Economic Quarterly—Volume 96, Number 1—First Quarter 2010—Pages 1–9

Introduction to the Special Issue on the Diamond-Dybvig Model

Edward Simpson Prescott

This special issue of the *Economic Quarterly* is dedicated to the 1983 model of bank runs developed by Douglas Diamond and Philip Dybvig.[1] Their model has been a workhorse of banking research over the last 25 years and during the recent financial crisis it has been one that researchers and policymakers consistently turn to when interpreting financial market phenomena.

The Diamond-Dybvig model has three basic elements:

- Long-term investments that are more productive than short-term investments;
- A random need for liquidity on the part of an individual; and
- Private information about an individual's need for liquidity.

With these elements, Diamond and Dybvig (DD hereafter) show that it is desirable for people to pool their funds and jointly invest in productive long-term investments, while allowing individuals to withdraw their funds on demand, even before the end of the life of the long-term investments. Furthermore, they show that it is also desirable to set payouts for early withdrawals high enough so that if every person in the pool withdrew his funds early, there would not be enough funds available to meet every withdrawal.

DD interpreted this arrangement as a bank because it contains two important characteristics that are typically identified with banks. First, it performs

The views expressed do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] The paper appeared in the *Journal of Political Economy*. The full citation is in the references as Diamond and Dybvig (1983). A freely available reprint is Diamond and Dybvig (2000). For a simple exposition of this model, see Diamond's 2007 *EQ* article.

maturity transformation, that is, it backs short-term liabilities with long-term illiquid assets. Second, it issues liabilities that are payable on demand, that is, bank deposits.[2]

According to DD, while this arrangement is effective at increasing output and providing liquidity insurance, it is also susceptible to a bank run. In their environment, there is a coordination problem among depositors. If too many people withdraw early, then the long-term investments are liquidated early causing a loss in output. DD show that there is such an equilibrium in that depositors who do not need early liquidity will still withdraw early because they think that other depositors without an early liquidity need are going to withdraw early. This inefficient allocation is an equilibrium (as is the efficient allocation) even if the bank is solvent.

Diamond and Dybvig also discuss several mechanisms for eliminating the run equilibrium. These include deposit insurance, suspension mechanisms, and central bank lending. All of these mechanisms have been used to various degrees over time. The United States has had federal government-provided deposit insurance since 1933. The precursors to central banks, the clearing-houses, often would suspend payments during a financial crisis (Timberlake 1984). Finally, the lender-of-last-resort justification of central bank lending has been used heavily in this crisis and it was heavily used historically. For example, Bagehot (1873), when writing about the Bank of England, gave his famous dictum that to prevent a financial panic, a central bank should freely lend at a penalty rate on good collateral.

1. DIAMOND-DYBVIG AND THE RECENT FINANCIAL CRISIS

Until recently, bank runs were not considered a major problem in the United States. The introduction of deposit insurance in the 1930s was considered to have essentially solved this problem. There had been very few bank runs since then.[3] Much of the academic literature instead focused on the sizeable costs of moral hazard that can come with a deposit insurance system, as was seen in the savings and loan crisis of the 1980s (see, for example, White [1991]).

What the academic and policy worlds missed was just how much some of the newer (since the 1970s) financial arrangements were starting to resemble banks in that they performed maturity transformation and financed assets with

[2] The other characteristics typically identified with banks are delegated monitoring and payment services. For a theory of the former, see Diamond (1984). As for the latter, there is an extensive literature on payments and monetary economics, a portion of which uses models closely related to DD. In this issue, Cavalcanti and Jack, Suri, and Townsend discuss this literature.

[3] The bank runs that did happen tended to be isolated and on a small scale. For example, in 2005, there was a run on Abacus, a small bank in Chinatown, New York (Campbell 2005). In the 1980s, there were runs on some savings and loans, but these operated under state-sponsored insurance schemes (Todd 1994).

liabilities that resembled demand deposits. Many of these arrangements ran into trouble during the financial crisis when they could not roll over their short-term debt. Whether these episodes match the DD equilibrium in which a solvent bank is run because of a panic is still a topic of debate. After all, a run on a bank is also perfectly consistent with a bank being insolvent.

What we do have now, however, are data that are much higher quality than are available on historical runs.[4] Furthermore, as we will see, these financial arrangements differ along dimensions such as how excess short-term withdrawals are managed. My conjecture is that these sources of variation along with the data will provide an important source of information for not only evaluating the DD model, but also evaluating methods for dealing with a potential run.

Bank Runs

In the recent crisis, there were several runs on traditional banks. In the United Kingdom, Northern Rock bank was unable to roll over its wholesale funding in the fall of 2007, and that led to large withdrawals by retail depositors who, at that time, were not protected by deposit insurance.[5] In the United States, there were large withdrawals from IndyMac, a bank that specialized in alt-A mortgages, many of which were made in California (Office of the Inspector General 2009). Washington Mutual experienced large withdrawals in July 2008 and shortly after the failure of Lehman Brothers in September 2008 (Grind 2009).

Auction Rate Securities

Auction rate securities (ARS) are long-term debt securities that are transformed into short-term securities through regular periodic auctions.[6] The auctions set the short-term interest rate and allow for the transfer of ownership. If a holder wants to sell the bond, he places a sell order, and if there are enough bids in the auction, he sells his security. If there are not enough bids, then he keeps the security and the issuer of the bond pays a predetermined rate in the contract, often one that is relatively high. ARS are issued by municipalities, student loan pools, and closed-end mutual funds.

At first glance, ARS look like any other security with varying liquidity. They were, however, marketed and treated as cash-like securities.

---

[4] There was a debate about whether the runs in the 1930s were due to a DD-like bank run equilibrium occurring for a solvent bank or whether they occurred because the bank was insolvent. See Calomiris and Mason (1997).

[5] For a description of the Northern Rock run, see Shin (2009).

[6] Information in this section is from Han and Li (2008).

Furthermore, if there were not enough bids to clear an auction, the sponsoring entity, which was either a large bank or investment bank, would often provide enough bids to clear the market.[7] However, in the spring of 2008, the sponsoring banks started pulling their support. This contributed to a sizeable demand by investors to pull out of the market, and there was a large increase in the number of auction fails. Han and Li (2008) interpret this event as a run.

Special Purpose Vehicles

Another group of bank-like entities that developed are trusts that hold securities and are financed by a mix of short- and long-term debt (along with a small amount of equity occasionally). These trusts, set up by banks and investment banks, are also known as structured investment vehicles and collateralized debt obligations. Many of these trusts hold long-term securities, such as mortgage-backed securities, and finance part of their investment with commercial paper, which is a short-term, cash-like liability. The commercial paper issued by these trusts is similar to bank deposits in that a lender who chooses to roll over the commercial paper is analogous to a depositor who withdraws his deposit from a bank.

Covitz, Liang, and Suarez (2009) use daily data from August 2007 to December 2007 on the ability of these vehicles to roll over their commercial paper. They found that specific features of the programs, such as the existence of liquidity support, affected the ability to roll over commercial paper. They also found difficulties in rolling over debt that are not explained by these differences and conclude that this is evidence of a bank-like run caused by a panic.

Repo Markets

Repo transactions are short-term agreements to sell and repurchase securities. They are essentially short-term collateralized loans. The loans are often made by wholesale institutions such as money market funds, corporations, hedge funds, and other entities that have a lot of cash to invest. Since their cash holdings are too large to benefit from deposit insurance, they instead make these collateralized loans.

The broker-dealer investment banks (e.g., Bear Stearns, Lehman Brothers, Merrill Lynch, Morgan Stanley, and Goldman Sachs) partially financed their investments with these repo transactions. They would invest in long-term

[7] Tender option bonds and variable rate demand obligations are similar to auction rate securities in that they are fundamentally long-term bonds that have a short-term interest rate determined through an auction mechanism. Unlike owners of ARS, owners of these securities have the option of putting the security back to the originator or marketer.

assets, often through securities, and partially finance the investment with the cash lent as part of the repo transactions. Gorton and Metrick (2009) argue that these repo transactions looked a lot like demand deposits. The lender could withdraw all his funds by not rolling over the repo or even partially withdraw his funds by requiring a large haircut on the valuation of the collateral. Gorton and Metrick also argue that there was a wide-scale panic in these markets as investors began to doubt the quality of collateral and shifted their funds to safer forms such as Treasury securities. Partly because of this movement, the five large investment banks either failed or converted into banks.

Money Market Mutual Funds and Other Investment Pools

Money market mutual funds (MMMFs) are investment pools that invest in short-term liquid assets such as Treasury securities, commercial paper, repos, and certificates of deposit. Unlike other mutual funds, however, they use an accounting method that allows them to keep a constant net asset value (NAV) per share of one dollar. This convention makes MMMFs easier to use for transaction purposes and thus a close substitute for bank deposits. In September 2008, after Lehman Brothers failed, there were sizeable withdrawals from MMMFs. The immediate cause was losses to the Reserve Primary MMMF, which had a sizeable exposure to Lehman Brothers commercial paper. This loss led the fund to "break the buck," that is, the NAV of the fund dropped below one dollar.[8] There were large withdrawals from this fund, followed soon after by large withdrawals from some other MMMFs. According to the Investment Company Institute (2009), there was a large shift of money market funds by institutional investors from prime MMMFs—those that could invest in nongovernment securities—to government MMMFs.[9]

One reason that the institutional investors ran is that money market fund accounting in certain cases can give an incentive to run. In order to preserve their stable NAV, MMMFs are not continuously marked to market. Instead, most use the "amortized cost" method to value their assets (Cook and Duffield 1993). This method values a security at its acquisition cost and accrues interest uniformly over the security's remaining maturity. If the probability of a security defaulting goes up or, worse, if a default occurs, the value of an MMMF share will be temporarily less than the NAV of one. Selling shares in anticipation of such an event would let an investor in the pool receive the NAV of one, leaving other investors to bear the full drop in value of the securities.

---

[8] Drops in the NAV have happened before to other funds, but the sponsor of the fund had always made a transfer to the fund to raise the NAV to one.

[9] Retail investors did not run their funds.

This was a factor in the large withdrawals from the Reserve Primary MMMF. Withdrawals from other funds may have been driven by similar concerns as well as a general concern that assets in prime MMMFs would end up illiquid or in default. Some funds suspended withdrawals and at least one liquidated in order to distribute its proceeds equally among its investors (Investment Company Institute 2009). Withdrawals from these funds were stopped with the government introduction of insurance for the MMMFs. Interestingly, according to Swagel (2009), a significant motivation in providing the insurance was the concern that *issuers* of commercial paper would not be able to roll it over and would be forced to make large draws on their lines of credit from banks, assuming they even had them.

Similar to MMMFs are government investment pools. Many states offer funds to their municipalities in which they can pool their funds to invest in cash-like instruments (Cook and Duffield 1993). The Florida investment pool ran into trouble when it took losses on its securities and some became illiquid. This led some of the Florida municipalities that participated in the fund to withdraw their investments. The Florida fund was unable to meet these redemptions, so it partially suspended redemption and worked out a long-term scheme to distribute its assets to its members (Evans 2007; Evans and Preston 2007).

The wide variety of financial arrangements that experienced run-like behavior demonstrate that the DD model is just as relevant today as it was historically. These arrangements also provide important data for evaluating the DD model and will motivate much future work on it.

2. THE ARTICLES IN THIS ISSUE

Since DD, a lot of work has gone into developing a better understanding of what is essential to Diamond and Dybvig's fragility result and what can be done to prevent it. This literature is large, spans a long period of time, and is often technical. The article by Huberto Ennis and Todd Keister gives people unfamiliar with DD a nontechnical overview of this literature. They pay special attention to the roles of sequential service and uncertainty about aggregate liquidity needs.

The article by Edward Green focuses on a more specific issue. He examines the role of limited liability and the optimality of bailouts for partially financing illiquid investments. He defines a bailout as a combination of early liquidation along with taxes and transfers that relax the limited liability constraint. In an economy with limited liability, he finds that state-contingent payments from the taxpayer to the banking system are part of an optimal allocation. He is careful to point out that he does not address moral hazard, which could significantly alter this conclusion.

Green's focus on the limited liability constraint is important, not only because of its implications for bailouts, but also because relaxing limited liability

was an important part of historical banking arrangements. Until the 1930s, equity owners of national banks in the United States had "double liability," that is, they could be required to contribute up to the par amount of their equity to meet the bank's obligations (Macey and Miller 1992). Furthermore, in the 18th and 19th centuries, many Scottish banks had unlimited liability (Cowen and Kroszner 1989). As we consider how to redesign the financial system, limited liability rules may be one direction worth exploring.

The final two articles are about monetary theory. Historically, monetary and banking economics are deeply connected. Circulating bank liabilities are often called "inside money," that is, circulating debt that is backed by private assets. Despite this connection, money and banks are often modeled in isolation. The article by Ricardo Cavalcanti bridges monetary and banking theory by providing some recent history of thought about the two areas. He discusses the precursors to the Diamond-Dybvig model in which the traditional strategy, still found in textbooks, was to append a banking sector onto a market model. Cavalcanti argues that one of DD's main contributions was to take the different strategy of mechanism design theory, which focuses on information frictions and does not take the market structure as exogenous. He then proceeds to connect this strategy with monetary theory, in particular, the random matching models in which related information and commitment issues make fiat money valuable. He concludes by pointing out how recent models in this literature are altering information assumptions in order to incorporate bank-like organizations.

The article by William Jack, Tavneet Suri, and Robert Townsend continues the monetary economics theme by describing the recent development of mobile phone banking in Kenya and juxtaposing these developments with monetary theory. One advantage of this strategy is that, by looking at an economy that is simpler on some dimensions than that of the United States, it is easier to measure and understand the forces at work. Indeed, a developing country economy can be viewed as a laboratory for understanding more complex environments, much like biologists study animal biology to understand human biology.

This line of research is very fruitful. Not only does it raise important monetary and banking policy questions for Kenya, but it also points to parallels with the United States. In Kenya, mobile phone e-money looks like inside money, just as some of the financial liabilities created by the U.S. financial sector, such as repos, also look a lot like inside money. One implication of the monetary theories that they describe is that there is not a simple monetary policy that is robust across the various classes of models. This has implications not only for Kenyan monetary policy but also for evaluating financial reform proposals in the United States.

3. CONCLUDING COMMENT

We in the research department of the Federal Reserve Bank of Richmond have been fortunate to have Doug Diamond as a visiting scholar for the last 20 years. Personally, I always look forward to his visits. He is full of ideas and energy and is a delight to talk to. This special issue is dedicated not only to honor his famous article with Philip Dybvig, but also Doug's many contributions to our research department and this journal over the years.

REFERENCES

Bagehot, Walter. 1873. *Lombard Street: A Description of the Money Market*. New York: John Wiley & Sons, Inc.

Calomiris, Charles W., and Joseph R. Mason. 1997. "Contagion and Bank Failures During the Great Depression: The June 1932 Chicago Banking Panic." *American Economic Review* 87 (December): 863–83.

Campbell, Doug. 2005. "Why Economists Still Worry about Bank Runs." Federal Reserve Bank of Richmond *Region Focus*, 36–8 (Fall).

Cook, Timothy Q., and Jeremy G. Duffield. 1993. "Money Market Mutual Funds and other Short-Term Investment Pools." In *Instruments of the Money Market*, edited by Timothy Q. Cook and Robert K. LaRoche. Richmond, Va.: Federal Reserve Bank of Richmond, 156–72.

Covitz, Daniel M., Nellie Liang, and Gustavo A. Suarez. 2009. "The Evolution of a Financial Crisis: Panic in the Asset-Backed Commercial Paper Market." Federal Reserve Board FEDS Working Paper 2009-36 (March).

Cowen, Tyler, and Randall Kroszner. 1989. "Scottish Banking Before 1845: A Model for Laissez-Faire?" *Journal of Money, Credit and Banking* 21 (May): 221–31.

Diamond, Douglas W. 1984. "Financial Intermediation and Delegated Monitoring." *Review of Economic Studies* 51 (July): 393–414.

Diamond, Douglas W. 2007. "Banks and Liquidity Creation: A Simple Exposition of the Diamond-Dybvig Model." Federal Reserve Bank of Richmond *Economic Quarterly* 93 (Spring): 189–200.

Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 91 (June): 401–419.

Diamond, Douglas W., and Philip H. Dybvig. 2000. "Bank Runs, Deposit Insurance, and Liquidity." Federal Reserve Bank of Minneapolis *Quarterly Review* 24 (Winter): 14–23.

Evans, David. 2007. "Florida School Fund Rocked by $8 Billion Pullout." www.bloomberg.com (November 28).

Evans, David, and Darrell Preston. 2007. "Florida Investment Chief Quits; Fund Rescue Approved." www.bloomberg.com (December 4).

Gorton, Gary B., and Andrew Metrick. 2009. "Securitized Banking and the Run on the Repo." Yale ICF Working Paper 09-14 (November).

Grind, Kirsten. 2009. "The Downfall of Washington Mutual: Inside the Frenzied Effort to Prevent the Largest Bank Failure in US History." http://seattle.bizjournals.com/seattle/stories/2009/09/28/story1.html (September 28).

Han, Song, and Dan Li. 2008. "Liquidity, Runs, and Security Design." SSRN Working Paper (January 15).

Investment Company Institute. 2009. "Report of the Money Market Working Group." Washington, D.C.: ICI (March 17).

Macey, Jonathan R., and Geoffrey P. Miller. 1992. "Double Liability of Bank Shareholders: History and Implications." *Wake Forest Law Review* 27: 31–62.

Office of the Inspector General. 2009. "Safety and Soundness: Material Loss Review of IndyMac Bank, FSB." Audit Report OIG-09-032. Washington, D.C.: U.S. Department of the Treasury (February 26).

Swagel, Phillip. 2009. "The Financial Crisis: An Inside View." *Brookings Papers on Economic Activity* Spring: 1–63.

Shin, Hyun Song. 2009. "Reflections on Northern Rock: The Bank Run that Heralded the Global Financial Crisis." *Journal of Economic Perspectives* 23 (Winter): 101–19.

Timberlake, Richard H., Jr. 1984. "The Central Banking Role of Clearinghouse Associations." *Journal of Money, Credit and Banking* 16 (February): 1–15.

Todd, Walker F. 1994. "Lessons from the Collapse of Three State-Chartered Private Deposit Insurance Funds." Federal Reserve Bank of Cleveland *Economic Commentary* May (1): 1–6.

White, Lawrence J. 1991. *The S&L Debacle: Public Policy Lessons for Bank and Thrift Regulation*. New York: Oxford University Press.

Economic Quarterly—Volume 96, Number 1—First Quarter 2010—Pages 11–32

Bailouts

Edward J. Green

In the United States during 2008–2009, as in previous episodes here and other countries, supplying funding to financial intermediaries and other firms was a component of the government's response to a financial crisis. Some of these funding initiatives have been characterized—and, in some quarters, heavily criticized—as being *bailouts*: transfers from the government, made to firms (and sometimes other entities such as city governments) or to their creditors in order to avert insolvency or mitigate its effects, that the recipients are not anticipated to repay. Note that this definition distinguishes bailouts from bona fide government loans.[1] Henry Thornton (1802) and Walter Bagehot (1877) explained why it is good public policy for government to lend to firms (particularly to banks) in a financial crisis, and today that justification is widely accepted. Bailouts remain highly controversial, however.

Many economists perceive bailouts to be a costly manifestation of time inconsistency on the part of policymakers. That is, the government threatens that an entity that becomes insolvent must fail rather than being rescued, but subsequently, perhaps out of fear that insolvency would harm many people who bear no responsibility for it, the entity will be rescued when push comes to shove. Anticipating this denouement, the owners and managers of the entity

This article extends research that was presented in the 2007 Arijit Mukherji Memorial Lecture at the University of Minnesota. Professor Mukherji's untimely death, in 2000, deprived the economics and accounting disciplines of an ascending leader. Research support for this article has been provided by a gift from the Human Capital Foundation to the Center for the Study of Auctions, Procurements, and Competition Policy at Penn State University. I am grateful to Huberto Ennis for help in properly formulating Proposition 4, and to both him and also Devin Reilly for advice regarding the example in Section 8. The views expressed in this article do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: eug2@psu.edu.

[1] Even if a transfer takes the explicit form of a loan, it has an implicit bailout component if the interest rate is too low to compensate the lender at market terms for the risk of default. In particular, if the market price of risk is set by risk-neutral traders, then a loan has a bailout component unless it is actuarially sound—that is, unless the lender is making a bet at fair odds that it will be repaid. Specifically, suppose that a loan of size $\$L$ is made at interest rate r, and is anticipated to be repaid with probability p when it is made. Assume for simplicity that the loan will be repaid either in full, or else not at all. Then the loan is actuarially sound if $p \cdot r \cdot L \geq (1-p) \cdot L$, or equivalently if $p \cdot r \geq 1 - p$.

make inadvisable investments, taking risks that they would have avoided if the threat not to assist had been taken seriously. This view, formalized by Kareken (1983) and more recently elaborated by Stern and Feldman (2004), is a cogent, prima facie reason to judge that bailouts are a socially inefficient form of government intervention in the economy.

Nevertheless, despite this logic, numerous academic economists, policymakers, and market participants argued publicly that the 2008–2009 bailout was an indispensable policy action. In their view, there was considerable risk that the economy would have suffered serious, long-term harm if the finance, automobile, and housing industries had not been subsidized. Presumably they were concerned that millions of people would face the sort of immediate harm that expositions of the time-inconsistency argument typically cite, but they spoke of a greater, more persistent harm. In their view, if government did not provide a bailout in circumstances where to do so was vital, then incentives for socially beneficial investment would be impaired in a way that might take decades to repair. This vision is the polar opposite of the time-inconsistency vision, which sees investment incentives being harmed by the occurrence of bailouts rather than by their nonoccurrence.

The goal of this article is to formulate an economic model, in terms of which the concern just described can be understood. This is a very limited goal. It is not even to provide a prima facie argument that conducting a bailout is likely to be good policy. To meet the goal, the model need only establish that a bailout would be economically efficient under some conceivable conditions in some economy that shares salient features of the actual one.

In an economy in which a bailout of firms might be efficient, there must be some reason for production to be undertaken by firms that issue financial claims against which they might default. This feature is necessary because, if there were no good reason for firms ever to become insolvent, then an optimal policy would be to prevent them from ever taking that risk, rather than to allow them to take it and to help them when insolvency occurs.

In particular, besides firms being able to do something for their investors that the investors cannot do for themselves, there must be some constraint on a firm's ability to issue financial claims that would only have to be paid in those states of nature where the firm had the capacity to pay them. The Modigliani-Miller theorem (cf. Stiglitz 1969) states that, if a firm could contract ex ante for the payments that it could make and receive at every date, in any state of the world, then any production plan could be financed in such a way that the firm could not possibly become insolvent. Thus, a threshold condition for an economic model to be suitable for studying insolvency is that it must rule out some contracts that a firm might make in principle so that the no-insolvency implication of the Modigliani-Miller theorem will be avoided.

A well-known model with these features is the model of bank runs formulated by John Bryant (1980) and Douglas Diamond and Philip Dybvig (1983).

A firm (which those authors interpret to be a bank) can improve on autarkic production by pooling its investors' risks of idiosyncratic shocks to their respective preferences. It is assumed that the firm can only fund its production by issuing standard debt contracts rather than by issuing financial claims in a completely flexible manner.

The model to be formulated here closely resembles the Diamond-Dybvig model. Although those authors (and also Bryant) were particularly concerned with the possibility that a solvent firm might become illiquid—indeed, to formalize that distinction was an important aspect of their contribution—the model can be parameterized in such a way that optimal financial and production decisions must lead to insolvency in some states of the world.

