial mortgages they held 39 percent of the total amount outstanding. Snowden (1995a), p. 242, reports that insurance companies held $4.4 billion in residential mortgages in 1930. Grebler, Blank, and Winnick (1956), p. 447, report that total nonfarm residential mortgage debt outstanding in 1930 was $30.2 billion. The percentage may be slightly overstated by the extent to which Snowden's figure includes home (residential) mortgages located on farms.

[101] As of June 30, 2017, reserves for future life insurance payments and annuity reserves accounted for 73 percent of all life insurance company liabilities (Board of Governors 2017). We could find no data for insurance company liabilities from the nineteenth century. Saulnier (1950) provides some nineteenth-century data on insurance company assets.

[102] For banks, interest rate risk poses a danger that a shift in market interest rates will reduce their earnings or even produce insolvency. Banks' heavy reliance on short-term deposits means that their funding tends to reprice quickly in response to shifts in market interest rates; if banks do not quickly respond to market rate movements, their depositors will withdraw their funds and move them elsewhere. As a result, banks must limit the maturities of their assets so that interest rates on these will also move with market rates. If banks fail to do so and market rates increase, the interest rates they earn on their assets will increase less than their interest cost for deposits, and they will suffer losses.

[103] Paulson et al. (2012) discuss and measure the liquidity of life insurance company liabilities and conclude, on page 2, that, "Overall, life insurers have less liquid liabilities than banks do. ...While life insurers have some demand deposit-like products, many of their products have limitations on withdrawals." Still, some modern large life insurers offer a wide range of financial products, a portion of which are quite liquid. For example,

Single-family home mortgages were an important part of life insurance company mortgage lending but were not the only mortgage lending that they did. According to a survey of twenty-four of the largest life insurance companies, responsible for 65 percent of urban mortgage loans of life insurance companies (covering the years 1920 through 1946), as of the early 1920s, single-family home loans accounted for 78 percent of all urban mortgages made by life companies in numbers of loans and 31 percent in terms of dollars. Mortgages on apartments, stores, "other income properties," 2-4 family homes, and "1-4 family dwellings with a business use" accounted for the remainder of life insurance company mortgages.[104]

Contrary to the typical bank loan of the period, which was non-amortizing and was paid off in full at maturity, in the early 1920s 83 percent of the mortgages made by sampled life insurance companies were fully (24 percent) or partially (59 percent) amortizing. Also, they tended to be longer-term than their bank equivalents, with 60 percent having five- to nine-year contract maturities and 27 percent with ten- to fourteen-year maturities. Insurance companies likely were more willing to extend longer-term mortgages than banks because of the long-term nature of insurance company liabilities.[105]

Following the creation of the Federal Housing Administration mortgage loan guarantee program in 1934, during the Great Depression, life insurance companies moved heavily into making FHA-insured and later VA-insured mortgages.[106] At the same time, these companies began offering fifteen- to twenty-year maturities and twenty-plus-year mortgages.[107]

4. HOW THESE INSTITUTIONS IMPROVED MORTGAGE TERMS

Diversification

As of the 1890s, in terms of dollars, 70 percent of all U.S. mortgage loans were made by individual investors.[108] Therefore, financial institutions, such as commercial banks, mortgage companies, savings and loans, building and loans, and insurance companies accounted for only

---

Paulson et al. note that, as of 2011, 11.1 percent of life insurance company liabilities have "high liquidity," liabilities with few limits on early withdrawal, according to their estimates (Table 3, p. 3).

[104] Saulnier (1950), p. 42.

[105] Saulnier (1950), pp. 44–45.

[106] U.S. Department of Housing and Urban Development (2017)

[107] Saulnier (1950), p. 45.

[108] Snowden (1995a), p. 220.

30 percent of mortgage holdings. Such a market seems to have been ripe for financial innovations that could allow greater diversification in lending (assuming that individual mortgage investors are unlikely to be well-diversified), access to a wider pool of funds, and therefore offer more affordable mortgages.

One can think in terms of two types of diversification: first, intraregional diversification (the type of diversification allowed by a local lending institution, such as building and loan associations); and second, interregional diversification (the type of diversification allowed by late nineteenth-century MBS, representing mortgages made to borrowers in the western portion of the United States but sold to investors in the East and internationally). By reducing the risk borne by lenders, both types of diversification can reduce mortgage interest rates paid by borrowers.