Rather than assuming that a standard debt contract is the only available financial claim, financial flexibility will be constrained in the present model by assuming that the firm is a limited-liability corporation. In fact, Modigliani and Miller cited limited liability as a consideration that arguably prevents their theorem from holding precisely in an actual economy. Like the Diamond-Dybvig model, the present model is a partial-equilibrium model in the sense that it assumes a constraint on financing opportunities, rather than deriving that constraint as an implication of, or as an optimal policy response to, economic primitives such as tastes, technology, and privacy of information. For an informal discussion of the history and economic rationale of limited liability, see Easterbrook and Fischel (1985).

The firm is modeled here as making payouts of the good it produces, and those payouts cannot exceed in the aggregate the firm's output in any state of nature. However, there is an equivalent way of describing the way in which the allocation is implemented. That is, the firm promises state-contingent payouts to investors that exceed, in the aggregate, its output in some states of nature. Then, in those states of nature, the firm receives a tax-funded subsidy to bridge the gap between its output and its aggregate liabilities. According to this description, the tax/transfer scheme is a bailout of the insolvent firm. The tax is collected on investors' endowments at the date when the subsidy is paid. The limited-liability constraint specifies that the firm cannot claim those endowments directly, so the government's authority to tax them must be invoked in order to substitute for the promised payouts that the firm is unable to make.

It might be asked, what sense does it make to tax investors' endowments and then return them? The answer is that the tax is a lump-sum tax but the indemnification is dependent on the investors' continued participation in the firm. Thus, the tax-subsidy scheme can affect incentives. From an ex-ante perspective, it may be essential for providing sufficient incentive to invest along with others, some of whom (that is, those who suffer an adverse preference shock) foreseeably will liquidate their investments prematurely.

Since there is only one firm in the entire economy in this model, it should be interpreted to represent the entire firm sector of the economy, including banks, other financial firms, and nonfinancial firms. A main insight of Diamond and Dybvig is that banks contribute to economic welfare by engaging in *maturity transformation*, that is, by "borrowing short and lending long." Banks are not the only firms that do this, however. Recent research (cf. Acharya, Gale, and Yorulmazer 2009) emphasizes that a nonfinancial firm can engage in maturity transformation on its own behalf through the market for short-term corporate debt, with essentially the same implication as if it had borrowed from a maturity-transforming bank. The terminology adopted in this paper—"firm," rather than "bank"—reflects a view that the welfare analysis of bailouts as public policy is largely the same, whether the recipient is a financial or a nonfinancial firm.[2] Like the Diamond-Dybvig model, the present model concerns a policy response to a problem in a broad sector of an economy. Regardless of whether there is one direct recipient of government funds or there are many, and regardless of whether those direct recipients are financial firms or nonfinancial ones, a bailout affects the position of the firm sector (including its investors) in the aggregate.

1. THE ENVIRONMENT

There are three dates, denoted by 0, 1, 2. There is a large population of investors, each of whom randomly has one of two utility functions. Each investor behaves atomistically, and in particular, ascribes zero probability to the event that he could be a pivotal liability holder whose decision to demand payment might force the firm to default.

There is one good at each date, which can be either consumed or, except at the terminal date 2, transformed into the good at the next date by the technologies described below.[3] Each investor is endowed with $\bar{x}_0 > 0$ units of good at date 0 and $\bar{x}_2 > 0$ units of good at date 2 but is not endowed with any of the date-1 good.

If he is *impatient* (type 1), then an investor wants to maximize his consumption at date 1 until it has reached a high threshold. If he is *patient* (type 2), then he wants to maximize the sum of his consumption at date 1 and date 2. Date 0 is a date at which each investor can invest his endowment or exchange for a liability of the firm, which invests it, but at which no consumption takes

[2] Diamond and Dybvig introduced the sequential-service constraint, a feature of their model environment that prevented market transactions from decentralizing the same allocation as banking contracts implement. Analogously, the limited-liability constraint prevents market transactions from substituting for a combination of contracts and bailouts in the model to be analyzed here.

[3] There will be no explicit technology for transforming date-1 consumption to date-2 consumption, but refraining from early liquidiation of illiquid investment is tantamount to such a technology.

place. Each agent privately learns his own type at date 1 but remains ignorant of others' types.

An investor's utility function has three arguments: consumption at date 1, consumption at date 2, and the investor's type. An impatient investor has utility

$$\begin{aligned} u(c_1, c_2, 1) &= v(c_1) + c_2 \\ v(x) &= \begin{cases} \eta x & \text{if } x \leq \theta; \\ \eta\theta + (x - \theta) & \text{if } x > \theta; \end{cases} \\ \eta &> 1. \end{aligned} \tag{1}$$

A patient investor has utility

$$u(c_1, c_2, 2) = c_1 + c_2. \tag{2}$$

At date 0, before he knows his type, and at date 1, if the consequences of a decision depend on other investors' types of which he is ignorant, an investor maximizes expected utility as explained below.

There is a risk that concerns the fraction of the population that is patient. In every state of the world, either a fraction μ_Γ or μ_B of the investors are impatient, where $0 < \mu_\Gamma < \mu_B < 1$. In subsequent analysis, it will be assumed that $\bar{x}_0/\mu_B < \theta < \bar{x}_0/\mu_\Gamma$.

An *aggregate state* of the world in which fraction μ_B of investors are impatient occurs with probability β, and $0 < \beta < 1$. Denote this aggregate state by B, and denote its complement by Γ. (Strictly speaking, B is the event that comprises all of the *bad* states in which some group of μ_B investors is impatient, and the *good* aggregate state Γ is the complementary event.)

All investors are equally likely to be patient, and no investor's type is more highly correlated with the aggregate state than any other's type is. Thus, if μ_1^* is the probability that a particular agent is impatient, then the following equation is satisfied:[4]

$$\mu_1^* = \beta\mu_B + (1 - \beta)\mu_\Gamma. \tag{3}$$

It follows that, if an investor knows that he is impatient but knows nothing of other investors' types, then his probability belief that event B has occurred is[5]

$$\beta_1^* = \beta\mu_B/\mu_1^*. \tag{4}$$

[4] An asterisk superscript indicates in this article that a probability will appear explicitly in an investor's expected-utility calculation.

[5] Probability β_1^* is calculated according to Bayes' Theorem. Game theorists call this the investor's *interim probability*, to distinguish it from *posterior probability*, which reflects knowledge of both the investor's own type and also the other investors' types. *Interim expected utility* is the mean of the investor's utility function with respect to the interim probability measure.

Similarly, if the probability that a particular investor is patient is denoted by μ_2^*, then

$$\mu_2^* = \beta(1 - \mu_B) + (1 - \beta)(1 - \mu_\Gamma), \tag{5}$$

and a patient investor believes with probability β_2^* that event B has occurred, where

$$\beta_2^* = \beta(1 - \mu B)/\mu_2^*. \tag{6}$$

Intuitively, an investor assigns higher probability to B if he discovers that he is impatient than if he finds that he is patient. A routine computation shows that, correspondingly, $\beta_1^* > \beta_2^*$.

Investment must be undertaken at date 0. There are two technologies, each of which has return that is linear in investment. *Liquid technology* is just storage: it returns one unit of output at date 1 for each unit of investment. *Illiquid technology* returns $R > 1$ units of consumable output at date 2 for each unit of investment. However, if a unit is withdrawn at date 1, then it only yields $r < 1$ units of consumable output. Assume that

$$R/r < \eta. \tag{7}$$

The economic implication of this inequality is that, for an impatient investor whose date-1 consumption is below θ (so that the marginal utility of consumption at date 1 is η), the marginal rate of substitution of c_1 for c_2 (that is, R/r) is higher than the marginal rate of transformation of c_1 into c_2 by choosing between alternate uses of the illiquid technology.

2. EXPECTED UTILITY

The probabilities defined and calculated above provide the basis for calculating investors' prior and conditional expected utilities of state-contingent allocations. In the calculations below, and throughout the rest of this article, s will denote an individual investor's state, σ will denote an aggregate state, and t will denote a date in the model economy. For $s \in \{1, 2\}\, \sigma \in \{B, \Gamma\}$, and $t \in \{1, 2\}$, let $c_t^{s\sigma}$ denote the consumption level at date t in the event that the investor's state is s and the aggregate state is σ. An *allocation* specifies the consumption level in each of the eight possible combinations of date, investor's state, and aggregate state.[6]

[6] As is typical in models in which there are many agents whose individual states are i.i.d. conditional on the aggregate state, so that a law of large numbers can be presumed to hold, it is unnecessary to distinguish formally between an economy-wide allocation in a generic state of the world and a bundle of state-contingent commodities for an individual agent in the economy.

Denote the prior expected utility of allocation c by $U_0(c)$, which is

$$\begin{aligned}&\beta\left[\mu_B u\left(c_1^{1B}, c_2^{1B}, 1\right) + (1-\mu_B)\, u\left(c_1^{2B}, c_2^{2B}, 2\right)\right]\\&+ (1-\beta)\left[\mu_\Gamma u\left(c_1^{1\Gamma}, c_2^{1\Gamma}, 1\right) + (1-\mu_\Gamma)\, u\left(c_1^{2\Gamma}, c_2^{2\Gamma}, 2\right)\right].\end{aligned} \tag{8}$$

If an agent learns that he is impatient (type 1), then his interim expected utility is $U_1(c)$, which is

$$\beta_1^* u\left(c_1^{1B}, c_2^{1B}, 1\right) + \left(1-\beta_1^*\right) u\left(c_1^{1\Gamma}, c_2^{1\Gamma}, 1\right). \tag{9}$$

If he learns that he is patient (type 2), then his interim expected utility is $U_2(c)$, which is

$$\beta_2^* u\left(c_1^{2B}, c_2^{2B}, 2\right) + \left(1-\beta_2^*\right) u\left(c_1^{2\Gamma}, c_2^{2\Gamma}, 2\right). \tag{10}$$

3. COOPERATIVE AND AUTARKIC PRODUCTION FEASIBILITY

The endowment and technologies described above imply a set of technically feasible cooperative production outcomes if all investors' endowments are invested jointly, and a set of technically feasible individual production outcomes if a single investor invests autarkically. Assume free disposal: If a production outcome is feasible, then any outcome that provides less consumption at each date (and, in the case of individual feasibility, for each type) is also feasible.

Since technology is linear, cooperative feasibility can be considered in per capita terms. A *cooperative production plan* specifies an amount, ι, of the endowment to be invested in the illiquid technology and amounts, ϵ_B and ϵ_Γ, of that illiquid investment to be liquidated at date 1 in the two possible aggregate states. *Technical feasibility* requires that, for each $\sigma \in \{B, \Gamma\}$,

$$0 \leq \epsilon_\sigma \leq \iota \leq \bar{x}_0. \tag{11}$$

Given production plan π, let $y_{\sigma t}$ denote the output at date t in aggregate state σ. Then, for $\pi = (\iota, \epsilon_B, \epsilon_\Gamma)$, y is the vector satisfying

$$\begin{aligned}y_{\sigma 1} &= (\bar{x}_0 - \iota) + \epsilon_\sigma r;\\ y_{\sigma 2} &= (\iota - \epsilon_\sigma)\, R.\end{aligned} \tag{12}$$

Proposition 1 *If $\pi = (\iota, \epsilon_B, \epsilon_\Gamma)$ is technically feasible and $\min(\epsilon_B, \epsilon_\Gamma) > 0$, then another technically feasible production plan provides strictly more output at both dates and in both aggregate states than π does.*

Proof. Let $\pi = (\iota, \epsilon_B, \epsilon_\Gamma)$ be a technically feasible production plan, and define $\pi' = (\iota - \min(\epsilon_B, \epsilon_\Gamma), \epsilon_B - \min(\epsilon_B, \epsilon_\Gamma), \epsilon_\Gamma - \min(\epsilon_B, \epsilon_\Gamma))$. Then

π' is also technically feasible, and it has weakly higher output than π at both dates and in both aggregate states, because resources that π allocates to early-liquidated illiquid-technology production are reallocated in π' to liquid-technology production of the same goods (that is, goods at date 1 in the two aggregate states). If both ϵ_B and ϵ_Γ are positive, then π' produces strictly more output at date 1 in each aggregate state than π does, and it produces identical output to π at date 2 in each aggregate state. Then, by devoting a slightly higher investment than $\iota - \min(\epsilon_B, \epsilon_\Gamma)$ to illiquid production, a new, technically feasible production plan can be constructed that provides strictly more output at both dates and in both aggregate states than π does. ■

An allocation is *technically feasible for cooperative production* if there is a technically feasible cooperative production plan such that, at each date and in each aggregate state, the impatient and patient agents together consume no more than the sum of the output of that plan and the endowment at that date (that is, 0 at date 1 or $\bar{x}_2$ at date 2). Specifically, c is technically feasible for cooperative production if, for some feasible production plan π,

$$\begin{aligned} \mu_\sigma c_1^{1\sigma} + (1-\mu_\sigma)\, c_1^{2\sigma} &\leq y_{\sigma 1}; \\ \mu_\sigma c_2^{1\sigma} + (1-\mu_\sigma)\, c_2^{2\sigma} &\leq y_{\sigma 2} + \bar{x}_2. \end{aligned} \tag{13}$$

An *autarkic production plan* specifies a fraction of $\bar{x}_0$ to be invested in the illiquid technology and fractions ϵ_1 and ϵ_2 of that investment to be liquidated at date 1 if the investor is impatient or patient, respectively. Note that, since an individual investor does not observe the aggregate state, an autarkic production plan cannot depend on it. The output of an autarkic production plan is defined analogously to (12):

$$\begin{aligned} y_{s1} &= (x_0 - \iota) + \epsilon_s r; \\ y_{s2} &= (\iota - \epsilon_s)\, R. \end{aligned} \tag{14}$$

Moreover, since the aggregate state is irrelevant to either the production possibility set or the preferences of an investor of either type, there is no reason for an autarkic investor's allocation to depend on it. Because an autarkic investor does not need to acquire private information from anyone else in order to implement his plan, technical constraints are the only feasibility constraints. Thus, define allocation c to be *feasible for autarkic production* if, for some autarkic production plan π,

$$\begin{aligned} c_1^{sB} &\leq y_{s1}; \\ c_2^{sB} &\leq y_{s2} + \bar{x}_2; \\ c_t^{s\Gamma} &= c_t^{sB}. \end{aligned} \tag{15}$$

Proposition 2 *If allocation c is feasible for autarkic production, then there is a technically feasible cooperative production plan with sufficient output per capita to provide every investor with the same level of consumption at both dates, in every state, as c provides. If at least one type of investor would liquidate a positive amount of illiquid investment at date 1 in an autarkic production plan for c, then there is a cooperative plan with sufficiently high output to provide every investor with higher consumption at both dates, in every state, than c provides.*

Proof. If $(\iota, \epsilon_1, \epsilon_2)$ is an autarkic production plan that produces sufficient output for c according to (15), then $(\iota, \mu_B\epsilon_1 + (1-\mu_B)\,\epsilon_2, \mu_\Gamma\epsilon_1 + (1-\mu_\Gamma)\,\epsilon_2)$ is a cooperative production plan that produces sufficient output for c according to (12). The second assertion in this proposition follows from Proposition 1 since, if at least one of ϵ_1 and ϵ_2 is positive, then $\epsilon_B = \mu_B\epsilon_1 + (1-\mu_B)\,\epsilon_2$ and $\epsilon_\Gamma = \mu_\Gamma\epsilon_1 + (1-\mu_\Gamma)\,\epsilon_2$ imply that both ϵ_B and ϵ_Γ are positive. ■

4. OPTIMAL AUTARKIC PRODUCTION

Consider the problem of optimizing expected utility, $U_0(c)$, among allocations that are feasible for autarkic production. Since the feasibility condition (15) requires that $c_t^{s\Gamma} = c_t^{sB}$, the definition (8) of $U_0(c)$ reduces to

$$U_0(c) = \mu_1^* u\left(c_1^{1B}, c_2^{1B}, 1\right) + \mu_2^* u\left(c_1^{2B}, c_2^{2B}, 2\right). \tag{16}$$

Since U_0 is strictly increasing in all of its consumption arguments, the feasibility constraints will all hold with equality in (15). That is, c is the entire output of some autarkic production plan $(\iota, \epsilon_1, \epsilon_2)$, together with the endowment $\bar{x}_2$ at date 2. Making this substitution into (16), and expanding u according to its defining equations (1) and (2) yields

$$\begin{aligned} U_0(c) \;=\; & \mu_1^* \left[\upsilon\left((\bar{x}_0 - \iota) + \epsilon_1 r\right) + (\iota - \epsilon_1)\, R + \bar{x}_2\right] \\ & + \mu_2^* \left[(\bar{x}_0 - \iota) + \epsilon_2 r + (\iota - \epsilon_2)\, R + \bar{x}_2\right]. \end{aligned} \tag{17}$$

Recall that, by (1), $1 \leq \upsilon' \leq \eta$, so, if Δ_ι denotes the derivative of the right side of (17) with respect to ι, then[7]

$$R - \left(\mu_1^*\eta + \mu_2^*\right) \leq \Delta_\iota \leq R - 1. \tag{18}$$

[7] Function υ is differentiable except at θ, where the rightmost and leftmost terms in (18) are the left and right directional derivatives of υ, respectively.

If

$$R - \left(\mu_1^* \eta + \mu_2^*\right) > 0, \tag{19}$$

then the optimal level of ι is the maximal investment level $\bar{x}_0$. Condition (19) will be assumed henceforth, in order to focus on this case.

The derivative of the right side of (17) with respect to ϵ_2 is $\mu_2^*(r - R) < 0$, so the optimal level of ϵ_2 is the minimum level 0. Let Δ_ϵ denote the derivative of the right side of (17) with respect to ϵ_1:

$$\Delta_\epsilon \begin{cases} \mu_1^* (\eta r - R) & \text{if } \epsilon_1 r < \theta; \\ \mu_1^* (r - R) & \text{if } \epsilon_1 r > \theta. \end{cases} \tag{20}$$

By assumption (7), $\Delta_\epsilon > 0$ if $\epsilon_1 r$ is to the left of θ, so the optimal autarkic production plan must set $\epsilon_1 = \bar{x}_0$ if $\bar{x}_0 r < \theta$. Subsequent analysis will focus on this case.

The following proposition recapitulates what has been established in this section.

Proposition 3 *If* $R - \left(\mu_1^* \eta + \mu_2^*\right) > 0$ *and* $\bar{x}_0 r < \theta$, *then* $(\bar{x}_0, \bar{x}_0, 0)$ *is the optimal autarkic production plan. If allocation* c *is the output of this plan, then* $U_0(c) = \bar{x}_0 \left(\mu_1^* \eta r + \mu_2^* R\right) + \bar{x}_2$.

5. OPTIMAL COOPERATIVE PRODUCTION

Consider the optimal allocation that is technically feasible for cooperative production. Recall that impatient investors who receive date-1 consumption less than θ have marginal utility $\eta > 1$ for consumption at that date, all other investors have marginal utility 1 for date-1 consumption, and all investors have marginal utility 1 for date-2 consumption. Recall that output is described in per capita terms, so the greatest amount of output $y_{\sigma 1}$ that can be given to each investor of type 1 at date 1 in aggregate state σ is $y_{\sigma 1}/\mu_\sigma$. It follows that, to maximize expected utility among allocations that distribute y, it is necessary and sufficient that, for $\sigma \in \{B, \Gamma\}$, $\min(y_{\sigma 1}/\mu_\sigma, \theta) \leq c_1^{1\sigma}$. In particular, it is optimal to allocate all production output to the impatient investors, at both dates and in both aggregate states, and to allow every investor to consume his own endowment, $\bar{x}_2$, of the date-2 good. The level of ex-ante expected utility that this allocation provides is

$$\begin{aligned} &\beta \mu_B \left(\eta \min\left(y_{B1}/\mu_B, \theta\right) + \max\left(y_{B1}/\mu_B - \theta, 0\right) + y_{B2}/\mu_B\right) \\ &+ (1 - \beta)\, \mu_\Gamma \left(\eta \min\left(y_{\Gamma 1}/\mu_\Gamma, \theta\right) + \max\left(y_{\Gamma 1}/\mu_\Gamma - \theta, 0\right) + y_{\Gamma 2}/\mu_\Gamma\right) \\ &+ \bar{x}_2. \end{aligned} \tag{21}$$

If y is the output of cooperative production plan $(\iota, \epsilon_B, \epsilon_\Gamma)$, then (21) is equivalent to

$$\begin{aligned}
&\beta\mu_B\eta \min\left(\left(\left(\bar{x}_0 - \iota\right) + \epsilon_B r\right)/\mu_B, \theta\right) \\
&+ \max\left(\left(\left(\bar{x}_0 - \iota\right) + \epsilon_B r\right)/\mu_B - \theta, 0\right) + \left(\iota - \epsilon_B\right) R/\mu_B\big) \\
&+ \left(1 - \beta\right)\mu_\Gamma\eta \min\left(\left(\left(\bar{x}_0 - \iota\right) + \epsilon_\Gamma r\right)/\mu_\Gamma, \theta\right) \\
&+ \max\left(\left(\left(\bar{x}_0 - \iota\right) + \epsilon_\Gamma r\right)/\mu_\Gamma - \theta, 0\right) + \left(\iota - \epsilon_\Gamma\right) R/\mu_\Gamma\big) + \bar{x}_2. \quad (22)
\end{aligned}$$

Assume that it is technically feasible to provide consumption at least as high as the low-marginal-utility threshold to impatient investors at date 1 in aggregate state Γ, but not in B. That is,

$$\bar{x}_0/\mu_B < \theta < \bar{x}_0/\mu_\Gamma. \quad (23)$$

It is optimal to liquidate all investment in state B. The reason is that, regardless of the value of ι, date-1 output with complete liquidation will be $(\bar{x}_0 - \iota) + \iota r$, which is not greater than $\bar{x}_0$. If this output is all given to impatient investors to consume, then each of them receives $[(\bar{x}_0 - \iota) + \iota r]/\mu_B \leq \bar{x}_0/\mu_B < \theta$, at which level the marginal utility of date-1 consumption in state B is η, versus 1 for date-2 consumption. The marginal rate of transformation of date-0 endowment to date-1 consumption by means of making illiquid investment but liquidating it early is r, while the marginal rate of transformation to date-2 consumption by not liquidating is R, so (7) entails that early liquidation is optimal.

Under some circumstances, it is optimal to make illiquid investment up to the point where just enough is left over to provide every impatient investor with θ units of consumption at date 1 in state Γ and not to liquidate any of the investment in that state. That is, $(\bar{x}_0 - \mu_\Gamma\theta, \bar{x}_0 - \mu_\Gamma\theta, 0)$ is the optimal cooperative investment plan. These circumstances are now characterized.

Let Δ^- and Δ^+ denote the left- and right-hand derivatives of (22) with respect to ι, evaluated at $(\iota, \epsilon_B, \epsilon_\Gamma) = (\bar{x}_0 - \mu_\Gamma\theta, \bar{x}_0 - \mu_\Gamma\theta, 0)$. Then

$$\begin{aligned}
\Delta^- &= \beta\left(R - \eta\right) + \left(1 - \beta\right)\left(R - 1\right); \\
\Delta^+ &= R - \eta. \quad (24)
\end{aligned}$$

Now assume that

$$R - \eta < 0 < R - \left(\beta\eta + 1 - \beta\right). \quad (25)$$

This entails that Δ^- and Δ^+ are positive and negative, respectively, so the maximum is achieved at $\bar{x}_0 - \mu_\Gamma\theta$, where these directional derivatives were evaluated. Also, at that level of ι, the right-hand derivative of (22) with respect to ϵ_Γ is $r - R < 0$, which is sufficient, given the concavity of the objective

function, for 0 to be the optimum. The following proposition restates the conclusions of this argument.

Proposition 4 *If* $\bar{x}_0/\mu_B < \theta < \bar{x}_0/\mu_\Gamma$ *and* $R - \eta < 0 < R - (\beta\eta + 1 - \beta)$, *then* $(\bar{x}_0 - \mu_\Gamma\theta, \bar{x}_0 - \mu_\Gamma\theta, 0)$ *is the unique cooperative production plan, the output of which can be allocated to maximize expected utility* U_0 *among the allocations that are technically feasible for aggregate production. Allocation* c *is optimal among allocations that are feasible from this plan if and only if* $c_1^{1B} = (\mu_\Gamma\theta + (\bar{x}_0 - \mu_\Gamma\theta)\, r)/\mu_B$, $c_1^{1\Gamma} = \theta$, $c_1^{2B} = c_1^{2\Gamma} = 0$, $\mu_B c_2^{1B} + (1 - \mu_B)\, c_2^{2B} = \bar{x}_2$, *and* $\mu_\Gamma c_2^{1\Gamma} + (1 - \mu_\Gamma)\, c_2^{2\Gamma} = \bar{x}_2 + R\,(\bar{x}_0 - \mu_\Gamma\theta)$.

Note that it is possible for both premise (19) of Proposition 3 and also (25) to be satisfied, that is,

$$R - \eta < 0 < R - \max\left(\mu_1^*\eta + \mu_2^*, \beta\eta + 1 - \beta\right). \tag{26}$$

If (23) and (26) both hold, then, by Proposition 2 and Proposition 3, the optimal level of expected utility that it is technically feasible to obtain from the cooperative production characterized in Proposition 4 is strictly higher than the optimal autarkic level.

6. FIRM SECTOR, GOVERNMENT, AND FORMALIZATION OF A BAILOUT

The model being formulated and analyzed here is a rather abstract one. It makes no explicit mention of institutions, particularly of firms or of a government. Yet, the model is being proposed as a tool for gaining insight about government bailouts of firms. It is now time to discuss the intended interpretation of the model, in order to justify how a bailout is formalized within it.

The intended interpretation of cooperative production is that it is the activity of a limited-liability firm. Investors voluntarily give their initial endowments, $\bar{x}_0$, to the firm in return for state-contingent claims against it—the firm's liabilities. However, the firm is not empowered to come back to the investors at date 2 and demand part or all of their endowments, $\bar{x}_2$. Neither are the firm's creditors so empowered, if the firm defaults on its liabilities.

A firm does not have to be incorporated so that its investors have limited liability, but this is the typical legal arrangement, especially for large firms, in the United States and other industrialized countries. Historically, the widespread existence of limited-liability firms only goes back for about a century and a half. Until well into the twentieth century, U.S. banks were required by law to be chartered with shareholders having "double liability," whereby they could be required to contribute up to the par value of their

equity, if necessary, toward meeting the bank's corporate liabilities.[8] Although some firms today continue to be chartered as general partnerships or other forms of company with at least some investors having unlimited liability, it is widely accepted that the corporate form of organization confers benefits that society would forgo in an unlimited-liability regime (cf. Easterbrook and Fischel 1985).

Alongside the limited-liability corporations in a modern economy is the government, which can tax some investors and redistribute the proceeds to others. In particular, taxation can force an investor's consumption at date 2 below $\bar{x}_2$. Given an allocation c, for each individual state s and aggregate state σ there are unique $\rho^{s\sigma} \geq 0$ and $\tau^{s\sigma} \geq 0$ such that

$$\begin{aligned} c_2^{s\sigma} &= \bar{x}_2 + \rho^{s\sigma} - \tau^{s\sigma}; \\ \rho^{s\sigma} + \tau^{s\sigma} &= \min\left\{\rho+\tau \,|\, \rho \geq 0 \text{ and } \tau \geq 0 \text{ and } c_2^{s\sigma} = \bar{x}_2+\rho-\tau\right\}. \end{aligned} \tag{27}$$

(The second equation means that, at most, one of $\rho^{s\sigma}$ and $\tau^{s\sigma}$ can be positive, and that both must be zero if $c_2^{s\sigma} = \bar{x}_2$.) The quantity $\rho^{s\sigma}$ represents the investor's receipts from both corporate payouts and government subsidies, and $\tau^{s\sigma}$ represents the amount of tax that the investor has paid. Feasibility of an allocation implies a government budget constraint that subsidies cannot exceed taxes. In particular, if $\tau^{1\sigma} = \tau^{2\sigma} = 0$, then no tax is collected in aggregate state σ and therefore no subsidy can be paid out in that state. Allocation c exhibits subsidy if, for some s and σ, $\tau^{s\sigma} > 0$.

Intuitively, not every subsidy is a bailout. A bailout occurs when an extraordinarily high level of liquidation occurs and also (perhaps subsequently to the liquidation) an extraordinarily high level of subsidy is provided. Formally, a *bailout* is an aggregate state $\sigma \in \{B, \Gamma\}$ such that

$$\begin{aligned} \text{Either } 0 &< \min(\epsilon_B, \epsilon_\Gamma) \text{ and } \mu_\sigma \tau^{1\sigma} + (1-\mu_\sigma)\tau^{2\sigma} > 0, \\ \text{or, for } \sigma' &\neq \sigma, \epsilon_\sigma > \epsilon_{\sigma'} \text{ and } \mu_\sigma \tau^{1\sigma} + (1-\mu_\sigma)\tau^{2\sigma} > \mu_{\sigma'}\tau^{1\sigma'} \\ &\quad + (1-\mu_{\sigma'})\tau^{2\sigma'}. \end{aligned} \tag{28}$$

The two clauses of this definition represent situations with different welfare characteristics. In the first clause, early liquidation occurs in both aggregate states, so the allocation is technically inefficient. The clause states that, in that context, every aggregate state in which there is positive taxation (and associated subsidy) is a bailout state. Such a bailout resembles a bank run in Diamond and Dybvig's model.[9] In contrast, the second clause stipulates that

[8] Macey and Miller (1992) provide a history of this requirement, and they argue that it worked reasonably well as a prudential regulatory regime for banks.

[9] In Diamond and Dybvig (1983), there is a "run equilibrium" in which early liquidation takes place in both aggregate states. The allocation resulting from this equilibrium is inefficient, by the same logic as applies here.

early liquidation occurs only in one of the two aggregate states, specifically in the one in which tax revenue is highest. That is, liquidation is accompanied by a higher level of taxation, measured as tax revenue *per capita*, than is imposed in the non-bailout state. An allocation in which such a bailout occurs might, or might not, be optimal. Proposition 5, to be proved in Section 8, states that there is an economy in which a bailout occurs in one of the optimal allocations. Proposition 6, to be proved in Section 9, states that, when the model is modified by positing a convex deadweight cost of taxation, there is an economy in which a bailout must occur in the unique optimal allocation.

The relationship between the formal definition of a bailout provided here and the informal definition stated in the introduction deserves comment. The informal definition refers to subsidy for the purpose of preventing or mitigating insolvency. The formal definition refers to a correlation between subsidy and early liquidation of investments. The idea that links the two definitions is that early liquidation is a drastic measure that must be taken to avert or minimize insolvency under laissez faire, and that it has adverse effects prima facie. Especially in the case that a subsidy would be completely successful in averting insolvency without having recourse to early liquidation, there could not be any correlation of the subsidy with early liquidation. Thus, a subsidy that fits the intuitive definition of subsidy would not fit the formal definition. Conversely, if a firm can meet its obligations by means of early liquidation, and if a subsidy is provided when early liquidation is used for that purpose, then—providing that the firm could be forced to liquidate rather than having to be bribed with the subsidy to do so—the subsidy is not necessary to avert insolvency. That is, the subsidy would fit the formal definition of a bailout but not the informal definition. Nevertheless, although some adjustment of both the informal and formal definitions of a bailout to reconcile their meanings would be desirable in principle, the formal definition succeeds well in capturing the intent of the informal definition in the examples to be studied below. The issue of definitional fit here is typical, not exceptional. Formal concepts introduced in scientific theories seldom match exactly the informal concepts that they supplant.

7. INCENTIVE COMPATIBILITY

Since each investor's information about his own state is private, investors must be willing to report it truthfully in order for an aggregate production plan and the allocation that distributes its output to distinguish between B and Γ (cf. Myerson 1979). That is, evaluated according to the conditional expected utility of the investor's true type, what the allocation gives to that type is better than what it gives to the opposite type (that is, than what he would get if he were to report his type falsely). To formalize this idea, let $\tilde{c}$ be the allocation that, at each date and in each aggregate state, gives an impatient investor what

c gives a patient investor and vice versa. That is, for all t and σ, and for each $s \in \{1, 2\}$, $\tilde{c}_t^{s\sigma} = c_t^{(3-s)\sigma}$.[10] Allocation c is incentive compatible if, for each $s \in \{1, 2\}$,[11]

$$U_s(c) \geq U_s(\tilde{c}). \tag{29}$$

Note that, according to the terminology adopted in this article, *feasible* means technically feasible, and does not imply incentive compatibility. *Optimal* means maximal with respect to expected utility among feasible production plans, rather than among plans that are both technically feasible and incentive compatible. However, an allocation cannot actually be implemented in a private-information environment unless it is both technically feasible and incentive compatible. The reason is that, unless the allocation is incentive compatible, investors will not voluntarily make the state-contingent choices that are required to implement it, and those choices must be made voluntarily because they depend on contingencies that would have to be—but cannot be—observed by some third party in order to be enforced coercively. The remainder of the article will be a study of the questions: Can an optimal allocation be incentive compatible and, if so, must a subsidy or even a bailout be provided in some state of the world in order to satisfy the incentive-compatibility constraint (29)?

8. OPTIMALITY, INCENTIVE COMPATIBILITY, AND SUBSIDY

Here is an example of an economy in which there is an allocation that is both optimal for cooperative production (as in Proposition 4) and incentive compatible, and in which every allocation that satisfies both of these conditions exhibits subsidy. Consider the following parameter values:

$$\bar{x}_0 = 3; \bar{x}_2 = 1; \beta = 0.06; \mu_B = \frac{5}{6}; \mu_\Gamma = \frac{1}{2}; \theta = 4; R = 2; r = \frac{1}{2}; \eta = 5. \tag{30}$$

[10] It would not necessarily be technically feasible to give the consumption specified by $\tilde{c}$ to every investor. The point of defining $\tilde{c}$ is to specify what an investor would get by deviating unilaterally from truthful revelation. Such a unilateral deviation, by an investor whose individual consumption is infinitesimal compared to aggregate consumption, would not cause incentive compatibility to be violated.

[11] To spell out condition (29),

$$\begin{aligned}
\beta_1^* u\left(c_1^{1B}, c_2^{1B}, 1\right) + \left(1-\beta_1^*\right) u\left(c_1^{1\Gamma}, c_2^{1\Gamma}, 1\right) &\geq \beta_1^* u\left(c_1^{2B}, c_2^{2B}, 1\right) + \left(1-\beta_1^*\right) u\left(c_1^{2\Gamma}, c_2^{2\Gamma}, 1\right); \\
\beta_2^* u\left(c_1^{2B}, c_2^{2B}, 2\right) + \left(1-\beta_2^*\right) u\left(c_1^{2\Gamma}, c_2^{2\Gamma}, 2\right) &\geq \beta_2^* u\left(c_1^{1B}, c_2^{1B}, 2\right) + \left(1-\beta_2^*\right) u\left(c_1^{1\Gamma}, c_2^{1\Gamma}, 2\right).
\end{aligned}$$

These values satisfy (23) and (26). By Proposition 4, the optimal cooperative production plan is $(\iota, \epsilon_B, \epsilon_\Gamma) = (1, 1, 0)$, and an optimal allocation, c, of the product, y, of this plan, if it does not exhibit subsidy, must satisfy

$$\begin{aligned} c_1^{1B} &= y_{B1}/\mu_B = 3; \\ c_1^{1\Gamma} &= \theta = 4; \\ c_1^{2B} &= c_1^{2\Gamma} = 0; \\ \mu_\Gamma c_2^{1\Gamma} + (1-\mu_\Gamma)\, c_2^{2\Gamma} &= \bar{x}_2 + y_{\Gamma 2} = 3; \\ c_2^{sB} &= 1 \text{ for } s \in \{1, 2\}; \\ c_2^{s\Gamma} &\geq 1 \text{ for } s \in \{1, 2\}. \end{aligned} \tag{31}$$

For the allocation to be incentive compatible for a patient investor, according to (29), $U_2(c) - U_2(\tilde{c}) \geq 0$ must be satisfied. $U_2(c) - U_2(\tilde{c})$ is increasing in $c_2^{2\Gamma}$ and decreasing in $c_2^{1\Gamma}$. In an optimal allocation, by (31), a patient investor can only consume at date 2 and must consume exactly one unit at that date in aggregate state B. Therefore, if any optimal allocation can be incentive compatible for the patient investor but not exhibit subsidy, then one such allocation will give all date-2 output to patient investors. That is, the following allocation should be checked for incentive compatibility:

$$\begin{aligned} c_1^{1B} &= y_{B1}/\mu_B = 3; \\ c_1^{1\Gamma} &= 4; \\ c_1^{2B} &= c_1^{2\Gamma} = 0; \\ c_2^{1B} &= c_2^{1\Gamma} = 1; \\ c_2^{2B} &= 1; \\ c_2^{2\Gamma} &= R\iota/\mu_\Gamma + 1 = 5. \end{aligned} \tag{32}$$

But this allocation is obviously not incentive compatible. A patient investor's utility function is $u(c_1, c_2, 2) = c_1 + c_2$, and this quantity is identical for a patient and an impatient investor in aggregate state Γ and strictly higher for an impatient investor in state B. The consequence for condition (32) is that

$$U_2(c) - U_2(\tilde{c}) = -3\beta_2^* \approx -.063. \tag{33}$$

Taxing impatient investors' endowments at date 2, and transferring the tax revenue to patient investors, converts c to a new allocation that is equal to c with regard to ex ante expected utility, and that is incentive compatible. Specifically, define allocation d by taxing one unit of impatient investors' endowment in state Γ and transferring it to patient investor. That is,

$$\begin{aligned} d_t^{s\sigma} &= c_t^{s\sigma} \text{ if } t = 1 \text{ or } \sigma = B; \\ d_2^{1\Gamma} &= 0; \\ d_2^{2\Gamma} &= 6. \end{aligned} \tag{34}$$

It is obvious from (32) that, when the date-2 consumption of patient investors in c is increased by two relative to impatient investors with probability close to one, the resulting allocation, d, is incentive compatible for patient investors. To be precise, the incentive-compatibility constraint (29) for patient investors evaluates (after rounding) to $5.9 > 4.0$, so the constraint is satisfied. For impatient investors, (29) evaluates to $6.9 > 5.5$, so their incentive-compatibility constraint is also satisfied.

There is no bailout in allocation d, however, because liquidation occurs in one state but a tax is levied (and subsidy is distributed) only in the other. Consider optimal allocation e, in which taxation occurs in the bad state along with liquidation. That is, a bailout occurs in this allocation:

$$\begin{aligned} e_t^{s\sigma} &= c_t^{s\sigma} \text{ if } t = 1 \text{ or } \sigma = \Gamma; \\ e_2^{1B} &= 0; \\ e_2^{2B} &= 6. \end{aligned} \tag{35}$$

The incentive-compatibility condition (29) evaluates to $5.02 > 4.96$ for patient investors and to $7.8 > 5.1$ for impatient investors. The following proposition summarizes these findings.

Proposition 5 *Every technically feasible allocation of the economy described by (30) either is suboptimal, violates incentive compatibility, or exhibits subsidy. The economy has some allocations that are technically feasible, optimal, and incentive compatible. All such allocations exhibit subsidy. In some of them, a bailout occurs.*

9. ESSENTIAL BAILOUTS

Bailouts are defined as essential in an economy if one occurs in every allocation of that economy that is optimal subject to incentive compatibility constraints. In this section, the model of an economy is modified in such a way that there is an example in which bailouts are essential.

One way to make such a change would be, in effect, to gerrymander the model. We specify that the marginal utility of consumption for impatient investors at date 2 is lower in aggregate state B, but higher in state Γ, than that for patient investors. We also specify that each investor's date-2 endowment is state contingent, and is perfectly correlated with the investor's preference type, specifically with an investor having a larger endowment when impatient

than when patient. Optimality subject to incentive compatibility would then require the transfer from impatient to patient investors to be maximized at date 2 in B, and it would require the transfer from patient to impatient agents at date 2 in Γ to be maximized, subject to incentive compatibility. If early liquidation is required in B to achieve the optimal level of ex-ante utility, subject only to technical feasibility, and if incentive-compatibility constraints do not bind in that allocation, then bailouts are essential in the economy.

This sketch of an example shows that, in principle, either the definition of a bailout or the definition of essentiality needs to be tightened. That is, the two definitions together should express the idea that subsidy is being used in the bailout state to solve an incentive problem created intrinsically by early liquidation, rather than playing a distinct role having to do with insurance. Since all investors' utility for consumption at date 2 is assumed to be linear and identical across individual states (implying that there is no possibility of increasing ex-ante welfare by equalizing different investors' marginal utilities at date 2) in the example studied in Section 8, the current definitions seem satisfactory, as long as that assumption is maintained.

Consider an alternative modification of the model: the introduction of a convex deadweight cost of taxation. Let δ be a convex function satisfying $\delta(\tau) = 0$ for all $\tau \leq 0$, and δ is strictly convex at positive tax levels. This function specifies, for each investor, how much consumption is lost to the economy when tax is collected from him.[12] To formalize this idea, replace the definition (13) of technical feasibility for cooperative production with

$$\begin{aligned} \mu_\sigma c_1^{1\sigma} + (1-\mu_\sigma)\, c_1^{2\sigma} &\leq y_{\sigma 1}; \\ \mu_\sigma c_2^{1\sigma} + (1-\mu_\sigma)\, c_2^{2\sigma} &\leq y_{\sigma 2} + \bar{x}_2 - \left(\mu_\sigma \delta\left(\tau^{1\sigma}\right) + (1-\mu_\sigma)\,\delta\left(\tau^{2\sigma}\right)\right). \end{aligned} \tag{36}$$

A calculus result, Jensen's inequality, implies the following lemma.

Lemma 1 *If δ is strictly convex for $\tau > 0$, and if technical feasibility of an allocation for aggregate production is defined by (36), then $\tau^{1B} = \tau^{1\Gamma}$ and $\tau^{2B} = \tau^{2\Gamma}$ in an allocation that is optimal subject to technical feasibility and incentive compatibility. By strict convexity of δ, this constrained-efficient allocation is generically unique.*[13]

Using this lemma, it is routine to calculate an allocation f, analogous to e in Section 8, that is optimal among technically feasible, incentive-compatible

[12] The cost includes the direct cost of collecting and enforcing taxes and the indirect cost (in an actual economy, as opposed to the highly simplified model economy) of agents shifting resources to low-productivity, but tax-favored, investments. Embedding a costly state verification model of tax collection (along the lines of Townsend [1979]) in the model economy would provide a foundation for this reduced-form specification.

[13] *Generically* means that, for any parameter vector having more than one such allocation, the economy corresponding to an arbitrarily small perturbation of that vector in a random direction will have a unique optimum.

allocations of the economy that is identical to the one studied in Section 8 (with parameters specified in [30]), except that technical feasibility is defined according to (36). By the lemma, this can be taken to be the unique constrained-efficient allocation of the economy. In the allocation, $\tau^{1\sigma} > 0 = \tau^{2\sigma}$. By the lemma, the tax does not depend on σ. Thus, let $\hat{\tau}$ denote the tax levied on impatient investors in both states. The amount of tax levied is $\hat{\tau}/2$ in state Γ and $5\hat{\tau}/6$ in state B. This means that B is the high-subsidy state, as well as being the early-liquidation state, so there is a bailout. Since the allocation in question is the unique constrained-efficient allocation, bailouts are essential in this economy.

To carry out the details of this construction, let ζ be a small, positive number, and define

$$\delta(\tau) = \begin{cases} 0 & \text{if } \tau < 0; \\ \zeta\tau^2 & \text{if } \tau \geq 0. \end{cases} \tag{37}$$

Consider the economy with parameters specified in (30) and with the set of feasible allocations specified to incorporate a deadweight cost of taxation according to (36) and (37). Modify allocation c, defined in (32), to specify a feasible allocation f of this economy, defined in terms of a positive parameter $\hat{\tau}$, as follows:

$$\begin{aligned} f_1^{s\sigma} &= c_1^{s\sigma}; \\ f_2^{1\sigma} &= \bar{x}_2 - \left(\hat{\tau} + \delta\left(\hat{\tau}\right)\right); \\ f_2^{2B} &= \bar{x}_2 + 5\hat{\tau}; \\ f_2^{2\Gamma} &= \bar{x}_2 + 2R\iota + \hat{\tau}. \end{aligned} \tag{38}$$

That is, set f equal to c at date 1, set the consumption level of an impatient investor at date 2 to be the investor's date-2 endowment minus the sum of a tax $\hat{\tau}$ and the deadweight cost of its imposition, and set the consumption level of a patient investor at date 2 to be the sum of the investor's endowment and the investor's share of both the date-2 investment proceeds from plan $(1, 1, 0)$ and the receipt from the taxation of impatient investors.

If $\zeta = 0$ and $\hat{\tau} < \bar{x}_2$, then allocation f is optimal. If $\zeta > 0$ and $\hat{\tau} > 0$, then f is not optimal because $\delta(\hat{\tau}) > 0$, and this deadweight cost must be deducted from consumption. However, for the parameter values specified in (30), a subsidy is necessary to achieve incentive compatibility, and logically this is true under an assumption that $\zeta > 0$, since the set of feasible allocations for positive ζ is a subset of those for $\zeta = 0$. Optimality subject to incentive compatibility is achieved when the tax, $\hat{\tau}$, is minimized, subject to the constraint that the resulting allocation should be incentive compatible. That value of $\hat{\tau}$ is the one that makes the incentive-compatibility constraint for patient investors hold with equality, that is,

$$\begin{aligned} 0 &= U_2(f) - U_2\left(\tilde{f}\right) \\ &= \beta_2^* \left[\left(f_1^{2B} - f_1^{1B}\right) + \left(6\hat{\tau} + \delta(\hat{\tau})\right)\right] \\ &\quad + \left(1 - \beta_2^*\right)\left[\left(f_1^{2\Gamma} - f_1^{1\Gamma}\right) + \left(2R\iota + 2\hat{\tau} + \delta(\hat{\tau})\right)\right] \\ &= \zeta\hat{\tau}^2 + \left(2 + 4\beta_2^*\right)\hat{\tau} - 4\beta_2^*. \end{aligned} \tag{39}$$

By the quadratic formula and the positivity of $\hat{\tau}$,

$$\hat{\tau} = \frac{-\left(2 + 4\beta_2^*\right) + \sqrt{\left(2 + 4\beta_2^*\right)^2 + 12\zeta\beta_2^*}}{2\zeta}. \tag{40}$$

Using Taylor's formula to approximate the square-root term in the numerator of (40),

$$\hat{\tau} = \frac{4\beta_2^*}{2 + 4\beta_2^*} = 0.03. \tag{41}$$

It can easily be computed that, when the tax is set at this level, the incentive-compatibility constraint for impatient investors does not bind. Thus, allocation f is the unique allocation that is technically feasible and is also optimal subject to the incentive-compatibility constraints for both patient and impatient investors. In f, since $\hat{\tau}$ is collected from $\frac{5}{6}$ of the investors in state B but only from $\frac{1}{2}$ of them in state Γ, aggregate tax revenue in B is $\frac{5}{3}$ times aggregate tax revenue in Γ. That is, given that early liquidation occurs in B but not in Γ, allocation f exhibits a bailout in B, and this bailout is essential. The following proposition summarizes this result.