Many of the individual mortgage investors of the late 1800s were "professional operators or real estate attorneys," or one-time lenders, such as individual home sellers providing purchase money for the buyer of their home, family members, or occasional investors.[109] Individual home sellers and occasional investors, unless extremely wealthy, and therefore able to make numerous mortgage loans, were likely to have an undiversified mortgage portfolio so that the default of one borrower would cause large proportional losses to their portfolio of mortgage investments. In consequence, such lenders will tend to charge high interest rates to compensate themselves for the substantial credit risk they face.

A lending institution, such as a local bank (or a building and loan association) could gather funds from a large pool of local savers, invest in numerous loans, and diversify away some of the credit risk. This intraregional diversification advantage was likely the genesis for the development of many local lending institutions.

According to estimates made at the time, as of the early 1890s, only 24 percent of individual investor funds came from out of state (meaning interregionally); the remainder came from investors in the same state as the borrower, implying that the providers of most mortgage funding in the United States—individuals—were subject to huge losses from local shocks that lenders with a more geographically diversified pool could avoid.[110] In 1890, 42 percent of the population lived on farms so that many of these individual investors were likely making mortgage loans to farmers, and likely, due to the size of individual investors'

[109] Grebler, Blank, and Winnick (1956), chapter 13, pp. 190–91.
[110] Frederiksen (1894), p. 209.

portfolios, loans concentrated in one or a few localities.[111] Therefore, these investors' incomes would be subject to large proportional losses if the region were struck with adverse weather—such as an unexpected freeze affecting citrus growers in Florida. Even in nonfarming areas, such as industrial areas, the failure of an important manufacturer would likely lead to trouble for many mortgage borrowers who were employed at a local factory, so that an individual investor's income would be heavily influenced by such shocks. In contrast, a regionally or nationally diversified lender would be better protected.

Similarly, if individual mortgage lenders, meaning individuals with large savings, tended to focus on lending near their home, and such savers were not evenly distributed around the nation, then interest rates could vary considerably from region to region. Regions in which savers were concentrated would have low interest rates and better terms—such as longer-term loans and lower down payments—as these concentrated savers competed among one another to lend to the available borrowers. This seems to have been the case in the northeast portion of the United States in the late nineteenth century, for example.[112] Further, intermediaries, agents who bought and sold home mortgage loans, apparently, tended to purchase local loans and sell them to local investors.[113] In contrast, interest rates would be higher in regions with few savers, which implies that more homes would be built (as well as farms established and mortgaged-buildings built) in areas in high-savings regions, say the Northeast, even though there might be greater demand for homes in other areas of the country.[114] As a result, even though there might be many more new households forming in other regions of the country, and ideally more homes built in other regions, instead households that were completely creditworthy would be unable to afford homes because of high interest rates in those areas where pools of savings were smaller.

The question of whether rates were too high in western and southern states compared to northeastern states has been investigated. Here, *too high* means borrowers who were equivalently creditworthy received

---

[111] U.S. Census Bureau (1975), pp. K1-16.

[112] See Frederiksen (1894), p. 206; Davis (1965), p. 375. Davis (1965), p. 370, notes that before 1890, national banks in the West sold CDs to investors in the East, but that this activity was criticized by Office of the Comptroller of the Currency examiners.

[113] Frederiksen (1894), p. 221.

[114] Frederiksen (1894), p. 209, notes the inefficiency created by the inability of mortgage funding to flow to its most valuable uses. "So that in America the making of a mortgage loan is essentially a local transaction. ... Under an ideal system of mortgage banking, the capital available for permanent investment would be distributed where most needed. ... In one part of the country the rate of interest paid on a mortgage loan is with equal security twice as high as another."

higher interest rate loans in southern and western states compared to northeastern borrowers. Davis (1965) argues that while a national market for mortgages had developed throughout most of the country by 1900, in the South especially and to some extent in the West, mortgage rates remained unusually high as late as 1900.[115] Eichengreen (1984) analyses the riskiness of farm mortgages and concludes that rate differentials between the East, on the one hand, and the West and South, on the other, can mostly be explained by foreclosure risk differences. Snowden (1987) disagrees with Eichengreen. Snowden analyzes interest rates paid on both farm and home mortgages in 1890 and found that differences could not be explained by default (foreclosure) risk differences and instead that there were remaining regional differences even after accounting for risk.

One reason that investors with available savings tended to lend locally, and especially so when long-range communication and travel was difficult and slow, as in the late 1800s, was that monitoring borrowers was—and still is—costly. Such monitoring could include multiple activities, such as gathering knowledge of local business conditions in order to forecast future land values and ensuring that the borrower maintained the mortgaged property's condition—thereby protecting the lender's collateral interest. But monitoring was less costly for local lenders, given that they could more easily check on the borrower and the collateral (home or farm) and perhaps knew the borrower personally.[116] When lending at a distance, monitoring collateral preservation meant slowly, and perhaps dangerously, traveling to distant areas to check on collateral or alternatively hiring others (so-called delegated monitoring) to perform this monitoring and then ensuring that these monitors were diligent. Costly investment mistakes arose for distant lenders when delegated monitors were careless or dishonest during the nineteenth century just as they did during the twenty-first century subprime crisis.[117]

Similarly, if the borrower were in default, the lender would be especially eager to keep an eye on the collateral—as in this situation the borrower may be less able (because of a lack of funds) or interested in preserving the property (because she knows that it is unlikely she will be living at the property for very long). The ability to inexpensively visit and check on a property on a very frequent basis would be especially valuable in such situations.

---

[115] Davis (1965), pp. 388–93.

[116] Snowden (1995a), p. 221.

[117] Snowden (1995a), pp. 221–30, discusses nineteenth-century arrangements by which distant lenders established and contracted with delegated monitors.

While monitoring costs tended to concentrate mortgage lending in regions with high savings, laws also restricted nationwide lending, which produced efficiency losses for this reason, as discussed earlier. And such laws were, in some cases, quite stringent. For example, in New York, building and loan associations were prohibited from making mortgage loans on properties that were more than fifty miles from the association's headquarters.[118] Similarly, New York law prohibited insurance companies—the third-largest provider of intermediated mortgages in the early 1890s—from lending outside of the state of New York until 1886, and New York-headquartered insurance companies held a significant share of insurance assets.[119] A number of other states also prohibited interstate mortgage lending by insurance companies headquartered in their states, though some large insurance companies in Connecticut and Wisconsin enjoyed interstate lending powers.[120] New York insurance companies saw the earnings that they were missing due to these restrictions and had lobbied aggressively in the 1870s and 1880s to have the restrictions removed.[121] New York-headquartered insurance companies were aware that interest rates were unusually high in the West and South and had observed insurance companies headquartered in Connecticut successfully provide distant mortgages. Still, even when legislated restrictions were removed, the costs of lending in distant markets were often found to be prohibitive.[122]

Liability Structure

While lending to borrowers who are unknown to the lender—as are distant borrowers—was costly for insurance companies and other types of mortgage lenders, insurers as well as MBS-issuing companies and building and loan associations likely had an advantage in mortgage lending not available to banks and other deposit-taking institutions. Banks and similar institutions typically fund their loans to a significant degree with short-term deposits. Some of these deposits can even be withdrawn on demand (known as demand deposits). Interest rates on short-term deposits must track market interest rates when rates increase or the bank's customers are likely to withdraw their deposits—when they mature, or immediately in the case of demand deposits—and take them somewhere that offers higher interest rates.

[118] Herrick and Ingalls (1915), p. 21.
[119] See Snowden (1995a), pp. 220, 235.
[120] Snowden (1995a), p. 230.
[121] Snowden (1995a), p. 235.
[122] Snowden (1995a), pp. 218–19 and 235.

The short-term nature of the liabilities of depositories (banks and others) means that these institutions face two risks: (1) the danger of runs—when depositors become frightened of the health of the bank or of the broader economy and suddenly, in force, show up at the bank and want to withdraw their deposits; and (2) interest rate risk.

To address these risks, depositories must take costly measures. To counter the risk of runs, depositories hold large amounts of low interest rate liquid assets (or noninterest earning assets, in the case of cash in the depository's vault), which can be sold quickly, without loss of value, to meet depositor demands during a run. As a result, preparing for runs limits depository institution interest earnings to a degree. To address interest rate risk, depositories tend to limit the maturity of their assets, sacrificing some expected return, to be closer to the maturity of the liabilities. The prohibition on national bank mortgage lending may have been, in part, an effort to control interest rate risk, given that mortgage loans tend to be fairly long term (maturities of at least several years).

Building and loan associations, MBS issuers, and insurance companies were in a better position to offer long-term mortgages than depositories. Building and loans and MBS issuers did not have to hold large amounts of low-earning liquid assets in order to meet runs, given that customers did not have the ability to withdraw on demand; building and loans were largely funded with shares that were quite long-term, and the bonds that funded MBS issuers had long maturities. Similarly, building and loans and MBS issuers faced little interest rate risk given that both their assets (mortgages) and their liabilities (shares and bonds) had similar long-term maturities. While members of many building and loan associations could redeem their shares with thirty days' or sixty days' notice, that notice period made those redemptions quite different in their effect from on-demand withdrawals, given that the associations could obtain the cash to meet a withdrawal in an almost leisurely manner. And the liabilities of insurance companies (and especially of life insurance companies) are fairly predictable streams of payments on insurance policies and annuity contracts, so they are not subject to runs and can easily be matched against long-term assets like mortgages.