Proposition 6 *Under the assumption that taxing an investor has convex deadweight cost, there is an economy in which bailouts are essential.*

10. CONCLUSION

Occasional bailouts of insolvent firms that are ultimately financed by taxation—notably including bailouts of financial intermediaries—are a fact of life in virtually every country. On one side of a debate about the welfare assessment of such bailouts are economists, such as Kareken (1983) and Stern and Feldman (2004), who emphasize that inefficient risk-taking results from a combination of time inconsistency on the part of the government and moral hazard on the part of firms' owners, liability holders, and managers. On the other side, there has been only an amorphous plea, albeit a sincere one from some distinguished economists and sophisticated policymakers and financial-market participants, that unspecified but very serious and long-term harms would result if government were to refrain from a bailout. At first sight, such a plea seems to be

a reflection of precisely the time inconsistency that is pivotal to the critics' arguments. However, there is another possible interpretation of the point that apologists for bailouts are trying to make. Namely, once a regime has been established that favors the incorporation of limited-liability firms, bailing out those firms in some states of the world may be the only way to make ex-ante efficient investments incentive compatible. While critics believe that it would be time inconsistent to conduct a bailout, apologists believe that it would be time inconsistent to refrain from a bailout in some circumstances. The long-term harm that they fear is impairment, after an ex-ante commitment to incentive-enhancing bailouts had been shown not to be credible, of investors' willingness to fund socially beneficial projects. This paper, particularly in Proposition 6, develops the logic of that position.

It should be kept in mind that this article has explored the logic of an economic argument, rather than having advocated a policy. Issues of first-rank importance in an actual economy, such as the effect that anticipating a bailout to be available will have on firm owners' and managers' incentive to take risk, do not arise in the model economy studied here. Nevertheless, this analysis shows that public discussion regarding the bailout of firms by the U.S. government during the financial crisis in 2008–2009 has had shortcomings. It has generally been asserted by critics of the bailout, and conceded by its proponents, that a tax-financed subsidy to firms is ex ante a bad policy. This assertion is not sound with respect to the model economy analyzed here. It may well be sound with respect to the U.S. economy, but that judgment should be given a supporting argument rather than taken as a starting point. A tradeoff has to be made between the potential benefits of a bailout emphasized in the present model and the costs that are emphasized in other models. It is an oversimplification to presume that a bailout is necessarily all bad.

REFERENCES

Acharya, Viral V., Douglas M. Gale, and Tanju Yorulmazer. 2009. "Rollover Risk and Market Freezes." NYU Stern School of Business Technical Report FIN-08-030.

Bagehot, Walter. 1877. *Lombard Street: A Description of the Money Market*. New York: Scribner, Armstrong & Co.

Bryant, John. 1980. "A Model of Reserves, Bank Runs, and Deposit Insurance." *Journal of Banking and Finance* 4 (December): 335–44.

Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 91 (June): 401–419.

Easterbrook, Frank H., and Daniel R. Fischel. 1985. "Limited Liability and the Corporation." *The University of Chicago Law Review* 52 (Winter): 89–117.

Kareken, John H. 1983. "Deposit Insurance Reform; Or, Deregulation is the Cart, not the Horse." Federal Reserve Bank of Minneapolis *Quarterly Review* 1–9.

Macey, Jonathan R., and Geoffrey P. Miller. 1992. "Double Liability of Bank Shareholders: History and Implications." *Wake Forest Law Review* 27: 31–62.

Myerson, Roger B. 1979. "Incentive Compatibility and the Bargaining Problem." *Econometrica* 47 (January): 61–73.

Stern, Gary H., and Ron J. Feldman. 2004. *Too Big to Fail: The Hazards of Bank Bailouts*. Washington, D.C.: Brookings Institution Press.

Stiglitz, Joseph E. 1969. "A Re-Examination of the Modigliani-Miller Theorem." *American Economic Review* 59 (December): 784–93.

Thornton, Henry. 1802. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. London: J. Hatchard and F. and C. Rivington.

Townsend, Robert M. 1979. "Optimal Contracts and Competitive Markets with Costly State Verification." *Journal of Economic Theory* 21 (October): 265–93.

Economic Quarterly—Volume 96, Number 1—First Quarter 2010—Pages 33–58

On the Fundamental Reasons for Bank Fragility

Huberto M. Ennis and Todd Keister

Over the course of the recent financial crisis, several large financial institutions experienced sudden, massive withdrawals of their usual funding sources. In the U.K., for example, depositors lost confidence in the bank Northern Rock and started a *run* of withdrawals that ended with the bank being taken into state ownership. In the United States, the investment bank Bear Stearns and the commercial bank Wachovia both experienced a rapid loss of funding and were taken over by other institutions to avoid their outright failure. This same phenomenon affected other types of institutions as well, including a large part of the money market mutual fund industry, which experienced heavy withdrawals following the failure of the Reserve Fund in September 2008.

These episodes are only the most recent examples of a phenomenon that has been a recurrent theme in the history of banking. Banking panics, with massive withdrawals often leading to widespread bank failures, were a regular occurrence in the United States prior to the advent of government-sponsored deposit insurance in 1933. Developing economies have also experienced runs on their banking system, including episodes in Ecuador (1999), Argentina (2001), and Russia (2004).

Observers of these episodes often claim that there is an important self-fulfilling component to the behavior of depositors and/or investors. In this view, each depositor fears that the withdrawals of *other* depositors will cause the bank to fail and rushes to withdraw her funds before this failure occurs. Collectively, these actions validate the original belief that a wave of withdrawals will cause the bank to fail. During the height of the Panic of 1907 in the United States, J.P. Morgan was reported in the *New York Times* to have

We would like to thank Borys Grochulski, Ned Prescott, and Juan Sánchez for comments on a previous draft. The views expressed here do not necessarily represent those of the Federal Reserve Bank of New York, the Federal Reserve Bank of Richmond, or the Federal Reserve System. E-mails: huberto.ennis@rich.frb.org; todd.keister@ny.frb.org.

said, "If the people would only leave their money in the banks instead of withdrawing it...everything would work out all right."[1] In other words, Morgan claimed that it was the behavior of the depositors themselves that was placing the largest strain on the banking system. If this strain were removed, individuals would be willing to leave their money deposited and a superior outcome would obtain.

This view of events implies that banks and other financial intermediaries are inherently *fragile*, in the sense of being susceptible to a self-fulfilling run by their depositors. The degree to which one accepts this view has strong implications for public policy. The desirability of government-provided deposit insurance, for example, and of other public interventions in the banking system depends in large part on whether banking crises do indeed have an important self-fulfilling component or whether they instead result from other, more fundamental causes.

A substantial economic literature has developed that attempts to identify the essential components that would justify a self-fulfilling interpretation of events. Bryant (1980) and Diamond and Dybvig (1983) provided the first steps in the development of a coherent theory along these lines. Subsequently, various authors have tried to understand if the set of elements included in these early contributions is sufficient to explain banking and the fragility of banks and other financial intermediaries, and which other elements, if any, may be missing.

The approach taken in this literature has been to specify a complete physical environment and to study economic outcomes that agents in such an environment could achieve without imposing any artificial restrictions on their ability to enter mutually beneficial arrangements. In following this approach, the literature has become fairly technical and intricate. In this article, we aim to provide an informal discussion of the issues and the results produced so far in this literature. We hope that our endeavor will make the lessons obtained from this body of work more readily accessible to readers who may be less inclined to endure over the many technical issues involved in the subject.

We begin our discussion by reviewing the key theoretical contribution of the seminal work by Diamond and Dybvig (1983). We discuss the basic elements of their banking theory and how subsequent researchers have addressed the technical difficulties involved in designing an equilibrium concept that allows for the possibility of a bank run. As will become clear in the discussion, one essential element of the theory is the existence of a first-come, first-served (or *sequential service*) constraint. In Section 2, we discuss how the literature has handled the specification of an explicit sequential service constraint. Several important recent contributions in this literature have resulted from the

[1] *New York Times*, October 26, 1907, "Bankers Calm; Sky Clearing."

efforts to combine explicitly modeled sequential service with the presence of aggregate uncertainty about the fundamental need for liquidity in the system. We review those contributions and how they relate to each other in detail. In Section 3 we discuss some potentially fruitful directions for further research and, finally, we close the article with some brief concluding remarks.

1. THE DIAMOND-DYBVIG MODEL

This section presents an overview of the seminal contribution by Diamond and Dybvig (1983) and sets the stage for the discussion of the more recent literature that explores the fundamental reasons for bank fragility. In Diamond and Dybvig's theory, banks play an essential role in the process of *maturity transformation*: they issue short-term (deposit) liabilities in order to finance long-term productive investment. While maturity transformation may happen through other channels in the economy, Diamond and Dybvig identify two other essential features of banking arrangements: the fact that agents' demands must be dealt with on a *first-come, first-served* basis, and the fact that agents' true liquidity needs remain *private information*. These three elements constitute the foundations of Diamond and Dybvig's theory of banking and are also the source for the potential of bank fragility in their model.

The Physical Environment

Diamond and Dybvig (1983) consider an environment where a large number of agents face idiosyncratic uncertainty about their intertemporal desire to consume. Agents have an initial endowment of goods and there is a technology that can be used to transform these goods into (potentially more) goods in the future. If investment is left in place long enough to mature, the net returns are positive. However, some agents will discover that they are *impatient* and need to consume before the investment matures. Other agents are *patient* and able to consume after investment has matured.

Investment takes place before agents discover their intertemporal preference for consumption. To the extent that the idiosyncratic desire to consume early is not perfectly correlated among agents, there are insurance possibilities to be exploited in this environment. In particular, there exists a clear social benefit from pooling resources ex ante, before preferences are realized, investing in the long-term technology, and then making payments ex post to agents, contingent on their needs.

Diamond and Dybvig (1983) assume that an agent's realized preference type (patient or impatient) is private information. Any attempt to provide consumption to agents in a way that depends on their intertemporal preference for consumption must, therefore, rely on reports from agents. This fact could complicate matters in two ways. First, the ex-post payments to agents must

be arranged in such a way as to create the right incentives for each individual agent to not misrepresent her consumption needs. Second, private information opens the door to the possibility of a coordinated misrepresentation by agents, which may be interpreted as a *run* to withdraw from the pool. The insurance possibilities associated with a pooling arrangement depend crucially on its ability to avoid these two types of misrepresentation.

In principle, it would be beneficial to collect as much information as possible about the total demand for withdrawals before making any payments from the resource pool. However, Diamond and Dybvig (1983) assume that agents who decide to withdraw early place their demands sequentially, and that payments from the pool must be made at the time each demand is placed. In other words, payments ought to respect a first-come, first-served rule, which they call a *sequential service constraint*. Diamond and Dybvig argue that this kind of restriction is a realistic description of how banks operate.[2]

Resource Allocation and Optimality

Diamond and Dybvig's simple environment provides a natural setup to think about the institution of banking. In the model, agents initially deposit their endowments in a pool, which can be interpreted as a "bank." In exchange for her deposit, an agent receives a claim to future consumption from this bank. After deposits are made, the bank invests in the long-term technology. Finally, agents discover their consumption needs and contact the bank sequentially to withdraw resources and consume. The bank makes payments to agents, on demand, in a pre-arranged manner.

From a theoretical point of view, it is appealing to abstract from institutional details and focus instead on allocations of consumption that are achievable while respecting the constraints imposed by the physical environment and the structure of information. Much of what is done in Diamond and Dybvig's (1983) article is consistent with this strategy. Following the basic principles in the theory of mechanism design, the way to proceed is to set up a planning problem that consists of choosing a (contingent) consumption allocation to maximize the ex ante expected utility of agents subject to incentive compatibility, sequential service, and resource feasibility constraints.[3] We will call this allocation the *constrained-efficient allocation*.

---

[2] An important component of a formal sequential service constraint is the specification of whether or not agents who decide to not withdraw early still contact the pool at that time. Diamond and Dybvig (1983) implicitly assume that only agents who are attempting to withdraw contact the pool. We return to this issue later in this article.

[3] Going back to the interpretation of the theoretical constructions in terms of the institutions of banking, it can be demonstrated that under certain conditions the solution to this planning problem is equivalent to the outcome that would obtain when profit-maximizing banks compete for deposits.

To understand the implications of agents possibly misrepresenting their consumption needs, it is useful to solve the same planning problem, but without imposing the incentive constraints. We will call the solution to this modified problem the *unconstrained-efficient allocation*.[4]

In general, the incentive compatibility constraint for an individual agent in this environment depends on the assumed behavior of the rest of the agents. While it may be incentive compatible for an agent to not misrepresent her consumption needs when all the other agents are also not misrepresenting, the situation may be different when the other agents are expected to misrepresent. This payoff complementarity is important because it creates the potential for strategic coordinated responses that may result in substantial inefficiencies.

The strategic interaction among agents takes place in the *withdrawal game* induced by a given contingent consumption allocation, i.e., a complete payment scheme. In the withdrawal game, agents decide when to contact the resource pool (the bank) to demand payment. An allocation is *implementable* (under truthful representation) if there is a Nash equilibrium of the induced withdrawal game in which all impatient agents withdraw early and all patient agents wait until the investment matures. An implementable allocation is often also called *incentive feasible*, in the sense that it satisfies the incentive compatibility constraint for each individual agent given that all the other agents are not misrepresenting their consumption needs. If the equilibrium of the withdrawal game is unique, we say that the allocation is *strongly* (or *fully*) *implementable*. As we will see, implementable allocations in the Diamond-Dybvig model are sometimes not strongly implementable. In those cases, there exists another Nash equilibrium of the withdrawal game in which some patient agents misrepresent their need to consume and attempt to withdraw early, in effect running to obtain payment from the pool before its resources are exhausted.

Diamond and Dybvig (1983) make some additional simplifying assumptions that turn out to have significant implications for their results. In particular, they assume that there is a continuum of agents in the economy and that preference types (patient or impatient) are independent and identically distributed (i.i.d.) across agents. The combination of these two assumptions and the law of large numbers implies that the total need for early consumption is completely predictable. In other words, if the bank believes that only impatient agents will withdraw before investment matures, then it knows the total demand for liquidity even before agents begin placing their requests.

[4] The unconstrained-efficient allocation is the best allocation that can be attained when preferences of agents are observable. Since we consider the sequential service constraint a reflection of a feature of the physical environment (Wallace 1988), the unconstrained-efficient allocation must satisfy sequential service in the same way that it must satisfy resource feasibility.

Diamond and Dybvig (1983) show that the unconstrained-efficient allocation is actually implementable in their environment.[5] Hence, the constrained-efficient allocation is equal to the unconstrained-efficient allocation, and the fact that agents' preferences are private information imposes no restrictions in terms of what is implementable in this environment (i.e., the incentive constraints in the planning problem are not binding at the solution). Furthermore, the fact that agents withdraw from the resource pool sequentially has no implications for the choice of the constrained-efficient allocation. In other words, the sequential service constraint is also nonbinding at the solution to the planning problem.

Under certain conditions on the relative risk aversion of agents, Diamond and Dybvig also show that in the unconstrained-efficient allocation, agents withdrawing early receive more than what they initially deposit at the bank. In other words, the best allocation provides some degree of insurance against the contingency that the agent becomes impatient and cannot wait for the investment to mature. This finding is important to understand the fundamental reasons for the possibility of bank fragility in the model.

Deposit Contracts and the Possibility of Runs

Interestingly, there are many possible payment schemes that can be used to implement the unconstrained-efficient allocation. One such scheme specifies that each agent, after depositing her resources in the pool, is entitled to a fixed payment if she withdraws early and a different fixed payment if she withdraws late. This arrangement resembles a simple demand deposit contract, commonly used in practice, in which agents experience a penalty for withdrawing early but their payment is otherwise not contingent on information that the bank might receive as (sequential) withdrawals occur. We call this scheme the *optimal simple demand deposit contract*.

This demand deposit contract must respect important restrictions imposed by the physical description of the environment. First, it must obviously conform with resource feasibility. This unavoidable constraint implies that payments will be fixed only as long as the bank does not run out of resources. Second, the contract must be consistent with the assumption that agents have private information about their own preferences. In particular, payments cannot be made contingent on the true preference of agents, since these are unknown to the bank.

When only impatient agents withdraw early, the bank does not run out of resources when following this contract and the payments generate the

[5] Making early payments is costly for the pool since it removes resources from investment before it has had time to mature. For this reason, it is always optimal to give agents who are withdrawing late at least as much utility as those withdrawing early and, hence, the unconstrained-efficient allocation always satisfies the incentive constraints.

unconstrained-efficient allocation. However, the demand deposit contract does not strongly implement the unconstrained-efficient allocation. There is another equilibrium of this withdrawal game in which all agents attempt to withdraw early and the bank runs out of resources before paying some agents.[6] This equilibrium resembles a self-fulfilling run on the bank.

What is the logic behind this run equilibrium? Agents have, sequentially, an opportunity to withdraw early from the bank. Those agents who become impatient have no real decision to make: they place a demand to withdraw when their turn comes. Patient agents, on the other hand, need to decide whether to try to withdraw early or wait until investment matures. If all patient agents expect that all other patient agents will try to withdraw early, then they also expect that the bank will run out of resources before all agents have been paid. Waiting until investment matures in such circumstances is pointless, since the bank's resources will be depleted before then. Hence, all patient agents attempt to withdraw early, fulfilling their beliefs and making this outcome consistent with equilibrium.

The possibility of this type of self-fulfilling run is a direct consequence of the presence of the sequential service constraint. Without sequential service, the bank could wait until all agents have placed their withdrawal requests before making any payments. Since the bank knows the number of impatient agents in the population, once requests pass this threshold the bank would be able to clearly identify that a run is taking place. Importantly, the bank would then know about the run before making any payments to agents. It is not hard to see that, once a run has been identified, the payment scheme in the simple demand deposit contract is no longer optimal. Because agents are risk averse and everyone is attempting to withdraw, the best way to allocate existing resources is to distribute them evenly among agents. In this case, however, patient agents would actually prefer not to participate in the run. By waiting and leaving their funds in the bank, patient agents will be able to receive a higher payment after the investment matures. In summary, the lack of sequential service would be sufficient to rule out runs as possible equilibrium phenomena.

Another critical assumption in the run situation described above is that only those agents who intend to withdraw are expected to contact the bank. If this were not the case, then it would be easy for the bank to realize that a run is taking place before any significant portion of agents have attempted to withdraw. In general, when no run is taking place, the bank would expect withdrawal demands to be scattered among nonwithdrawal demands. If every agent contacting the bank places a demand for withdrawal, the bank can quickly infer that a run is taking place. In the case of the continuum of

[6] This is a consequence of the provision of insurance in the unconstrained-efficient (which, in this case, is also equal to the constrained-efficient) allocation.

agents, this logic is extreme and the run could be identified before any significant payments have been made.[7] In a sense, with a continuum of agents the sequential service constraint is only relevant when not all agents contact the bank in the early period. When there is a finite number of agents, however, things are different. As we will see in the next section, the sequential service constraint can be meaningfully specified in either way in this case, with differing implications for bank fragility.

An unsettling characteristic of the run situation under the optimal simple demand deposit contract is that once the number of withdrawals surpasses the number of impatient agents in the population, which is nonstochastic, the bank is certain that a run is underway. This information could potentially be used to design a more robust payment scheme. In fact, there is a payment scheme that strongly implements the unconstrained-efficient allocation by modifying only payments that will then lie off the equilibrium path of play and, hence, never be made. This payment scheme involves a suspension of convertibility clause, which says that after a certain number of withdrawals the bank will suspend payments and wait until investment has matured. If this suspension is designed to take place only after the number of withdrawals is larger than the number of impatient agents in the population, but not too much after that, then it will never occur in equilibrium; the expectation that it would occur if needed is sufficient to rule out a possible run on the bank.

In summary, even when payments are required to be made sequentially, there is a scheme involving an (off-equilibrium) suspension of convertibility that rules out runs and strongly implements the unconstrained-efficient allocation. In that sense, the presence of sequential service does not change the configuration of equilibrium outcomes in the benchmark version of the Diamond-Dybvig model.

One crucial feature that allows the suspension of payments to work so effectively, without any cost, is the absence of aggregate uncertainty about the total number of impatient agents in the economy. In other words, the model described above has no uncertainty about the total *fundamental* need for early liquidity. If the bank were unsure about the true aggregate need for early liquidity, it would be much more difficult to choose the right time to suspend payments. Suspending too soon may leave some impatient agents without precious resources at the time that they truly need to consume. Suspending too late may leave resources sufficiently depleted to make the run consistent with equilibrium. Diamond and Dybvig (1983) recognize this important limitation in their analysis and give some preliminary steps in the direction of relaxing the assumption of no aggregate uncertainty.

[7] De Nicoló (1996) exploits this idea to design a contract that strongly implements an allocation arbitrarily close to the constrained-efficient allocation.

The presence of aggregate uncertainty, together with sequential service, significantly complicates the analysis. Not only is solving and characterizing the unconstrained-efficient allocation a much more complex problem, but studying the strategic interaction among agents also involves more sophisticated techniques and logic. Diamond and Dybvig (1983) only hint at these issues in their seminal analysis. They abstract from incentive compatibility and sequential service constraints to solve for a benchmark allocation under aggregate uncertainty.[8] They then demonstrate that this benchmark allocation is not implementable under private information and sequential service. Whether the unconstrained-efficient allocation (which takes into account sequential service) could be implemented and/or strongly implemented in the presence of aggregate uncertainty was left as an open question in the literature for a long time. Only 20 years later was the first detailed analysis of this question in the Diamond-Dybvig framework provided by Green and Lin (2003). We discuss their contribution in Section 2.

Runs and the Equilibrium Concept

Before we conclude our discussion of Diamond and Dybvig's (1983) initial contribution, it is worth mentioning some important issues related to the formal treatment of bank fragility that originated in their work.[9] It is easy to see that in the Diamond-Dybvig model a bank run can happen only if the agents and the bank are not certain ex ante that one will occur. If the bank is certain that a run will happen, it will make payments to agents without providing insurance, which makes the run strategy of agents inconsistent with equilibrium. If the bank believes that a run will not occur, but the agents are certain that one will, then agents will not choose to deposit their resources at the bank. Hence, runs can occur in equilibrium only if they are expected to happen with some probability strictly less than unity. Formally, this kind of uncertainty can be captured by introducing an *extrinsic random variable* in the model, which allows agents to condition their behavior on the realization of such a variable. This modeling strategy was suggested by Diamond and Dybvig (1983) and subsequently formalized by Cooper and Ross (1998) (see also Peck and Shell [2003]).[10]

---

[8] Note that without aggregate uncertainty, this strategy delivers the unconstrained-efficient allocation. With aggregate uncertainty, however, this is no longer the case.

[9] Postlewaite and Vives (1987) propose a related model that does not rely on multiplicity of equilibria as an explanation for bank runs. In their model, there is aggregate uncertainty about agents' preferences over intertemporal consumption and, in some cases, agents strategically rush to withdraw their funds before they have a true need to consume. Postlewaite and Vives do not have a sequential service constraint in their analysis.

[10] Gu (forthcoming) studies the case when different groups of agents observe the realization of different extrinsic random variables. She constructs run equilibria in which only a subgroup of the patient agents chooses to misrepresent preferences and withdraw.