Therefore, building and loans, MBS issuers, and insurance companies did not face the risks that depositories faced from issuing long-term assets. As a result, these nondepository financial institutions were likely to be able to make mortgages at lower interest rates than were depositories. On the other hand, once federal deposit insurance for banks and for savings and loans was created in 1933 and 1934, respectively, the risk of runs was greatly reduced, taking away some of the advantage enjoyed by nondepositories—which may help to explain the historical decline of these nondepository institutions as sources of mortgage loans relative to depositories.

5. CONCLUSION

The history set out here highlights a variety of mechanisms through which private institutions have participated in providing affordable mortgage lending: resale or securitization of mortgages (mortgage companies and MBS), mutual self-help (building and loan associations), and portfolio lending (insurance companies). Among the economic efficiencies through which they were able to improve mortgage terms were local-level diversification of mortgage portfolios, interregional diversification of portfolios, and better matching of their liabilities with their long-term assets.

The success of these institutions, over periods of some decades, raises a question of whether they may be relevant to contemporary policy debates. Today, as policymakers consider the prospect of winding down the GSEs, there is concern about the extent to which home mort-gage products that are perceived as desirable—particularly long-term, fixed-rate mortgages—would continue to be available in the absence of the GSEs and their implied federal guarantees. In addition, there is concern that without the facilitative role of GSEs in maintaining MBS markets, there may be a major reduction in the extent to which home-mortgage markets have access to funding from capital markets. These, of course, are complex issues that cannot be resolved on the basis of the historical record alone. The history of U.S. mortgage finance does, however, illustrate some of the ways in which private efforts may arise to address demand for affordable mortgages in the absence of public guarantees or subsidies.

REFERENCES

Alger, George W., and Alfred A. Cook. 1934. "Report to His Excellency Herbert H. Lehman, Governor of the State of New York." Albany: New York state government document.

Behrens, Carl F. 1952. *Commercial Bank Activities in Urban Mortgage Financing*. Cambridge, Mass.: National Bureau of Economic Research.

Board of Governors of the Federal Reserve System. 2017. "Financial Accounts of the United States – Z.1. 2017: Q2 Release." Available at: https://www.federalreserve.gov/releases/z1/current/default.htm [September 21].

Bodfish, Henry Morton, ed. 1931. *History of Building and Loan in the United States*. Chicago: U.S. Building and Loan League.

Brewer, H. Peers. 1976. "Eastern Money and Western Mortgages in the 1870s." *Business History Review* 50 (Autumn): 356–80.

Byers, C. Floyd. 1927. "Building and Loan Associations: The Principle Plans of Operation." M.A. thesis, Ohio State University.

Davis, Lance E. 1965. "The Investment Market, 1870-1914: The Evolution of a National Market." *Journal of Economic History* 25 (September): 355–99.

Dexter, Seymour. 1889. "Co-Operative Savings and Loan Associations." *Quarterly Journal of Economics* 3 (April): 315–35.

"Downfall of the 'Nationals.'" 1896. *Financial Review and American Building Association News* 15 (March).

Edson, Charles L. 2011. "Affordable Housing — An Intimate History." In *The Legal Guide to Affordable Housing Development*, edited by Tim Iglesias and Rochelle E. Lento. Chicago: American Bar Association, 3–20.

Eichengreen, Barry. 1984. "Mortgage Interest Rates in the Populist Era." *American Economic Review* 74 (December): 995–1015.

Foulke, Roy A. 1941. *The Sinews of American Commerce*. New York: Dun & Bradstreet.

Frederiksen, D. M. 1894. "Mortgage Banking in America." *Journal of Political Economy* 2 (March): 203–34.

Grebler, Leo, David M. Blank, and Louis Winnick, eds. 1956. *Capital Formation in Residential Real Estate*. Princeton, N.J.: Princeton University Press.

Griffin, Emma. 2013. *Liberty's Dawn: A People's History of the Industrial Revolution.* New Haven, Conn.: Yale University Press.

Hansell, Saul. 1995. "Banking's New Giant: The Deal; Chase and Chemical Agree to Merge in $10 Billion Deal Creating Largest U.S. Bank." *New York Times*, August 29.

Haveman, Heather A., and Hayagreeva Rao. 1997. "Structuring a Theory of Moral Sentiments: Institutional and Organizational Coevolution in the Early Thrift Industry." *American Journal of Sociology* 102 (May): 1606–51.