As discussed above, suspension of convertibility rules out runs altogether in the standard Diamond-Dybvig framework. For this reason, Cooper and Ross (1998) restrict the possible set of banking contracts to those that take the form of a demand deposit contract without a suspension clause. Given this restriction, they show that the optimal demand deposit contract is consistent with the possibility of runs if the probability of a run is small enough. The extrinsic random variable in their model acts as a coordinating device. Agents observe the realization of the random variable and, for some realizations, play the run action. Interestingly, this modeling procedure works only if the bank does not observe the realization of the random variable. In this way, the bank remains uncertain about the motivation of the initial group of agents who attempt to withdraw: they may need to consume or they may be part of a run. If the probability that the bank assigns to experiencing a run is small enough, it will make fairly generous payments to early withdrawers, compromising the availability of resources for payment to those who wait. It is the anticipation of this situation by patient agents that, in turn, makes the run strategy consistent with equilibrium.[11]

While the findings of Cooper and Ross (1998) are quite interesting, their restriction to demand deposit contracts without a suspension clause is unsatisfactory when trying to identify the fundamental reasons for bank fragility. In Cooper-Ross' model, as in Diamond-Dybvig's, the fully unrestricted optimal banking contract rules out runs. Even if one does not go so far as to rule out runs completely, it is easy to see how their demand deposit contract without suspension would clearly be suboptimal and, hence, unlikely to materialize. At some point in the withdrawal process, the bank should be expected to realize that a run is taking place. In this (predictable) contingency, the simple demand deposit contract is easily seen to be suboptimal. Reducing the amount of resources paid to early withdrawers after that point would allow the bank to spread consumption more evenly among the remaining withdrawers, which would clearly improve the allocation (compared to keeping the payment constant and then running out of resources before some agents have been paid). As it turns out, this type of "partial suspension" (Wallace 1988, 1990) is also a feature of the optimal banking contract when there is uncertainty about the aggregate need for early liquidity in the economy, as we discuss in the next section.

In summary, Diamond and Dybvig (1983) identify three basic elements of a plausible theory of banking and bank fragility: (1) maturity transformation; (2) private information; and (3) sequential service. Uncertainty about the agents' total need for early liquidity could also be an important ingredient of a successful theory. Studying an explicit model of banking that incorporates

[11] Ennis and Keister (2006) clarify some aspects of the analysis in Cooper and Ross (1998) and derive additional results in their framework.

these components has proved to be a challenging task. Only recently has there been significant progress in understanding the implications for banking and bank fragility of combining all four components. We will review this research next.

2. TAKING SEQUENTIAL SERVICE SERIOUSLY

In this section, we discuss a series of papers that study versions of the Diamond-Dybvig model and in which special attention is devoted to the explicit specification of the sequential service constraint. We highlight (i) the interaction of aggregate uncertainty with the details of the environment that motivate the sequential service constraint and (ii) the implications of these assumptions for the possibility of bank fragility.

The Wallace Critique

In an influential article, Jacklin (1987) clarifies the role of trading restrictions in the Diamond-Dybvig model. He demonstrates that if agents are allowed to interact in a market after they discover the timing of their consumptions needs, there is an alternative arrangement that implements the unconstrained-efficient allocation without any possibility of runs. In this mechanism, agents initially buy shares in a firm that invests in the long-term technology. After discovering their consumption needs, impatient agents trade their shares with patient agents in exchange for consumption. Jacklin (1987) shows that this arrangement is capable of delivering the unconstrained-efficient allocation in the Diamond and Dybvig model, leaving no essential role for the institution of banking.

The market arrangement in Jacklin (1987), however, requires that the sequential service constraint be considered a restriction on the banking mechanism rather than a feature of the environment. The basic logic that allows the market arrangement to work requires that agents wait until all of them have discovered their consumption needs before they trade and consume. Under such a specification, however, it is not clear why a bank should be subject to sequential service. In principle, the bank could also wait before making any payments. In a way, assuming that banks make payments sequentially, as they do in real life, seems ad hoc in Jacklin's version of the Diamond-Dybvig model.

Wallace (1988) argues that the sequential service constraint should be considered a direct consequence of some frictions in the environment. If this were not the case, Jacklin's results imply that we should expect to see maturity transformation taking place solely in market-based arrangements and not in banks. Wallace interprets the fact that banks do perform a significant amount of maturity transformation as clear evidence of fundamental frictions that prevent

markets from playing this role. He describes an environment in which agents are isolated from each other when the early consumption opportunities arise and cannot meet to trade in a market. Agents are, however, able to contact the bank and they do so sequentially. Wallace assumes that all agents contact the bank before investment matures: some agents make an early withdrawal and others inform the bank that they will not withdraw until after investment has matured.[12]

These assumptions could be regarded, a priori, as fairly restrictive. The key to understanding their role is to realize that without these (or similar) assumptions, the Diamond-Dybvig model is unable to explain banking, illiquidity, or excess fragility. In a sense, these assumptions are necessary to have a successful theory of banking in the Diamond-Dybvig tradition. With this stipulation in mind, we can consider the isolation assumption a reasonable approach to capture, in a stylized manner, the fact that agents often have limited access to financial and asset markets when consumption opportunities arise. Banks, then, help agents overcome this kind of financial friction by providing a more reliable source of on-demand liquidity.

Wallace (1988) also emphasizes that once the sequential service constraint is considered a feature of the environment, it implies that payments to agents cannot be recalled at a later time. One can imagine that when a payment is made, the agent consumes these resources immediately. This approach implies that the type of deposit insurance scheme discussed by Diamond and Dybvig (1983) is infeasible in an environment with sequential service. Diamond and Dybvig assume that the government can tax agents after the opportunities to withdraw from the bank have passed. Wallace argues that if such taxation is possible, then agents must not need immediate access to their funds and the bank could wait until it has received all of the withdrawal requests before making any payments. If the sequential service constraint is truly a feature of the environment, it must apply to the government as well as to private institutions.

As we mentioned before, solving for the constrained-efficient allocation in the presence of an explicit sequential service constraint and aggregate uncertainty is a complicated matter. Wallace (1988, 1990) identifies some relevant features of such a solution. The basic insight is that each payment can only be contingent on information revealed up to the point when this payment is made. While the probability distribution over the possible values of the aggregate need for (early) liquidity is known a priori, the actual realization must be inferred from the withdrawal demands of agents. In other words,

[12] In the Diamond-Dybvig tradition, the order in which agents get an opportunity to withdraw is assumed to be exogenously given (generally determined by a random draw). In other words, agents in the model are not allowed to take explicit actions to change their order of arrival. This assumption is, of course, extreme and, unfortunately, not much is known so far about the case where it is not made.

the allocation must reflect the gradual process of information revelation that results from an explicit sequential service constraint.

Wallace (1988) shows that the constrained-efficient allocation under aggregate uncertainty must have early payments that depend on the order in which they occur. As more agents place withdrawal demands, the probability that the final number of impatient agents is large increases and the size of the payment to early withdrawers tends to decrease. This adjustment in the size of payments is the upshot from the fact that higher aggregate need for early liquidity implies less investment left to mature and, hence, a smaller total amount of resources available to distribute. Wallace (1990) calls the decreasing size of early payments a "partial suspension of convertibility."

Wallace (1990) studies a particular case of aggregate uncertainty that, at the cost of appearing somewhat artificial, provides a clear illustration of the forces influencing the determination of the efficient allocation. In particular, he considers a situation in which there are two groups of agents: one group that contacts the bank first (still sequentially) and has a known proportion of patient and impatient members, and another group that contacts the bank afterward and has either all patient or all impatient agents. This second group is the driver of aggregate uncertainty in the model.

Wallace demonstrates that the optimal payments to the first group of agents do not depend on the order in which the agents are paid (as in the Diamond-Dybvig model without aggregate uncertainty). However, once the first agent of the second group reveals his preferences, the efficient payment to him, and the payments to the rest of the agents that have not yet withdrawn from the bank, adjust significantly. The reason for this adjustment is that when the first agent of the second group contacts the bank, he reveals crucial information about the aggregate state, and this new knowledge renders necessary an adjustment to the pattern of payments. In more general (and, perhaps, realistic) cases of aggregate uncertainty, a similar logic applies: Payments to subsequent agents adjust if the information provided by the new agent contacting the bank reveals substantial information about the realization of the aggregate state.

Note that these articles, and indeed the entire literature we review here, do not explicitly consider a deposit insurance system. As mentioned above, Wallace's specification of the sequential service constraint prevents the government from being able to finance deposit insurance by taxing agents who have already withdrawn. In line with the mechanism design literature, one way to interpret the exercise in these articles is by asking: What is the optimal way to distribute whatever resources are available in the economy given the constraints imposed by the physical environment (and, in particular, sequential service)? Wallace's results suggest that complete deposit insurance is unlikely to be optimal; when there is an unusually large number of early withdrawals, the efficient allocation gives less consumption to those depositors who are relatively late in the order induced by sequential service.

The Green-Lin Model

In an influential article, Green and Lin (2003) pick up, basically, where Wallace leaves off. They write down an environment in the Diamond-Dybvig tradition with a finite number of agents and i.i.d. preference shocks, and they study the possibility of banking fragility in such a setup. They first study an environment without sequential service and show that the unconstrained-efficient allocation is strongly implementable.[13] This result is not very surprising, but confirms the need to deal with sequential service if the theory is to have any hope of addressing issues associated with the possibility of bank fragility.

After dealing with the simple case with no sequential service, Green and Lin (2003) specify a Wallace-style, explicit sequential service constraint and prove a remarkable result. They show that the unconstrained-efficient allocation (which takes into account sequential service but not incentive compatibility) is also strongly implementable. In other words, under their specification of the environment (including a specific form for the sequential service constraint), there is no room for bank fragility in the model.

The details of the sequential service constraint specified by Green and Lin are important for our discussion. Following Wallace (1988, 1990), Green and Lin assume that agents are isolated from each other during the early period and cannot observe other agents' actions during that time. Furthermore, as in Wallace, all agents contact the bank during the early period (i.e., before investment has had time to mature), either to demand a withdrawal or to inform the bank of their decision not to withdraw. Lastly, Green and Lin introduce a novel element into the picture: They assume that the order in which agents contact the bank is known to them with some degree of accuracy; in the extreme and simplest case, each agent exactly knows his or her place in the sequence of contacts with the bank. In the more complicated case, agents observe their "time" of arrival to the bank, which allows them to estimate their approximate position in the order. As it turns out, nothing of substance is lost from adopting the extreme case of perfect knowledge of the position in the order (see Green and Lin [2000]).

Several important implications arise from the particular assumptions used by Green and Lin in their specification of the explicit sequential service constraint. We briefly discuss these implications here since they help one appreciate the nature of the results and the way those results change when alternative specifications of the environment are used.

The combination of a finite population with i.i.d. preference shocks allows aggregate uncertainty to play a significant role in the determination of the outcomes in the model. In fact, the i.i.d. assumption implies that all possible partitions of the set of agents between patient and impatient occur with positive

[13] To prove this result, the i.i.d. assumption is actually not needed.

probability and that the bank can never fully discover the aggregate state until all agents have had a chance to withdraw. In other words, as each new agent contacts the bank, additional information becomes available that must be taken into account in designing the optimal allocation. As a result, the sequential service constraint is always binding in the unconstrained-efficient allocation in their environment.

Even though the sequential order of withdrawals gives the environment a certain degree of "dynamics" during the early period, the isolation assumption implies that the withdrawal game played by agents is a simultaneous-move, *static* game. Agents simultaneously decide on their strategies that, in combination with the particular realization of agents' preferences, will determine the final allocation of resources across the population. A strategy for an agent in the withdrawal game is a contingent plan that specifies whether or not to withdraw when contacting the bank in the early period, depending on the agent's realized preferences and (expected) place in the order of arrivals. The simultaneous-move, static nature of the game eliminates several technical complications like the need to specify off-equilibrium beliefs or to consider the possibility that agents would want to influence the decisions of other agents that come later in the order of withdrawals.

The remarkable result in Green and Lin (2003) relies on a type of backward-induction logic that comes into play once the agents receive reliable information about their order of withdrawal. Consider an agent who knows she will be the last one to contact the bank. By the time her opportunity to withdraw arrives, all of the other agents will have already taken their actions. Suppose, for example, that all of these agents have chosen to withdraw early. Then this last agent knows that if she chooses to withdraw early, she will receive whatever resources are left in the bank.[14] If she chooses to wait, however, she will receive the matured value of these assets in the later period, which is larger. Hence, if she is patient, she is strictly better off waiting to withdraw.

Now consider the penultimate agent to contact the bank. From the reasoning above, he knows that the agent who comes after him will only withdraw if she is truly impatient. He does not know her preferences, of course, but he knows the probability of her being impatient. The unconstrained-efficient allocation has the property that this agent will always be strictly better off waiting if he is patient. The heart of Green and Lin's proof that the unconstrained-efficient allocation is strongly implementable consists of showing that this property holds in general: If any agent believes that all agents whose opportunity to withdraw arrives after hers will report truthfully, she strictly prefers to report truthfully herself, regardless of the reports of those who contact the

[14] Green and Lin show that, because all sequences of preference types are possible and agents' marginal utility of consumption is assumed to be unbounded at zero, the resources available for the last agent are always strictly positive, even if all previous agents have chosen to withdraw.

intermediary before her. It is important to note that the unconstrained-efficient allocation is not chosen to satisfy this property, and hence the reasons why this property holds are far from straightforward. Once this property is established, however, their main result follows from using iterated deletion of strictly dominated strategies to arrive at the strategy profile in which all agents report truthfully.

Green and Lin (2003) conclude from their analysis that something is missing in the Diamond-Dybvig theory of banking fragility. In their specification of the model, a bank can ensure that resources are allocated efficiently across depositors without introducing the type of fragility highlighted by Diamond and Dybvig.

Extensions and Clarifications

Andolfatto, Nosal, and Wallace (2007) study a modified version of the Green-Lin model in which they allow for a more general class of utility functions and clarify the importance of the i.i.d. assumption for obtaining the strong implementation result. They also (implicitly) change the sequential service constraint so that it differs in important ways from the one used by Green and Lin (2003). In the Green-Lin model, an agent does not observe the actions of those agents that have contacted the bank before her. Andolfatto, Nosal, and Wallace (2007) instead assume the bank informs each agent of the complete profile of actions taken by the agents before her, which allows an agent's action to be contingent on the actions of (a subset of) the other agents. This change in the environment makes the incentive compatibility constraints stronger, in the sense that fewer allocations are implementable.

Andolfatto, Nosal, and Wallace (2007) show that, in this modified environment, any allocation that is implementable is also strongly implementable. The logic of their proof is simple but powerful. In order for an allocation to be implementable in their environment, it must be the case that an agent, following *any* sequence of reports by the agents who have preceded her, prefers to report truthfully when all other agents report truthfully. Suppose now that an agent believes that some of the agents who preceded her have lied about their types, but that all agents who come after her will report truthfully. Under the assumption that preference types are i.i.d., the fact that some agents may have lied has no impact on her payoffs—all that matters is the sequence of actual reports. The fact that the allocation is implementable, therefore, implies that an agent will prefer to report truthfully as long as she believes that those who follow her will also report truthfully. Given this fact, the same type of backward-induction argument used by Green and Lin (2003) can be used to show that the allocation is strongly implementable. Since the constrained-efficient allocation is, by definition, incentive compatible and, hence, implementable, a corollary to the main result in Andolfatto, Nosal, and

Wallace is that the constrained-efficient allocation is strongly implementable and that there is no room for fragility in their model.

If preferences are of the type used by Diamond and Dybvig (1983), then Green and Lin's (2003) proof of their main result is actually powerful enough to establish the strong implementability of the unconstrained-efficient allocation even when the sequential service constraint is specified as in Andolfatto, Nosal, and Wallace. While the analysis presented by Andolfatto, Nosal, and Wallace is more general in that it allows for a wider range of preferences than does the analysis by Green and Lin, it does not focus on the unconstrained-efficient allocation; the results only apply to implementable allocations.

An important clarification should be made at this point. Green and Lin (2003) find the constrained-efficient allocation by first solving an auxiliary problem without the incentive compatibility constraints and then showing that the solution is, actually, incentive compatible. For the general class of utility functions considered by Andolfatto, Nosal, and Wallace (2007), the incentive compatibility constraints are likely to be binding in many cases, even if agents' preference shocks are independent. For this reason, the methodology employed by Green and Lin (2003) to find the constrained-efficient allocation is likely to fail in many of the cases considered by Andolfatto, Nosal, and Wallace (i.e., the solution to the planning problem without the incentive constraints may not be incentive compatible). Finding the constrained-efficient allocation, then, may involve additional complications like identifying which incentive compatibility constraints are likely to be binding and then "reshaping" the payment scheme to minimize the distortions induced by the incentive compatibility requirement.

Ennis and Keister (2009a) modify the Green-Lin model in a different way by relaxing the assumption that preference types are independent across agents. All other elements of the model, including the specification of the sequential service constraint, are exactly as in the Green and Lin analysis. Under the assumption that preferences exhibit constant relative risk aversion, they derive the unconstrained-efficient allocation in closed form, which allows them to calculate examples with more agents than had been done in the previous literature. They present a series of examples that show how the results of Green and Lin (2003) can break down when types are correlated. In these examples, there exists an equilibrium of the withdrawal game in which some, but not all, agents run on the bank by withdrawing early regardless of their true consumption needs.

The logic used by Green and Lin (2003) to show that the last agent to contact the bank has no incentive to misreport her type still holds in this setting. For this reason, there cannot be an equilibrium in which all agents run on the bank. The equilibria constructed in Ennis and Keister (2009a) have the property that those agents who have a relatively early opportunity to withdraw choose to run, while those who are relatively late withdraw only

if they have a true consumption need. An essential feature of these examples is that the key property identified by Green and Lin fails to hold—an agent who believes that everyone who arrives after her will report truthfully may nevertheless prefer to misrepresent her type. These results show that the strong-implementability result of Green and Lin (2003) relies on more than a simple use of backward-induction logic; it depends critically on properties of the unconstrained-efficient allocation that may not hold when agents' preference types are not i.i.d.

Alternative Approaches to Sequential Service

The Green-Lin formulation of the sequential service constraint is appealing in several dimensions. To begin with, it is clearly specified and helps the reader view the allocation problem in the Diamond-Dybvig model in terms of the standard theory of mechanism design. In addition, their specification shows how important "dynamic" features of bank runs can be captured in a model without bringing in the complications associated with dynamic games. Several subsequent articles have investigated how much the particular assumptions Green and Lin made matter for their strong implementation result. Peck and Shell (2003) modify the Green and Lin environment in two ways, considering both a more general specification of agents' preferences and a different specification of the sequential service constraint. They find that the strong implementation result of Green and Lin goes away under this alternative set of (also reasonable) assumptions.

With respect to agents' preferences, Peck and Shell allow the marginal utility of impatient agents to differ from that of patient agents. When impatient agents have a high marginal value of consumption, the bank will want to give relatively large payments to those agents who withdraw early. If this effect is strong enough, the incentive constraint for patient agents will be binding in the constrained-efficient allocation, something that could not happen in the Green-Lin model. The relatively large payments made on early withdrawals increases the incentive of patient agents to misrepresent their type if they expect others to do so.

The second change introduced by Peck and Shell is in the way the sequential service constraint is specified. The agents in Peck-Shell do not observe any information about their position in the order of arrival at the bank before making their withdrawal decision. Instead, each agent views the positions as being randomly assigned after withdrawal decisions have been made. Under this approach, the backward-induction logic used by Green and Lin cannot be applied since no agent is confident that she will be the last one to contact the bank.

These two changes—in preferences and in the specification of sequential service—are both important for the examples of run equilibrium constructed

by Peck and Shell. The change in the sequential service constraint enlarges the set of implementable allocations relative to the Green-Lin model, since now there is a single incentive compatibility constraint rather than a separate constraint following each possible history of reports leading up to an agent's decision. The change in preferences implies that the constrained-efficient allocation in the Peck-Shell setting may not be implementable in the Green-Lin specification of sequential service and, in fact, the examples in Peck and Shell have this feature. It remained an open question whether both of these elements were needed to overturn the strong implementation result of Green and Lin.

Ennis and Keister (2009a) answer this question by constructing examples of run equilibria in which the environment is identical to that in Green and Lin's article except that agents do not know their position in the order of arrival at the bank. There is no change in preferences and, as a result, the constrained-efficient allocation is exactly as in Green and Lin's model. These results show that it is the change in the sequential service constraint, and not the nonstandard specification of preferences, that is at the heart of the Peck-Shell result.

Peck and Shell make another interesting change to the sequential service constraint, although they show it is not important for their results. Green and Lin assume that all agents contact the bank during the first round of withdrawals, regardless of whether the agent wishes to withdraw or not. Peck and Shell, instead, assume that only agents who wish to withdraw contact the bank. This change results in a more coarse information structure for the bank. In particular, the bank only observes withdrawals and, as a consequence, the efficient allocation is less responsive to the type realizations of those agents who are early in the line. In the Green and Lin setup, when the bank observes that the first agent in the line is impatient, it adjusts the constrained-efficient allocation by reducing the early payments. In Peck and Shell, information arrives to the bank more slowly, leading the bank to make fewer adjustments to the allocation early on in the process. However, Peck and Shell show that their same result obtains when all agents report to the bank regardless of whether they want to withdraw or not (see their Appendix B).

The banking contract that implements the optimal allocation in the Green-Lin setup generally involves payments to agents that are highly contingent on the information collected by the bank up to the point of actually making the payment. This feature is a consequence of the combination of aggregate uncertainty with sequential service and seems counter to common practice in banking where the face value of deposits is respected under most circumstances. This counterfactual implication of the Diamond-Dybvig theory was, in fact, recognized since its inception (see, for example, Postlewaite and Vives [1987]).

A plausible modification of the details involved in the specification of Green and Lin's sequential service constraint may move the theory closer to reality in this respect. In particular, some preliminary results from our own research (see Ennis and Keister [2008]) suggest that when the bank only observes withdrawals as they occur (following the specification in Peck and Shell [2003]), but obtains no information about the realized preferences of agents who do not intend to withdraw, the constrained-efficient allocation more closely resembles a demand deposit contract. This result holds even when agents know their place in the order at the time of their early withdrawal decision (an assumption made by Green and Lin [2003] that was not present in Peck and Shell [2003]). Interestingly enough, when this modification to Green and Lin's specification of the sequential service constraint is introduced, the efficient allocation may no longer be strongly implementable for some parameter configurations and the possibility of bank fragility reappears in the model.

As we have seen, one of the main differences among the alternative specifications of the sequential service constraint lies on the amount of information that an agent has at the time of deciding whether or not to withdraw. In the version studied by Green and Lin (2003) the agent knows if she is patient or impatient and her place in the order of sequential contacts with the bank. In the version of Andolfatto, Nosal, and Wallace (2007) the agent knows more (the actions of those agents prior to her in the line) and in the Peck-Shell version the agent knows less (only whether she is patient or impatient).

In a recent article, Nosal and Wallace (2009) propose an alternative interpretation of the various specifications of the sequential service constraint in this dimension. In particular, they assume that the agent directly receives information only about his preferences, and that the bank can communicate to the agent (before he chooses whether or not to withdraw) information that it may have about the agent's place in the order and what the other agents before him have done. This way of thinking about the model provides a unified way of viewing the alternative specifications that have been studied in the literature, each corresponding to a different assumption about the amount of information the bank is revealing to the agent.

A natural question to ask under this approach is how much information the bank would reveal to agents if it were allowed to choose. Nosal and Wallace (2009) study this question when the bank is a benevolent entity (a planner). An interesting complication arises at this point. The set of implementable allocations is strictly larger when agents do not have any information about the order, which would in principle give the planner more flexibility in designing the payoff schedule. However, as Peck and Shell (2003) have shown, the constrained-efficient allocation may not be strongly implementable in this case. Nosal and Wallace show that if the planner is only concerned with implementation (but not with strong implementation) then, under some

parameter values, it will not want to reveal information about the order to the agents.

This finding has important implications for the possibility of a bank run in the model. If the bank believes a run is very unlikely to occur, even when one is consistent with equilibrium, then it may choose not to reveal information that could rule out the possibility of a run. This happens because, by not revealing information, the bank improves the outcome that obtains when agents do not run. In other words, there is a tradeoff in the model between efficiency when a run does not occur and eliminating the possibility of a run altogether.

3. OTHER POSSIBLE INGREDIENTS

The Green-Lin model and the modifications of it that we have discussed so far describe a very basic environment that abstracts from many other features that are typically associated with the workings of banking institutions. A natural question, then, is to ask whether or not there might be additional ingredients that are relevant to explain banking and bank fragility in models within the Diamond-Dybvig tradition. In this section, we discuss three possibilities that have been recently examined: self-interested bankers, limited commitment, and investment restrictions.

Self-Interested Bankers and Moral Hazard

In all of the discussion above, the bank is operated with the objective of maximizing the welfare of its depositors. It seems more in line with reality, however, to explicitly model situations in which the banker does not always act in depositors' best interests. In Green and Lin's environment, the banker centralizes the information about the aggregate state as it is gradually revealed by the sequential decisions of agents. While the banker may be able to commit to a payment contract, the contract may give the banker incentives to manipulate the information provided to the remaining agents after each withdrawal. Based on this logic, Andolfatto and Nosal (2008) illustrate how a self-interested banker in charge of delivering a contract of the type studied by Green and Lin may want to misrepresent the situation and artificially reduce payouts to depositors.

After establishing this fact, Andolfatto and Nosal investigate alternative schemes that could be used to give proper incentives to the banker. To benefit from a banking arrangement depositors must, eventually, be able to compare the claims of the banker with some relevant information about the true aggregate state. As a consequence, new assumptions are needed about the accessibility to information by agents and the banker. Unfortunately, there is no clear natural way to proceed in formalizing this issue. Andolfatto and Nosal pick one particular configuration—agents can convene after investment

matures and collect information about their actual preferences. In this case, Andolfatto and Nosal show that the payments in the best contract delivered by a self-interested banker may be less sensitive to the aggregate state than the Green-Lin contract and, hence, may appear more in line with the type of demand deposit contracts that are common in real-world banking. However, this result depends on parameters, and in certain cases the contract actually becomes more complex than the Green-Lin contract (with new contingencies and positive early payments to patient depositors).

Andolfatto and Nosal also study the implications for financial fragility of considering explicitly the incentives of the banker. They conclude that it may be harder to construct equilibria in which patient agents misreport their types. However, their analysis is far from conclusive. Overall, their work demonstrates that analyzing the effects of bankers' agency problems in the Green-Lin model, while potentially important, is not a straightforward task. This line of inquiry, though, seems to us potentially very fruitful and deserving of further attention.

Limited Commitment

Another ingredient that may be important for explaining financial fragility is limited commitment on the part of the bank or on the part of policymakers more generally. Ennis and Keister (2010) study a version of the Diamond-Dybvig model in which the bank cannot commit to a plan of action; rather, the payment to each agent is only decided when that agent arrives to withdraw. The key aspect of this lack of commitment power is that it prevents the bank from being able to credibly use a suspension of convertibility clause to uniquely implement the constrained-efficient allocation.

In the environment studied by Ennis and Keister (2010), there is no aggregate uncertainty and the bank knows precisely how many agents will be impatient.[15] The sequential service constraint follows Peck and Shell (2003) in assuming that only those agents seeking to withdraw contact the bank. Once the number of withdrawals passes a certain threshold, therefore, the bank will know for sure that a run is underway. However, once this situation is reached, it will not be ex-post optimal for the bank to follow through with a suspension, since doing so would imply giving no consumption to some agents who are truly impatient (as in Ennis and Keister [2009b]). Instead, the bank continues making payments to some of the withdrawing agents, compromising resource availability in the later period.

Ennis and Keister (2010) demonstrate that when the bank initially believes that a run is unlikely, it will in some cases choose payments that make a run

[15] An interesting avenue for future research would be to apply the techniques developed in Ennis and Keister (2010) to the model with a finite number of agents and aggregate uncertainty.

consistent with equilibrium. In other words, bank fragility is possible in the Ennis-Keister version of the Diamond-Dybvig model with limited commitment, even though there is no fundamental source of aggregate uncertainty in the model. Interestingly, the equilibria of the model have a natural "dynamic" structure, which derives from the fact that agents have information about their position in the order of early withdrawal opportunities (as in Green and Lin [2003]). An equilibrium bank run consists of an initial wave of withdrawals, which is followed by a reaction from policymakers. Following this reaction, the run may end or it may continue with another wave of withdrawals taking place, which would lead to another reaction from policymakers, and so on. This interplay between the withdrawal decisions of agents and the reaction of policymakers seems to be an important feature of real-world banking crises.

Investment Restrictions

There is a long tradition in policy of regulating the activity of banking. One common approach has been to restrict the type of investments that banks are allowed to undertake. For example, for more than 50 years, banks in the United States that accepted deposits from the public were prohibited from engaging in certain asset management activities, which were reserved for a different set of institutions called investment banks. These restrictions were imposed partly as a way to address the possibility of bank fragility. When those policies were designed, a formal theory of banking was not available. Diamond and Dybvig (1983) and the literature that followed have provided such theory and, hence, it is natural to ask how this kind of policy influences outcomes in the models within this tradition. Peck and Shell (2010) address this question.

Peck and Shell (2010) consider an environment with an indivisibility in consumption, which is aimed at capturing the payment function of demand deposits: A check written for a purchase, for example, either pays the bearer at par or may not be useful for exchange. Peck and Shell consider an environment with two investment technologies: one technology is as in the standard Diamond-Dybvig model and the other has higher long-run return but is completely illiquid in the short run. They analyze two regulatory systems for banks—a unified system and a separated system. In the unified system, banks are allowed to invest in both technologies on behalf of agents. In the separated system, however, banks cannot invest in the illiquid technology and agents do that directly. Somewhat surprisingly, Peck and Shell show that runs can happen in the separated system but not in the unified system. They conclude that policies that impose restrictions on the investment strategies of banks can actually have unexpected, counterproductive effects by inducing fragility in the system.

4. CONCLUSION

Understanding the root causes of the banking crises that have been observed around the world is an extremely difficult task. Some commentators claim that self-fulfilling behavior on the part of depositors and investors plays a critical role, while others emphasize more fundamental factors related to the value of banks' assets. Banking crises are complex phenomena that typically occur in conjunction with a variety of unfavorable financial and macroeconomic factors, making it difficult to determine the true underlying cause of an event. In spite of these difficulties, progress has recently been made in several directions. This article reviews the progress in one of these directions.

The literature we have discussed shows that it is possible to provide an internally consistent explanation for the self-fulfilling interpretation of bank runs. However, this literature also shows that the details of the environment are important. In other words, the fragility of banks in these models is the result of physical and informational frictions, but only specific combinations of these frictions lead to fragility. In particular, information about the actions of agents must not flow too quickly, so that the bank makes a significant amount of payments to depositors before discovering whether or not a run is underway. In addition, some feature of the environment must make suspension of convertibility clauses in deposit contracts either undesirable or ineffective.

How important are self-fulfilling factors in the explanation of observed crises? It may very well be the case that the types of frictions described in this paper were present in the real economy and that observed financial crises have had a considerable self-fulfilling component. If these theories are a useful reflection of reality, however, it is important to realize that natural changes in the way information flows in the economy (because of, for example, technological innovation) could have substantial implications for bank fragility in the future. In addition, it seems important to recognize that our understanding of the issues involved remains fairly limited. Identifying appropriate policies to deal with bank fragility, then, must be an ever-evolving activity that takes into account changes in the structure of the financial system as well as further developments in our understanding of the issues. The theories we have discussed here provide a solid foundation for pursuing these important and pressing issues.

REFERENCES

Andolfatto, David, and Ed Nosal. 2008. "Bank Incentives, Contract Design, and Bank Runs." *Journal of Economic Theory* 142 (September): 28–47.

Andolfatto, David, Ed Nosal, and Neil Wallace. 2007. "The Role of Independence in the Green-Lin Diamond-Dybvig Model." *Journal of Economic Theory* 137 (November): 709–15.

Bryant, John. 1980. "A Model of Reserves, Bank Runs, and Deposit Insurance." *Journal of Banking and Finance* 4 (December): 335–44.

Cooper, Russell, and Thomas W. Ross. 1998. "Bank Runs: Liquidity Costs and Investment Distortions." *Journal of Monetary Economics* 41 (February): 27–38.

De Nicoló, Gianni. 1996. "Run-Proof Banking Without Suspension or Deposit Insurance." *Journal of Monetary Economics* 38 (October): 377–90.

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

Ennis, Huberto M., and Todd Keister. 2006. "Bank Runs and Investment Decisions Revisited." *Journal of Monetary Economics* 53 (March): 217–32.

Ennis, Huberto M., and Todd Keister. 2008. "Run Equilibria in a Model of Financial Intermediation." Federal Reserve Bank of New York Staff Report No. 312 (January).

Ennis, Huberto M., and Todd Keister. 2009a. "Run Equilibria in the Green-Lin Model of Financial Intermediation." *Journal of Economic Theory* 144 (September): 1,996–2,020.

Ennis, Huberto M., and Todd Keister. 2009b. "Bank Runs and Institutions: The Perils of Intervention." *American Economic Review* 99 (September): 1,588–607.

Ennis, Huberto M., and Todd Keister. 2010. "Banking Panics and Policy Responses." *Journal of Monetary Economics* 57 (May): 404–19.

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.

Gu, Chao. Forthcoming. "Partial Bank Runs Triggered by Noisy Sunspots." *Macroeconomic Dynamics*.

Jacklin, Charles J. 1987. "Demand Deposits, Trading Restrictions, and Risk Sharing." In *Contractual Arrangements for Intertemporal Trade*, edited by Edward C. Prescott and Neil Wallace. Minneapolis: University of Minnesota Press, 26–47.

Nosal, Ed, and Neil Wallace. 2009. "Information Revelation in the Diamond-Dybvig Banking Model." Federal Reserve Bank of Chicago Policy Discussion Paper Series (December).

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

Peck, James, and Karl Shell. 2010. "Could Making Banks Hold Only Liquid Assets Induce Bank Runs?" *Journal of Monetary Economics* 57 (May): 420–7.

Postlewaite, Andrew, and Xavier Vives. 1987. "Bank Runs as an Equilibrium Phenomenon." *Journal of Political Economy* 95 (June): 485–91.

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.

Wallace, Neil. 1990. "A Banking Model in which Partial Suspension is Best." Federal Reserve Bank of Minneapolis *Quarterly Review* 14 (Fall): 11–23.

Economic Quarterly—Volume 96, Number 1—First Quarter 2010—Pages 59–82

Inside-Money Theory after Diamond and Dybvig

Ricardo de O. Cavalcanti

This article argues that the model in Diamond and Dybvig (1983, DD hereafter) was a significant conceptual and methodological advance in studying banking arrangements. Its methodological contribution was the use of mechanism-design theory rather than the old strategy, still prevalent in textbooks and some of macro, of tacking a banking sector onto a model of market exchange. A great deal of attention has been given to the model's multiple equilibria and interpreting them as financial fragility. This attention is warranted, but there are other less recognized implications of the model. I provide examples in which the model is used to address banker incentives and means of payment. I also show how its methodology is related to recent work that uses monetary models to consider money, credit, and imperfect monitoring.

Recent events provide a good opportunity to put into perspective progress in the field of money and banking. The idea that banks are inherently unstable is as old as the field itself. The recent economic crisis in the United States and around the world was met by a new generation of central bankers familiar with the notion of financial fragility in DD. The dramatic increase in the balance sheet of the Fed, for instance, led by the purchasing of private securities whose markets had virtually disappeared, seems to indicate that financial meltdowns may not be restricted to unique conditions like those of the Great Depression.[1]

I am grateful to Huberto Ennis, Juan Sánchez, and Ned Prescott for helpful comments. I also thank graduate students Jefferson Bertolai, Artur Carvalho, Murilo Ferreira, and Bruno Teixeira, who agreed to present several of the papers discussed in this article in our money and banking course at FGV, as well as the *EQ* managing editor, Amanda Kramer, for her invaluable help. The views expressed in this article are not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: ricardo.cavalcanti@fgv.br.

[1] In this sense, macroeconomics is far from a "solved problem." Even though bank runs, as traditionally described, were not significant, the fact that returns on short-term Treasury bonds approached zero indicates a generalized lack of confidence in the strategy of depositing funds at private institutions.

Overall, the DD framework is a model of intertemporal trade with relatively few variables. It allows for a sharp description of frictions that private information and sequentiality of transactions imposed on the provision of insurance against preference shocks. This tractability emphasizes the question of whether or not optimal allocations are implemented uniquely, that is, whether or not the optimum is fragile to runs. A less appreciated issue, and my focus here, is how the DD framework suits developments in monetary theory that emphasize imperfect monitoring.

Micro and Macro Mechanisms

Findings about the multiplicity of equilibrium outcomes are common in the field of money and banking. Expectations about future behavior are important for the current value of money and its substitutes. While this property tends to raise the possibility of multiple outcomes, precise conclusions depend heavily on assumptions about what is traded and how markets are organized. There is an old habit in macroeconomics of building models around institutions seen empirically as important devices with which to organize savings: deposits, bonds, capital, alternative currencies, etc. DD represent a new modeling approach with their use of a very compact model of the pooling and intertemporal redistribution of resources in a single resource-constraint world. A bonus of this minimalist approach is the speed at which the literature identifies key elements that drive the role for liquidity provision and the possibility of financial fragility in their theory. These elements, taken as immutable primitives, rule out remedies like deposit insurance that even DD point to in their title.

Reviewing the DD contribution on the basis of this sort of "micro" problem alone is attractive because of the sharp results achieved by follow-up work. This can be appreciated by the material I present in Section 3, which follows an informal summary of model choices in Section 2. In addition to the emphasis on the issue of multiplicity guided by the model of Green and Lin (2003), I make considerations about imperfect monitoring that are inspired by the model of Prescott and Weinberg (2003). In Section 4, those considerations provide a bridge to identify monitoring in models where banks are people, which is different from the original DD setup. Examples include the models of Calomiris and Kahn (1991) and Deviatov and Wallace (2009). The latter, a monetary model, leads us to much larger mechanisms in macroeconomics.

I find it important, however, to also offer some history of thought perspective by starting in Section 1 with brief discussions of the models of Shubik and Wilson (1977) and Bryant (1980). These early models illustrate the difficulty of applying general-equilibrium theory to money and banking. These attempts were too early to benefit from the mechanism-design approach, and imposing banking exogenously meant they could not answer some important fundamental questions about the role of imperfect monitoring and sequential

service in the provision of insurance. In their defense, the mechanism-design approach of today is more abstract and takes the model further away from the data than does the traditional approach.[2]

These early general-equilibrium models of money and banking address a rich set of questions that the DD methodology can only partially answer. I provide below a very short description of a planner problem that I find useful and serves the purpose of introducing the liquidity problem set by DD.

A Benchmark

Let us first consider a static endowment economy with two goods, grapes and wine, and a finite population (as I note below, a more precise analogy to the DD model would have the supply of these goods competing for the same resources). Individuals are initially identical in their preferences and endowments, but an exogenous stochastic process generates a distribution of marginal utilities across people. A benevolent social planner is called to organize the best allocation of goods, constrained by the fact that marginal utilities are private information.

This can be called a liquidity problem to the extent that the designed system first acquires control of all endowments and then makes transfers of goods based on individual announcements about preferences, the way actual banking systems seem to operate when we interpret withdrawals as announcements of impatience. Incidentally, in monetary theory, people are subject to numerous periods of shocks and transfers. When the history of transfers to particular individuals becomes difficult to monitor, money becomes necessary as a way to introduce recordkeeping. In our benchmark economy, in contrast, individuals are identified and their announcements are perfectly recorded. If the planner can collect all announcements before making the transfers of goods, contract theory can be used to design constrained-efficient allocations, as in the literature on adverse selection, but without a convincing notion of financial fragility. DD, and the literature that follows, identify a key aspect of banking that is capable of tying together liquidity and fragility. This aspect is the sequential service constraint, which is the counterpart in banking to anonymity in monetary theory.

Sequentiality is an assumption about the physical environment of grapes and wine that makes transfers more difficult to organize. The exogenous stochastic process now includes the assembly of a queue and forces the allocation of grapes to be made sequentially, that is, according to the flow of announcements. The allocation of wine is made after all transfers of grapes

[2] This explains why the agenda of pursuing general-equilibrium theory with limited forms of institutional design is experiencing a revival.

take place. As a consequence, consider an individual with high marginal utility for wine relative to grapes. This individual would not be concerned if people ahead in the queue state a preference for grapes, unless this behavior has implications for the aggregate endowment of wine. The DD model introduces this concern by replacing grapes and wine for, respectively, date-1 and date-2 goods that compete for the same resource constraint. Now an above-average desire for grapes creates a concern about the scarcity of wine.

Special assumptions about preferences and the stochastic process governing the queue and the distribution of marginal utilities give rise to a tractable model. In this case, the optimum is identified by a separating equilibria of the mechanism chosen by the planner. A bank run is identified by a pooling equilibria (for a population subset) in which the planner fails to observe the true types in sequence because individuals misrepresent their preferences by announcing a high desire to consume good 1, the grapes. This is an outcome of low welfare.

Notice that the value of truthful reporting is twofold. By declaring a true type, an individual not only helps with the allocation of scarce resources today, but he also provides invaluable information about the aggregate state of nature when preferences are correlated across people. Truth-telling behavior thus gives rise to a positive externality by helping the planner estimate the distribution of marginal utilities for the whole population, improving future transfers. A question emerges about whether or not individuals find incentives for providing this externality. DD give us a significant conceptual and methodological advance in studying banking arrangements because questions like this can now be studied formally.

A final note may help the reader to evaluate model choices in monetary models. Money models usually assume the existence of a continuum of individuals because deviation payoffs become computable by simple reference to equilibrium outcomes. But there is no reason for the money literature to insist on total anonymity. Hybrid models can allow for imperfect monitoring of money traders, bringing the analysis closer to banking. Some recent models achieve a co-existence of money and credit by restricting monitoring to cover only a subset of the population. Like in DD, the mechanism-design approach has an interesting implication: Optimal allocations have subtle features that are difficult to anticipate in terms of standard market outcomes.[3]

1. BUILDING DICHOTOMIES

A review linking DD to inside-money theory requires a perspective on the field of money and banking and the early tradition that they replace. The

[3] Future research could also consider the option of introducing the friction of sequential service in monetary models, raising the issue of financial fragility in the provision of inside money.

models proposed by Shubik and Wilson (1977) and Bryant (1980) are early and influential attempts to give money and banking important roles in the allocation of resources in endowment economies. There is, however, a kind of dichotomy that is common to both approaches. The authors build on basic models for which price theory works well, adding an exogenous financial sector, similar in spirit to the way cash-in-advance constraints have been used to give money a role in dynamic general-equilibrium economies. These models illustrate how DD break away from the dichotomy tradition of imposing a trade game to an otherwise well-organized economy.

The Shubik-Wilson Model

Shubik and Wilson (1977, SW hereafter) pioneered a formulation of optimal bankruptcy rules in a monetary setting. Although the trading games proposed by Shubik in his extensive work can be criticized as having been chosen arbitrarily, it is clear that penalty devices used by SW to constrain credit reappear in a variety of borrowing constraints in modern general-equilibrium models.[4]

In the SW model, individuals inhabit a pure endowment economy with a single date and two consumption goods. The population is equally divided into two types. The marginal utility of their Cobb-Douglas preferences and the initial endowments vary with types so there is a strong incentive to trade. It is assumed that individuals allocate a fraction of their endowments for sale and bid units of bank money for goods in two trading posts, one for each good. The corresponding prices are given by the ratio of total monetary bids to quantities of goods put up for sale in each market.

An "outside bank" appears as a mechanism to supply bank money in exchange for payment promises issued by individuals, since individuals are not endowed with money.[5] By introducing a sequence of events to the process for borrowing from the bank, making bids, receiving proceeds from sales, and repaying bank loans, SW give strategic choices a sequential interpretation. The implicit assumption is that individuals cannot commit to future actions. The mechanism or bank fixes the per capita amount of money it issues. Then individuals bid quantities of personal notes in this third trading post set by the bank and called the money market. Likewise, the money price is computed as the ratio of total bids to money supply. This ratio can also be identified as the gross rate of interest on loanable funds, a fee intended to "protect" the bank against default.

---

[4] See, for instance, Dubey, Geanakoplos, and Shubik (2005) and the extensive literature on endogenous borrowing constraints.

[5] Bank money, being part of an allocation, can be viewed as inside money.

Bankruptcy is allowed in the sense that an individual who is unable to pay his debt in full receives a utility penalty, which takes the form of a linear function and maps the unpaid fraction of promises into a utility loss. Utility penalties are exogenous and are allowed to vary with preference types. Money leftovers provide no utility so that individuals plan bankruptcy to different degrees as these parametric penalties change.

SW use their model to compare symmetric (across types) equilibrium allocations with those that would attain in the absence of intermediation (allocations in the contract curve of the Edgeworth box) for alternative assumptions about the information structure on which strategies are made contingent. In one specification, individuals receive some information about prices and bids in the money market prior to selecting bids and sales in the markets for goods. In another specification, SW assume that all bids and sales quantities are chosen simultaneously in noncooperative fashion and where deviations cannot affect prices because the population size is taken to infinity. They find combinations of penalties for which individuals of different types alternate in being solvent, as well as a region with bankruptcy for both types in which trade and prices are the same as the competitive equilibrium. If penalties are high enough there is no bankruptcy and equilibrium is again competitive.

SW defend their approach as a complete and consistent microeconomic foundation for monetary economics (the reason to build on the basic general-equilibrium model) where there are rules that describe all moves and cover all contingencies.[6] The role of money is to finance the float in trading posts requiring, by construction, bids and offers be made simultaneously. Yet, the role of credit is a passive one: The bank auctions enough universally accepted means of payment so that competitive-equilibrium optimality is restored. Because penalties are exogenous, the model is not meant to provide a theory of inside money in its true sense: a prediction of how much credit society is able to accommodate and how much money is needed to organize trade.

The Bryant Model

Compared with SW, the approach presented by Bryant (1980) also gives bank firms special abilities in channeling intertemporal savings. While written in less formal terms, Bryant's article is also more ambitious because it introduces bank fragility, and it does so using a model featuring monetary equilibria derived from first principles. Bryant builds on the overlapping-generations model of money, introducing a role for bank deposits, fractional reserves, and

[6] That aspect is also cited by Amir et al. (2009), in a special issue of *Games and Economic Behavior* in honor of Shubik, as the reason for a "natural" emergence of financial institutions such as money, credit, and bankruptcy rules.

anticipation of withdrawals. The analysis of his model is simplified by the emphasis on steady states without inflation.

There are two parts in Bryant's analysis. In the first part, a basic model without uncertainty is presented. It assumes that young individuals can acquire money by selling goods to old individuals, but that credit transactions must be done through an intermediary, a bank firm. The bank is valuable because two types of individuals are born in a given period and one of the types is composed of people who are only endowed with goods when old. They must thus borrow to finance consumption when young. The bank maximizes profits but its linear lending technology yields zero profits in the steady-state competitive equilibrium. This part of the model delivers predictions about consumption of the two types of individuals, holdings of money and deposits, and debt positions. It also discusses how lump-sum taxation of those receiving positive endowments when young can be used to finance payments of interest on government bonds that can only be held by the bank.