Herrick, Myron T., and R. Ingalls. 1915. *How to Finance the Farmer: Private Enterprise—Not State Aid*. Cleveland: The Ohio State Committee on Rural Credits and Cooperation.

JPMorgan. 2017. "About Us, Company History." https://www.jpmorgan.com/country/US/EN/company-history [May 11].

Loucks, William N. 1929. *The Philadelphia Plan of Home Financing: A Study of the Second Mortgage Lending of Philadelphia Building and Loan Associations*. Chicago: Institute for Research in Land Economics and Public Utilities.

Mason, David L. 2004. *From Buildings and Loans to Bail-Outs: A History of the American Savings and Loan Industry, 1831-1995*. Cambridge, U.K.: Cambridge University Press.

Mason, David L. 2012. "The Rise and Fall of the Cooperative Spirit: The Evolution of Organizational Structures in American Thrifts, 1831-1939." *Business History* 54 (June): 381–98.

Nicholas, Tom, and Anna Scherbina. 2013. "Real Estate Prices During the Roaring Twenties and the Great Depression." *Real Estate Economics* 41 (Summer): 278–309.

Paulson, Anna, Richard Rosen, Zain Mohey-Deen, and Robert McMenamin. 2012. "How Liquid are U.S. Life Insurance Liabilities?" Federal Reserve Bank of Chicago *Fed Letter* 302 (September).

Price, Seymour J. 1958. *Building Societies: Their Origin and History*, London: Franey & Co.

Rose, Jonathan D. 2014. "The Prolonged Resolution of Troubled Real Estate Lenders During the 1930s." In *Housing and Mortgage Markets in Historical Perspective,* edited by Eugene N. White, et al. Chicago: University of Chicago Press, 245–84.

Saulnier, R. J., ed. 1950. *Urban Mortgage Lending by Life Insurance Companies*. Cambridge, Mass.: National Bureau of Economic Research.

Snowden, Kenneth A. 1987. "Mortgage Rates and American Capital Market Development in the Late Nineteenth Century." *Journal of Economic History* 47 (September): 671–91.

Snowden, Kenneth A. 1995a. "The Evolution of Interregional Mortgage Lending Channels, 1870-1940: The Life Insurance-Mortgage Company Connection." In *Coordination and Information: Historical Perspectives on the Organization of Enterprise*, edited by Naomi R. Lamoreaux and Daniel M.G. Raff. Chicago: University of Chicago Press, 209–56.

Snowden, Kenneth A. 1995b. "Mortgage Securitization in the United States: Twentieth Century Developments in Historical Perspective." In *Anglo-American Financial Systems: Institutions and Markets in the Twentieth Century*, edited by Michael D. Bordo and Richard Sylla. New York: Irwin, 261–98.

Snowden, Kenneth A. 1997. "Building and Loan Associations in the U.S., 1880-1893: The Origins of Localization in the Residential Real Estate Market." *Research in Economics* 51 (September): 227–50.

Snowden, Kenneth A. 2010. "The Anatomy of a Residential Mortgage Crisis: A Look Back to the 1930s." Working Paper 16244. Cambridge, Mass.: National Bureau of Economic Research. (July).

Tcherkinsky, M. 1922. *The Landschaften and Their Mortgage Credit Operations in Germany, 1770-1920*. Rome: Printing Office of the International Institute of Agriculture, Bureau of Economic and Social Intelligence.

U.S. Bureau of the Census. 1975. *Historical Statistics of the United States: Colonial Times to 1970*. Washington, D.C.: U.S. Department of Commerce.

U.S. Department of Housing and Urban Development. 2017. "The Federal Housing Administration (FHA)." https://portal.hud.gov/hudportal/HUD?src=/program_offices/housing/fhahistory [May 11].

Wandschneider, Kirsten. 2014. "Lending to Lemons: Landschaft Credit in Eighteenth-Century Prussia." In *Housing and Mortgage Markets in Historical Perspective*, edited by Eugene N. White, et al. Chicago: University of Chicago Press, 305–25.

Wandschneider, Kirsten. 2015. "Landschaften as Credit Purveyors — The Example of East Prussia." *Journal of Economic History* 75 (September): 791–818.

White, Eugene N. 2014. "Lessons from the Great American Real Estate Boom and Bust of the 1920s." In *Housing and Mortgage Markets in Historical Perspective*, edited by Eugene N. White, et al. Chicago: University of Chicago Press, 115–58.

Wrigley, Edmund. 1869. *The Working Man's Way to Wealth; A Practical Treatise on Building Associations: What They Are and How to Use Them.* Philadelphia: James K. Simon.