A good sense of Bryant's model can be captured in a version with two-period lived generations; no uncertainty; a continuum of newborns who are equally divided into two subgroups of unity-measure population, types 1 and 2; and where there is one consumption good per date. The per capita endowment for a type 1 is k units of goods when young and zero when old, while the reverse holds for a type 2. There is also a quantity of m units of fiat money evenly distributed among the initial old. In addition to the option of acquiring money from the old by selling goods at price $1/p$, the type-1 individuals use part of the money acquired to hold deposits with the bank. The variables indicating their final holdings of money and deposits are m^1 and d, respectively. The bank can then lend money to type-2 individuals, charging an interest, r, to be paid in money at the next date. The per capita debt of type-2 individuals is m^2.

Because the creation of deposits is costly, with costs in proportion $g \in (0, 1)$ to the goods value of dollars deposited, type-2 individuals only borrow what they need to consume when young and hold no money in equilibrium. In addition to the creation of deposits and loans, the bank can buy b units of government bond: a promise of a dollar next period, at price s. The per capita tax levied on type-1 individuals is, in equilibrium, $p(1 - s)b$.

Individual type 1 chooses (m^1, d) in order to maximize his utility in the budget constraint set defined by $c_1^1 \leq k - pm^1 - pd - p(1 - s)b$ and $c_2^1 \leq pm^1 + pd$. Individual type 2 sets m^2 so as to maximize her utility in the budget given by $c_1^2 \leq pm^2$ and $c_2^2 \leq k - (1 + r)pm^2$. In each period, the bank chooses levels of deposits, loans, and bond holdings so as to maximize profits. Its maximization problem is linear and the necessary conditions for zero profits with positive intermediation and bond holdings can be shown to be $(1 + r)^{-1} = 1 - g$ and $s = 1 - g$. In equilibrium, all goods are consumed or used up in intermediation, all the supply, m, of money is held by the type-1

young ($m^1 = m$), and all the supply, b, of bonds is held by the bank. As the government chooses different levels of b, keeping $m + b$ constant, more of the social cost of intermediation is transferred from borrowers to lenders (the taxpayers).

This structure guides a discussion of how to finance excess withdrawals in a (not fully detailed) stochastic version of the model in the second part of the article. The new ingredients are as follows. Each period is divided into two subperiods, early and late. Type-1 individuals receive a privately observed preference shock that mandates consumption early in their second period of life with probability α, and learn about the realization of the shock after all markets close when young. Type-2 individuals receive their endowment of goods only late in their second period of life. Young type-1 individuals are the only ones who can supply goods early to the old who receive the liquidity shock. The next assumptions are that banks only receive proceeds from bond investments late and that trading deposit claims is prohibitively costly. As a result, the old who must consume early need money to buy goods from the young.

Because claims to deposits have no use for the old with liquidity needs, the need appears for the bank to keep money reserves in order to accommodate early withdrawals. Without aggregate uncertainty, the mass of such withdrawals is α and the framework is well-suited for predicting equilibrium outcomes. A new problem arises, however, when the (late) endowment of type-2 individuals becomes random. An excess redemption can now take place under the assumption that an additional fraction, β, of type-1 individuals does not draw the realization of early consumption but receives instead a signal about the quantity of the random endowment. In events when an early withdrawal is advantageous to the bank, a bank that has planned money reserves for α withdrawal requests will face withdrawals from a measure of $\alpha + \beta$ individuals. These events are called bank runs.

This model illustrates several important issues debated by the field. For instance, Bryant discusses how a random taxation scheme could ensure the real value of deposits against variations in loan repayments, given that type-2 individuals cannot share risk with type-1 individuals of the next generation. There is also the possibility of government money being created to cover excess redemptions in a form of deposit insurance that can dominate, for the impatient, a partial-suspension scheme. Another issue discussed is that, given high payoffs accruing to deviations, coordination of behavior around particular no-run outcomes is difficult to implement. Early redemption of bonds is yet another possible intervention to inject liquidity in the system.[7] One conclusion

[7] The kind of insurances discussed may not prevent runs. Another difficulty is the need to model how uninformed individuals learn about runs and, in particular, if early prices reveal runs for all individuals alive at the moment or just those holding money and making early purchases.

is that reactions by the private sector may partially offset redistributive effects of interventions. Another is that it may be desirable to occasionally let the government print money instead of inducing banks to store costly reserves.

2. THE NEW TRADITION

The transition from the models of SW and Bryant to that of DD outlined in the introduction can be motivated in many ways. A great deal of effort goes into simplifying the liquidity problem but maintaining a coherent description of the physical environment. Instead of focusing on the infinite horizon of Bryant or the static setting of SW, a natural choice becomes focusing on the two-period economy with a single resource constraint. This choice reveals an interest in the provision of insurance against privately observed shocks, as described by Bryant, but without his emphasis on monetary allocations. As I point out below, Green and Lin (2003) propose simplifying the setting even further by adopting a finite trader specification with important consequences. In some aspects, the bank in SW is a technology: Individuals in their economy do not have access to some goods without the assistance of the bank. This feature is less present in DD, where the central question becomes the level of insurance that the bank can provide. But the bank in SW, like the one in DD, is operated by a benevolent social planner. In contrast, bank deposits in Bryant have to compete in rate of return with outside money as alternative assets.

Bryant's point is that runs are distortions that undermine the desired insurance protection against preference shocks and are triggered by the attempts of informed individuals to gain on the uninformed through early purchases when prices are likely to be low. His discussion of how efficiency can be restored, perhaps partially, with the help of government interventions is limited by the class of contracts banks are offering in the first place. His view that withdrawals must be paid on a first-come, first-served basis is important to the concept of fragility. He also notes the second-best aspect of regulation: Suspension of payments or conversion of deposits to currency at a much-reduced rate would transfer too much of the burden to those with real liquidity needs.

In what has become a key feature of the DD model, the true state of nature is revealed partially and sequentially according to the volume of withdrawals. This property lends support to the now-accepted view that bank contracts have a second-best nature. Because optimality is not discussed formally in Bryant's model, it is unclear to what extent desirable allocations incorporate some financial fragility or if what is called a run is compensation for having access to early information.

Another distinguishing feature of the literature that starts with DD is the limited reference to devices that can regulate bank fragility. This is explained by the emphasis on efficient allocations and the consequent limits to the range of admissible policies and interventions. Although the discussion of deposit

insurance and partial suspension pointed out by Bryant made its way to the original DD setup, extensions focusing on aggregate uncertainty and sequential service skip considerations to such devices. Wallace (1988) was pivotal in stressing the importance of sequential service for the notion of fragility, as well as the need to study the contract that best deals with the imperfect flow of information and, thus, when deposit insurance is not feasible. Green and Lin (2003) successfully implement that agenda in a tractable version with finite traders. They rule out bank runs with specific assumptions about preferences and independency of shocks. Peck and Shell (2003) reinstate multiplicity by resorting to limited information about positions in the queue formed at the bank and to preferences leading to active truth-telling constraints.

In summary, the generality of key results about financial fragility have been revisited and is the subject of ongoing research. In order to provide some perspective on this research, I present below a particular formalization of the DD model that I find suitably short. Before doing that, let me stress some common elements of follow-up articles.

Four points on model choices can be highlighted. First and foremost, markets are no longer a primitive in the setup. The premise is that, if we are to find financial fragility as a result of asymmetric information, we should do so in the best way society can find to process information and to organize transfers while respecting individuals' incentives to truthfully reveal information. Imposing particular market organizations may hide better ways to organize exchanges.

Second, for reasons of tractability, it becomes important to abstract from payment instruments, like the use of money, the existence of which may depend on additional assumptions that would complicate the analysis. As a result, all the action in the model is restricted to a two-period structure in which a "bank" is a programmable technology that can commit to making transfers of real goods in these two periods according to announcements of preference shocks. In this sense, the bank becomes an aggregation of the intermediary, the productive sector, and the government.

Third, since it is not reasonable to assume that the government has more information about preferences than the individuals themselves, the bank machine is restricted to making transfers that are contingent on the flow of information provided by the requests to withdraw—the first-come, first-served structure of Bryant (1980). A combined bank-government, making payments in real goods, will then find it impossible to promote deposit insurance unless it can bypass the sequentiality of consumption (soliciting announcements first and then making payments only after collecting all answers). But if it can bypass sequentiality then a bank is never fragile. Thus, it becomes evident that sequential service must be taken as part of the physical environment. As Wallace (1988) discusses further, this observation moves the analysis definitively away from initial attempts of mixing banks and markets.

Fourth, once banking, production, and the benevolent government are aggregated into a mechanism that must respect sequential service, it follows that the best strategy for an individual is a function of the previous announcements made by those already serviced according to their position in line. If there is aggregate uncertainty about these requests then consumption is a random variable. One trivial case appears in the absence of aggregate uncertainty (as in Bryant [1980], in case there is no risk about future endowments and thus no inside information). The bank does not need to make payments contingent on line position and can suspend payments after the known fraction of impatient people has withdrawn. This scheme rules out any meaningful fragility.

3. SOME EXTENSIONS

In this section I provide a more formal description of the benchmark planner problem outlined in the introduction. This kind of specification has paved the way for very sharp results on the existence of multiple equilibria for the optimal deposit contract. Apart from the problem of runs, I shall also discuss how an element of imperfect monitoring, taking the form of delayed communication, sheds new light on the means of payment required to implement the optimum.

The Green-Lin Diamond-Dybvig Model

In the benchmark economy, bank runs can appear in the form of an early withdrawal by a patient individual concerned about the behavior of other patient individuals in the presence of aggregate uncertainty. Green and Lin (2003) demonstrate, however, that, with a finite number of individuals who receive shock realizations independently from each other, the solution of the optimal problem proposed by DD defines a deposit game that has the optimum as the unique equilibria. Green and Lin's demonstration relies on a class of preferences for which the optimum does not feature active truth-telling constraints. Their specification assumes simultaneous play: Individuals know their position in the queue but cannot choose strategies contingent on previous announcements.

Follow-up work produced at least three important results. First, Peck and Shell (2003) restored multiplicity with preferences that imply active truth-telling constraints when individuals are not informed about their position in line. Second, Ennis and Keister (2009b) return to the Green-Lin preferences and are able to provide a construction of the optimum explicitly even when shocks are correlated in a particular fashion. They shed light on the incentives for providing the informational externality alluded to in the introduction. In a run, the first individual in the queue announces his desire for early consumption even when truly patient. He thus fails to "inform" the planner about an important signal concerning the overall distribution of tastes when shocks are

correlated. The second individual in the queue, when patient, is concerned about the likelihood that many other individuals are impatient since the optimum was designed with truth-telling, that is, under the assumption of the best provision of the informational externality. The second individual thus cannot use the best conditional distribution for the preferences of others. As Ennis and Keister (2009b) find, the concern is justified as a bank-run equilibrium attains in some numerical examples.

Third, Andolfato, Nosal, and Wallace (2007) provide clarification of several points. They start by modifying the Green-Lin specification to include more general preferences and to let individuals know the announcements of others holding previous positions in the queue. They present a different demonstration of the result that, with independent shocks, individuals do not need the externality to predict the distribution of preferences of those ahead. The conditional distribution is the same regardless of whether or not initial traders have chosen to misrepresent their types. Thus, that distribution is the same as the one used to define truth-telling constraints under the assumption that all individuals reveal their type. Since the optimum is constructed so as to respect those constraints, telling the truth is a maximizing choice even when initial players choose differently. Since the population is finite, a backward induction argument can be used to demonstrate uniqueness.

Andolfato, Nosal, and Wallace (2007) also raise questions about signalling in this setup. Although the formulation adopted by Green-Lin and Ennis-Keister has simultaneous play, it is not clear whether or not this specification is necessarily mandated by the DD environment. This issue is relevant since sequentiality of plays creates the opportunity for individuals to signal their true type. In addition, if beliefs are required to satisfy the intuitive criterion of Cho and Kreps (1987), among others, then it can be argued that the run equilibria do not survive the refinement: Given reasonable beliefs, a patient individual has no conflict of interest in providing the informational externality because his announcement reinforces truth-telling, which tends to save resources for the future (relative to runs). In summary, if the planner can inform individuals about announcements made by others, and if doing so eliminates fragility, then there are grounds for taking the sequential formulation of Andolfato, Nosal, and Wallace (2007) as a welcome improvement.[8]

Based on the arguments above, including the footnote, I take Andolfato, Nosal, and Wallace (2007) as a reasonable formulation of what is called the Green-Lin Diamond-Dybvig model, presented below. It is important to keep in mind, however, that a great deal about runs and their relationship to the

[8] Truth-telling constraints change when play is simultaneous. While the examples constructed in Ennis and Keister (2009b) feature inactive truth-telling constraints, a separating equilibrium may not satisfy constraints in the version with sequential play. But if it does satisfy constraints, say for some parameters, then the models would be observationally equivalent.

provisions of information and shock structures can be learned from the models of Peck-Shell and Ennis-Keister.

The problem considered by Andolfato, Nosal, and Wallace is more easily presented by referencing their deterministic case, ignoring possible welfare improvements associated with the use of lotteries. There is one consumption good for each of two dates and the number of individuals is N. There is an aggregate endowment of Y units of good 1 and a linear technology that transforms x units of good 1 into R units of good 2 ("rate of return," $R-1$). Each individual becomes type $t \in T$, where $T = \{ i$ (impatient), p (patient)$\}$, and is assigned utility $u(c_1, c_2, t)$ according to a stochastic process and the mechanism assigning announcements to consumption bundles (c_1, c_2). Each individual maximizes expected utility. The environment also includes a random process that determines the vector, $\mathbf{t}$, representing the queue $(t_1, t_2, .., t_N)$, with the understanding that types are private information and t_i is the type of the jth individual in line. The stochastic process for $\mathbf{t}$ is specified by the probability measure $\pi = (\pi_0, \pi_1, ..., \pi_N)$ that describes the distribution of realizations of the total number of patient people, considering that all permutations determining place in line are equally likely. The draw of k patient individuals occurs with probability π_k.

Allocations are allowed to depend on positions in line and announcements, but are otherwise symmetric regarding identities. The mechanism reveals to each individual earlier announcements (so that the quantity of resources left is easily inferred). A strategy for an individual with place n is a function, $\mathbf{s}_n$, mapping $T^{n-1} \times T$ into announcements of types in T. The second argument is his or her true type, while the first is the vector of earlier announcements. A mechanism is a mapping (c_1^n, c_2^n) for each n, where c_1^n maps an announcement list of size n in T^n into good-1 consumption, and c_2^n maps each list, $\mathbf{t}$, into good-2 consumption, when all announcements become known.

Associated with mechanism $\mathbf{c}$, representing (c_1^n, c_2^n) for each n, and a strategy profile $\mathbf{s} = (\mathbf{s}_1, ..., \mathbf{s}_N)$, there corresponds an ex-ante expected utility

$$w(\mathbf{c}, \mathbf{s}) = \sum_{k,\mathbf{t},n} \pi_k \binom{N}{k}^{-1} u(c_1^n(\mathbf{s}_n), c_2^n(\mathbf{s}_N), t_n). \tag{1}$$

The planner's problem is to choose $\mathbf{c}$ in order to maximize ex-ante utility (1) under truth-telling, $w(\mathbf{c}, \mathbf{t})$, subject to feasibility,

$$R(Y - \sum_n c_1^n) \geq \sum_n c_2^n, \tag{2}$$

and truth-telling constraints. The latter are written according to beliefs about $\mathbf{t}_{n+1}$, the vector of types of those in line after n. For an individual, n, with type t_n, the probability of outcome $\mathbf{t}_{n+1}$ conditional on his or her type, as well as on earlier announcements, defines belief $\phi(\mathbf{t}_{n+1}; \mathbf{s}_{n-1}, t_n)$. The truth-telling constraint for individual n experiencing line history $\mathbf{t}_n$, including own type t,

is

$$\sum_{\mathbf{t}_{n+1}} \phi(\mathbf{t}_{n+1}; \mathbf{t}_n) u(c_1^n(\mathbf{t}_n), c_2^n(\mathbf{t}_n, \mathbf{t}_{n+1}), t) \geq$$
$$\sum_{\mathbf{t}_{n+1}} \phi(\mathbf{t}_{n+1}; \mathbf{t}_n) u(c_1^n(\mathbf{t}_{n-1}, s), c_2^n(\mathbf{t}_{n-1}, s, \mathbf{t}_{n+1}), t) \quad (3)$$

for all $s \in T$.

It is required that ϕ be consistent with Bayes' rule. When shocks are independent across individuals, $\phi(\mathbf{t}_{n+1}; \mathbf{t}_n)$ is constant in $\mathbf{t}_n$. This allows the induction argument that rules out multiplicity.

Green and Lin (2003) also present a dynamic programming problem that corresponds to a relaxed planning problem that follows from ignoring truth-telling constraints. Green and Lin (2000) explicitly solve this problem for a class of preferences with linear indifference curves and three traders. The class of preferences is such that the constraint (3) does not bind when other individuals follow truth-telling strategies and the solution of the problem is the optimal allocation. As hinted above, Ennis and Keister (2009b) generalize their programming problem for an arbitrary number of traders and a particular structure of correlated shocks: The probability that a person in position n is patient is a function of the number of previous patient draws, not of their order among the $n - 1$ people. They are able to construct a solution recursively and to show, by means of examples, that the associated mechanism can also implement an equilibrium with misrepresentation (run).

Imperfect Monitoring

The formal emphasis on optimality pursued by DD on the issue of bank illiquidity has of course initiated many developments in the field that cannot be covered here. Alternative formulations of aggregate and extrinsic uncertainty have been proposed.[9] Weakening of the ability of the bank to commit has been pursued in fruitful ways.[10]

Prescott and Weinberg (2003, PW hereafter), for instance, propose a new extension. They compare two payment instruments, in a version of DD without aggregate uncertainty, where the use of payment devices can be distorted by opportunistic behavior. If we ignore the initial planning period, there are two dates and one consumption good per date in their model. The counterparts in their model for the DD consumers ("buyers" in their language) have preferences $tu(c_1) + v(c_2)$, where c_1 is consumption of date-1 good, c_2 is

[9] See Hellwigh (1994) for the case of stochastic last-period endowments.

[10] See Ennis and Keister (2009a) for a re-examination of suspension schemes when the planner cannot commit.

consumption of date-2 good, $v(0) = 0$, and t is a preference shock defining types and takes values in a finite set.

PW depart from the direct contact between consumers and the DD bank machine. Their basic goal is to compare the performance of bank drafts and checks as communication devices between consumers and the machine (the planner). The typical DD allocation could be implemented with the use of drafts, that is, pieces of paper that communicate that the buyer has funds available to make a purchase. If enough multiple drafts, with pre-set amounts are initially distributed to consumers, then no noise can occur in the communication with the machine. A problem appears because drafts are assumed to be costly to produce.

Checks are an alternative communication device that can be produced and distributed at zero cost. Check technology is, however, subject to fraud. There are two ways to describe the communication problem produced by checks. In the version detailed by PW, some consumers are randomly able to use checks with many sellers in a way that multiple purchases temporarily become private information to chosen individuals.

Because sellers are agents with trivial choices in their model, there is another description of the communication inefficiency. In this alternative version, consumers contact the bank machine directly and write checks according to their privately observed type, t. Consumers, however, also draw an opportunity to re-enter the bank line and make new withdrawal requests. The fraud of entering in line multiple times is only detected in period 2. Because $v(0) = 0$, there is limited punishment that can be imposed on period 2. As a result, optimality requires the imposition of an upper bound on the values that consumers can write on their checks.

PW restrict attention to symmetric allocations in the sense that consumers with different realizations of fraud opportunities are treated equally. A similar outcome could attain under the assumption that even a small fraud is too costly for society. Hence, the model is used to predict allocations under the *threat* of fraud (or "bingeing").[11]

The framework allows for differentiated initial individual wealth, w. In the simpler case, w is public information and a bank is formed to deal exclusively with "population w" in isolation. That is, w becomes a parameter for the comparative statics predicting the use of drafts or checks in that population.

Assuming the existence of a continuum of consumers, an allocation is a pair of functions $(c_1(t, w), c_2(t, w))$ defined on the Cartesian product of the set of types and the set of initial endowments. Feasibility requires that aggregate

[11] Cavalcanti and Nosal (2010) endow to a subset of pairwise traders a particularly low cost of falsifying money, finding optima with a positive mass of counterfeits. A monetary model with closer links to the PW idea would, however, be the model of counterfeiting threats of Nosal and Wallace (2007).

endowment finances expect total consumption $c_1(t, w) + c_2(t, w)$ on a linear basis, net of the costs of using drafts.

When w is public information, the truth-telling constraints are common to both draft and check economies,

$$tu(c_1(t, w)) + v(c_2(t, w)) \geq tu(c_1(t', w)) + v(c_2(t', w))$$

for all (t, t'), but checks require a new constraint, because of the upper bound referred to above,

$$tu(c_1(t, w)) + v(c_2(t, w)) \geq tu(d(w)).$$

The right-hand side is the deviation payoff that follows from the worst fraud of appearing in line N times (consuming from N sellers, instead of one) and announcing the highest discount factor at each time, so that the property, d, of an allocation is defined by

$$d(w) = N \max_t c_1(t, w).$$

Two distortions implied by the upper-bound constraint are easily seen. First-period consumption in a check economy cannot vary with t as much as in a draft economy, otherwise an undesirable increase in payoff d is produced. Also, by moving from a draft economy (d identically zero) to a check economy, the upper-bound constraint becomes tighter and more consumption on date 2 needs to be allocated, further reducing the bank's ability to insure risk. PW also deliver results about the choice of payments as individuals become wealthier. Since, in their setup, wealthy individuals plan higher levels of date-2 consumption, there is a sufficient increase of punishment to fraud in date-2 ($c_2 = 0$) so as to increase check limits as w grows. Thus, the tendency is to shift from drafts to checks with increases in wealth. It is clear that many more payment questions can be asked in this line of research that managed to stay so close to the original DD model.

4. BANKS AS PEOPLE

There is no need to limit attention to banks that can be easily programmed to perform intermediation duties. Important progress has been made by models that give banks incentive constraints. As we shall see, this has become a useful device for introducing banking in less centralized environments where money plays a role of medium of exchange.

Endogenous Sequential Service

Calomiris and Kahn (1991) propose treating banks as individuals with commitment difficulties and who can hide resources. Fraud outcomes can be mitigated by investments on information acquisition by depositors, as well as

by the employment of an additional technology that removes control of the bank's assets at a cost. Under the interpretation that the use of this technology corresponds to an early withdrawal or liquidation, they conclude that optima require bank liabilities in the form of demand deposits.[12]

The approach in Calomiris and Kahn (1991) leads to models in which sequential service is not an ingredient for explanations of fragility as in DD but is instead an equilibrium feature that exists to discipline banks. One is led to the conclusion that there appears to exist a choice between studying the DD view of fragility when the emphasis is on the behavior of other depositors and the costly state-verification model that applies not only to banking but also to optimal contracts in abstract principal-agent problems.[13]

While there are many ways to introduce elements of fraud in economic models, it is also important to work with abstractions that capture the essence of what financial markets do. SW and Bryant have pointed to basic issues: attempts to understand the co-existence between money and credit or between banks and payment instruments. One can return to these fundamental issues knowing that future developments can always add further considerations about fraud.

Money and Credit

The agenda of making explicit the role of money with endogenous supply borrows a great deal from mechanism design. The influence of DD on this agenda can be illustrated by the importance given in the monetary literature to the concept of the essentiality of money. The analogous question in the DD model is whether or not banking arrangements are indeed fragile. It is important to rule out other arrangements, such as deposit insurance or partial suspension, that can provide the same levels of utility as in the optimum but without being exposed to multiplicity. By looking for a physical environment with such properties, one is identifying primitives that give rise to the notion of financial fragility. As discussed above, sequential service has been identified as a necessary friction, although not a sufficient one, for the fragility feature. When runs are present, sequential service is one of the conditions that make fragility *essential* in the DD model (without fragile allocations welfare would not be maximized).

Likewise, in monetary theory, one is always looking for conditions that make money essential. It has been shown that imperfect monitoring and the

---

[12] Another advantage of fragility, pursued by Diamond and Rajan (2001), is the notion that when creditors can commit to a run, the bank is encouraged to monitor borrowers. There is also a link in their work to the human capital interpretation of collection technologies emphasized by Kiyotaki and Moore (2000).

[13] See Townsend (1979), Diamond (1984), and Dewatripont and Tirole (1994), including references in the latter to models of entrepreneurial banks.

absence of commitment to future actions are necessary conditions to rule out trigger strategies and other credit arrangements that can substitute for money. Like sequential service, those conditions are frictions that impede smooth operations of markets and call into question the validity of dichotomies alluded to in Section 1. In this sense, money and banking have a lot in common and should, perhaps, be studied in a unified framework.

Notice, however, that the strategy in early work on models of medium of exchange, influenced by this essentiality reasoning, was to adopt a strong form of imperfect monitoring: total anonymity. Although anonymity preserves money, it rules out all forms of credit. A natural step in this literature was to look for weaker primitives that make the use of money essential in trade but facilitate credit in ways that do not eliminate money from the optimum. That strategy has the chance of producing banks according to their interaction with money-creation mechanisms.[14]

Such a link between money and banking appears in Deviatov and Wallace (2009, DW hereafter), provided that we interpret their monitored individuals as banks. Their mechanism is a representation of a central bank that can inject and destroy money in a particular fashion over seasons. Individuals are prohibited from creating money (notes) in their computed examples, but extending the model to an inside-money version is conceptually straightforward.

The model neither assumes that all individuals accomplish intertemporal trade based on announcements nor that they are all anonymous (like Kiyotaki and Wright [1989] or Levine [1991]).[15] It assumes instead that an exogenous fraction of the population is perfectly monitored (the m people) and that the remainder is not monitored at all (the n people). Only the planner can commit to future actions.[16]

In other respects, DW build on a typical random-matching model of money with seasons (it is useful to think of random matching as a restriction on the physical movement of goods between people).[17] The continuum of people rules out aggregate uncertainty, the horizon is infinite, the common discount factor is $\beta \in (0, 1)$, and goods are perishable. There is a symmetric division of people according to the goods they like and produce. There are two stages

[14] Another avenue is the study of how sequential service would change established *real* models of intertemporal trade. The problem solved by Green (1987), for instance, predicts transfers across a continuum of individuals facing privately observed endowment shocks. The optimal allocation could be considered an illiquid one and it is not known how sequential service would change the predictions.

[15] The environments compared in Kocherlakota (1998) are also extreme cases.

[16] The setup follows from Cavalcanti and Wallace (1999a, 1999b, 2008). Cavalcanti (2004) considers a version in which banks store capital and take announcements from the nonbank public. Cavalcanti, Erosa, and Temzelides (1999, 2005) study an equilibrium version in which banks are monitored by the creation and redemption of notes against reserves.

[17] See Cavalcanti and Nosal (2009) for a random-matching model with seasons in which the optimum requires ongoing interventions in the money supply because money can get "stuck in the wrong hands."

at each date. The first stage has pairwise meetings and the second stage has a centralized meeting. In the first stage an individual meets randomly a producer with probability $1/K$ or a consumer of the good he can produce with probability $1/K$ (no relevant meeting with probability $1 - 2/K$). The period utility from consuming q units of the desired good is $u(q)$. The period utility from producing q units for a capable producer is $-q/\delta_t$, where the productivity parameter moves seasonally: $\delta_t = \delta_l$ at odd dates (low aggregate productivity) and $\delta_t = \delta_h$ at even dates (high aggregate productivity). The monitored status and the consumer-producer status in meetings are common knowledge. Holdings of money are private for n people and observable for m people.

An n person must receive money in order to consume. For simplicity, money holdings are assumed indivisible and restricted to $\{0, 1\}$ at stage 1 and to $\{0, 1, 2\}$ at stage 2. An n person with money is thus so rich that he cannot be induced to produce, creating a nontrivial problem of liquidity distribution for this economy (he cannot lend his money in period 2 because the lending action of a person n cannot be recorded). There is, however, a need to arrange borrowing and lending among the m people, which can be done at stage 2.

DW study allocations that are two-date periodic (stationary) and that treat the same people in the same state: a point in $\{m, n\} \times \{0, 1\}$ at the beginning of stage 1 and a point in $\{m, n\} \times \{0, 1, 2\}$ at the beginning of stage 2. An allocation describes trade in output and money in stage 1, according to states and season. It also describes transfers of money in stage 2 and the fraction of each type who has money at the start of a season.

DW use lotteries to model transfers of money because of indivisibilities and consider that m people can be punished with banishment to the set of n people, although that never happens in equilibrium. Only individual defection is allowed in stage 2 but deviation by the pair is also allowed in stage 1. Since the holdings of n people are private information, an allocation must propose a menu of trades to which they self-select. The objective of the planner is average expected utility, assuming that the initial date has low productivity.

The optimum is computed numerically for an arbitrary example with high discount factor. DW find that the optimum has injection of money (stage 2) at dates of high productivity. There are no stage-2 transfers of money to n people (in order to preserve incentives for them to acquire money). Put another way, m people spend more than they earn at high dates (the opposite for low dates). DW also find that m people always start a date with money, and thus the threat of punishment is important to force them to produce. Computed welfare indicates that the main beneficiaries of the money interventions are the n people. Type-m people can be instructed to give gifts, that is, to produce to n people without money.

The seasonal monetary policy provides smoother output in meetings where the producer is type n (and has no money) and the consumer is type m (and

has money). The computed lotteries in those meetings predict higher spending during the high season compensated by injections of money that increase the fractions holding money. The injection must be offset by destruction during low seasons in order to preserve stationarity.

Although the model is formulated so that strong stationarity assumptions deliver a small state space, it documents that general principles about the nature of the optimum are difficult to anticipate, in contrast, for instance, to old-style monetary models for which one can guess that the Friedman rule is optimal from the start. DW conjecture that this difficulty is here to stay, at least if one wants to preserve a model capable of addressing the circulation of private bank notes.

5. FINAL REMARKS

In this article I compare the approach of Diamond and Dybvig with some other influential work in the field of money and banking. It is natural for reviews of this topic to mention previous attempts to mix general-equilibrium theory with banking. Having presented reasons for avoiding the old strategy of mixing banks and markets, one also has to explain why banking models seem so distant from monetary theory. I provide a unifying explanation. The DD approach has been successful in its choice for mechanism design. Without it, conclusions about fragility would be controversial. One has to give serious consideration to the possibility that financial fragility is an intrinsic consequence of trying to extend to individuals the best liquidity arrangements possible. This favors mechanism design over general-equilibrium theory. In addition, banking models become closer to monetary theory when monitoring becomes weaker.

There are, of course, negative aspects to consider. By posing a well-specified model of liquidity provision in which mechanism design and new formalizations of sequential service could be easily adopted, DD depart from the macroeconomic tradition of staying close to the data in the way it is traditionally presented, that is, with a great deal of reference to market statistics like rates of returns, and instead focus on payoffs accruing to individuals.

The dichotomy-tradition alternative is reflected in the way most textbooks in money and banking are organized.[18] Empirical observations about the functions performed by financial systems motivate "extensions" of the competitive-equilibrium model so that they incorporate payment services, risk management, collateral arrangements, etc. In this tradition, the DD model falls in the chapter of liquidity risk, which is to say it is supposed to provide guidelines

[18] See, for instance, Niehans (1978) and Freixas and Rochet (1997).

about deposit insurance or suspension of bank payments in contingencies of instability.[19] I have discussed concerns about this tradition on many grounds.

It is, however, difficult to predict how macroeconomics will incorporate lessons provided by the Diamond-Dybvig structure and other models of inside money. Presently, a variety of approaches are currently being considered. The validity of building market games with weak commitment assumptions, an approach dating back to the pioneering work of Shubik, is as debatable now as it was 30 years ago. For the moment, much can be learned from understanding that the assumptions that make banks important in the Diamond-Dybvig model are not very far from those that make inside money important in exchange models.

REFERENCES

Amir, Rabah, Robert J. Aumann, James Peck, and Myrna Wooders. 2009. "Introduction to the Special Issue of *Games and Economic Behavior* in Honor of Martin Shubik." *Games and Economic Behavior* 65 (January): 1–6.

Andolfato, David, Ed Nosal, and Neil Wallace. 2007. "The Role of Independence in the Green-Lin Diamond-Dybvig Model." *Journal of Economic Theory* 137 (November): 709–15.

Bryant, John. 1980. "A Model of Reserves, Bank Runs and Deposit Insurance." *Journal of Banking and Finance* 4 (December): 335–44.

Calomiris, Charles W., and Charles M. Kahn. 1991. "The Role of Demandable Debt in Structuring Optimal Banking Arrangements." *American Economic Review* 81 (June): 497–513.

Cavalcanti, Ricardo. 2004. "A Monetary Mechanism for Sharing Capital: Diamond and Dybvig Meet Kiyotaki and Wright." *Economic Theory* 24 (November): 769–88.

Cavalcanti, Ricardo, Andrés Erosa, and Ted Temzelides. 1999. "Private Money and Reserve Management in a Random-Matching Model." *Journal of Political Economy* 107 (October): 929–45.

[19] Some authors believe that an "asymmetric paradigm" can handle banks just as well as it can handle managers and firms. These partial-equilibrium models would be very difficult to blend with monetary models.

Cavalcanti, Ricardo, Andrés Erosa, and Ted Temzelides. 2005. "Liquidity, Money Creation and Destruction, and the Returns to Banking." *International Economic Review* 46: 675–706.

Cavalcanti, Ricardo, and Ed Nosal. 2009. "Some Benefits of Cyclical Monetary Policy." *Economic Theory* 39 (May): 195–216.

Cavalcanti, Ricardo, and Ed Nosal. 2010. "Counterfeiting as Private Money in Mechanism Design." JMCB conference mimeo.

Cavalcanti, Ricardo, and Neil Wallace. 1999a. "Inside and Outside Money as Alternative Media of Exchange." *Journal of Money, Credit and Banking* 31 (August): 443–57.

Cavalcanti, Ricardo, and Neil Wallace. 1999b. "A Model of Private Bank-Note Issue." *Review of Economic Dynamics* 2 (January): 104–36.

Cavalcanti, Ricardo, and Neil Wallace. 2008. "New Models of Old(?) Payment Questions." In *The Future of Payment Systems*, edited by Andrew G. Haldane, Stephen Millard, and Victoria Saporta. New York: Routledge, 75–86.

Cho, In-Koo, and David M. Kreps. 1987. "Signaling Games and Stable Equilibria." *Quarterly Journal of Economics* 102 (May): 179–221.

Deviatov, Alexei, and Neil Wallace. 2009. "A Model in which Monetary Policy is about Money." *Journal of Monetary Economics* 56 (April): 283–8.

Dewatripont, Mathias, and Jean Tirole. 1994. *The Prudential Regulation of Banks*. London: The MIT Press.

Diamond, Douglas W. 1984. "Financial Intermediation and Delegated Monitoring." *Review of Economic Studies* 51 (July): 393–414.

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

Diamond, Douglas W., and Raghuram G. Rajan. 2001. "Liquidity Risk, Liquidity Creation and Financial Fragility: A Theory of Banking." *Journal of Political Economy* 109 (April): 287–327.

Dubey, Pradeep, John Geanakoplos, and Martin Shubik. 2005. "Default and Punishment in General Equilibrium." *Econometrica* 73: 1–37.

Ennis, Huberto, and Todd Keister. 2009a. "Bank Runs and Institutions: The Perils of Intervention." *American Economic Review* 99 (September): 1,588–607.

Ennis, Huberto, and Todd Keister. 2009b. "Run Equilibria in the Green-Lin Model of Financial Intermediation." *Journal of Economic Theory* 144 (September): 1,996–2,020.

Freixas, Xavier, and Jean-Charles Rochet. 1997. *Microeconomics of Banking*. London: The MIT Press.

Green, Edward J. 1987. "Lending and the Smoothing of Uninsurable Income." In *Contractual Arrangements for Intertemporal Trade*, edited by Edward C. Prescott and Neil Wallace. Minneapolis: University of Minnesota Press, 3–25.

Green, Edward J., and Ping Lin. 2000. "Diamond and Dybvig's Classical Theory of Financial Intermediation: What's Missing?" Federal Reserve Bank of Minneapolis *Quarterly Review* 24: 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.

Hellwig, Martin. 1994. "Liquidity Provision, Banking, and the Allocation of Interest Rate Risk." *European Economic Review* 38 (August): 1,363–89.

Kiyotaki, Nobuhiro, and John Moore. 2000. "Inside Money and Liquidity." London School of Economics mimeo.

Kiyotaki, Nobuhiro, and Randall Wright. 1989. "On Money as a Medium of Exchange." *Journal of Political Economy* 97 (August): 927–54.

Kocherlakota, Narayana. 1998. "Money is Memory." *Journal of Economic Theory* 81 (August): 232–51.

Levine, David K. 1991. "Asset Trading Mechanisms and Expansionary Policy." *Journal of Economic Theory* 54 (June): 148–64.

Niehans, Jurg. 1978. *The Theory of Money*. Baltimore, Md.: The John Hopkins University Press.

Nosal, Ed, and Neil Wallace. 2007. "A Model of (the Threat of) Counterfeiting." *Journal of Monetary Economics* 54 (May): 994–1,001.

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

Prescott, Edward S., and John A. Weinberg. 2003. "Incentives, Communication, and Payment Instruments." *Journal of Monetary Economics* 50 (March): 433–54.

Shubik, Martin, and Charles Wilson. 1977. "The Optimal Bankruptcy Rule in a Trading Economy using Fiat Money." *Journal of Economics* 37 (September): 337–54.

Townsend, Robert M. 1979. "Optimal Contracts and Competitive Markets with Costly State Verification." *Journal of Economic Theory* 21 (October): 265–93.

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 96, Number 1—First Quarter 2010—Pages 83–122

Monetary Theory and Electronic Money: Reflections on the Kenyan Experience

William Jack, Tavneet Suri, and Robert Townsend

In 2007, the leading cell phone company in Kenya, Safaricom Ltd., launched M-PESA, a short message service (SMS)-based money transfer system that allows individuals to deposit, send, and withdraw funds from a virtual account on their cell phones and that is separate from the banking system. M-PESA has grown rapidly, currently reaching more than seven million users, approximately 38 percent of Kenya's adult population, and it is widely viewed as a success story to be emulated across the developing world. Indeed, similar products have recently been launched in a growing number of countries across Africa, Asia, and Latin America, with the intent of expanding financial services to previously unreached populations.[1]

M-PESA is used not only for remittance purposes, but also to save, to purchase pre-paid phone credit and other goods and services, to pay bills, and to execute bank account transactions. However, consumers do not need bank accounts in order to use M-PESA, and Jack and Suri (2009) found it was used by more than half of the unbanked in their sample. It is used by a broad cross section of Kenyan society, but has increasingly been adopted by those at the

William Jack is at the Department of Economics, Georgetown University; Tavneet Suri is at MIT Sloan School of Management; and Robert Townsend is at the Department of Economics, MIT. The authors would like to thank the editor, Ned Prescott; the referees, Huberto M. Ennis and Kartik Athreya; and Ignacio Mas, Carlos Perez-Verdia, and Tom Sargent for detailed comments and feedback. This research is supported by the Gates Foundation through the University of Chicago Consortium on Financial Systems and Poverty. The views expressed in this article are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mails: wgj@georgetown.edu; tavneet@mit.edu; rtownsen@mit.edu.

[1] For example, M-PESA is similar to the service called G-CASH in the Philippines.

lower end of the income distribution, as is evidenced by the steady reduction in the average transaction size since its inception. A large part of M-PESA's success is attributed to the broad and dense network of over 16,000 agents across Kenya, which provides the retail interface with consumers.

In this article, we examine the role of monetary theory in understanding this new generation of mobile banking products, especially those that, like M-PESA, do not simply provide electronic access to existing bank accounts. Deposits of money in a mobile phone-based account reflect holdings by the account owner of a commodity we refer to as e-money. Because e-money can be easily transferred from one individual to another, as long as it is expected to retain its value, it can be used in equilibrium as a means of exchange, as well as to transfer purchasing power between individuals.

Indeed, at the time of the launch of M-PESA, the service was seen as a means of overcoming the high transactions costs associated with sending cash remittances that faced the 80 percent of individuals in the economy without bank accounts.[2] But since then, e-money has been increasingly used both as a store of value and as a means of exchange, with users able to pay utility bills, make loan repayments, and even pay for taxi rides with it. The co-existence of essentially two forms of cash, even if closely related and linked, raises certain theoretical modeling issues in itself. But when one form of cash is issued by a profit-maximizing entity and the other by the central bank, further issues of competition, regulation, and coordination naturally emerge.

Although mobile banking is in its infancy in the United States, payroll cards have provided a similar payment function, albeit without the geographic reach of mobile phone communication. Funds are typically deposited by an employer into the account of an employee, who can either withdraw cash at an automated teller machine (ATM) or use the card to make purchases at stores possessing debit card machines. As in the case of mobile banking, payroll card users do not need a bank account, and transactions are executed using an existing communication network. In addition, payroll cards in the United States are generally cheaper than check-cashing services and money orders, just as M-PESA in Kenya is cheaper than most alternatives. Foster et al. (2010) describe the use of various payments systems in the United States—they find that 93.4 percent of consumers in the United States have adopted a payment card, but only 17.2 percent have a prepaid card.

This article presents a first look at how existing models of monetary theory can be used to think about the impact of mobile banking on the operations of the financial system and the implications for monetary and regulatory policy

[2] The original pilot program, supported by the U.K. government and the mobile phone provider Vodafone, was aimed at increasing the efficiency of microfinance products by allowing borrowers to make repayments more easily. However, by the time of the full launch, the focus had shifted to facilitating the sending of remittances more generally.

decisions that face the central bank. We are not yet in a position to develop a fully articulated model of mobile banking, but we hope this discussion will be a first step in this process. In addition, this article is not an exhaustive discussion of all models of money, but more of a focus on a subset of models that have different implications for the role of e-money in an economy.

Most theoretical models of money and credit include both a temporal dimension and some kind of generalized locational heterogeneity. Sequential trades over time require promises to be made (and kept) and records to be maintained. On the other hand, spatial separation can mean that it is not always possible for two parties to a trade to meet each other at the right time, so more complicated multilateral chains of individuals are required to effect the desired net trades.

In these environments, financial instruments such as fiat money and private debt can sometimes improve the efficiency of resource allocations by facilitating intertemporal and interspatial trades. However, equilibrium allocations may continue to be inefficient without the intervention of either a public institution (such as a central bank) or a well-regulated private agent (such as a clearinghouse).

Mobile banking has the potential to effectively reduce the distances that separate individuals, both literally and figuratively, thereby lessening the frictions that characterize models of incomplete intermediation, relaxing liquidity constraints, and reducing the need for monetary interventions. On the other hand, new liquidity constraints could arise that are binding for individuals who trade with the new financial instrument, e-money.

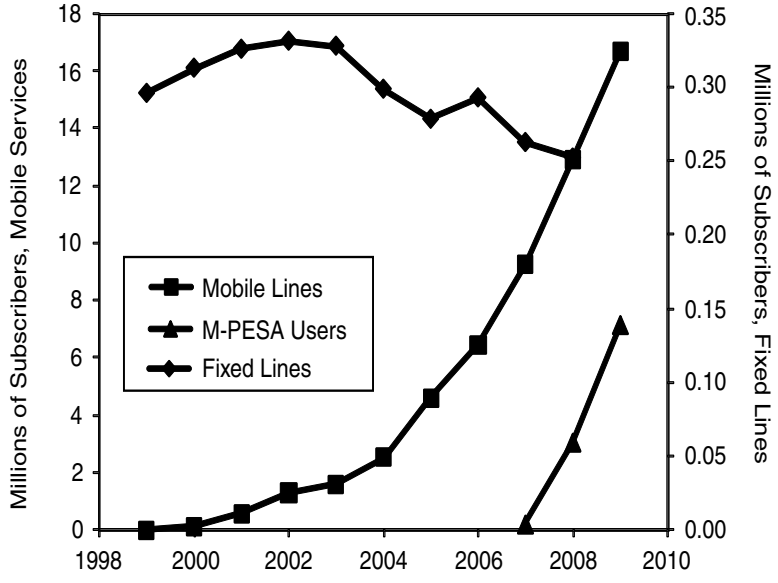
The article proceeds as follows: Section 1, which draws heavily on Jack and Suri (2009), provides background on the recent evolution of mobile technology and mobile banking in Kenya and on the practical operational features of M-PESA. Section 2 reviews a number of strands of the literature and discusses the specific lessons that we might learn regarding both the equilibrium impact of mobile banking and its implications for policy. Section 3 presents some empirical facts from a survey on M-PESA customers and agents that provide some insights into the implications from the models and lessons in Section 2. Section 4 concludes.

1. BACKGROUND ON M-PESA

Mobile Money in Kenya: An Introduction

Mobile phone technology has reduced communication costs in many parts of the developing world from prohibitive levels to amounts that are, in comparison, virtually trivial. Nowhere has this transformation been as acute as in sub-Saharan Africa, where networks of both fixed line communication and physical transportation infrastructure are often inadequate, unreliable, and dilapidated. While mobile phone calling rates remain high by world standards,

Figure 1 Phone Use in Kenya



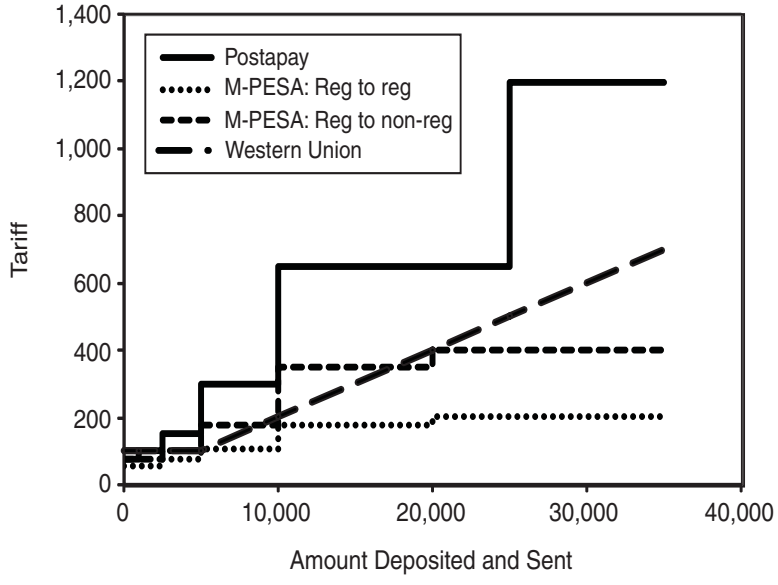
the technology has allowed millions of Africans to leap-frog the landline en route to 21st century connectivity. As the number of landlines in Kenya fell from about 300,000 in 1999 to around 250,000 by 2008, mobile phone subscriptions increased from virtually zero to nearly 17 million over the same time period (Figure 1). Assuming an individual has at most one cell phone,[3] 47 percent of the population, or fully 83 percent of the population 15 years and older, have access to mobile phone technology. In March 2007, following a donor-funded pilot project, Safaricom launched a new mobile phone-based payment and money transfer service, known as M-PESA.[4] The service allows users to deposit money into accounts linked to their cell phones, to send balances using SMS technology to other users (including sellers of goods and services), and to redeem deposits for regular money. Charges, deducted from users' accounts, are levied when e-money is sent and when cash is withdrawn.[5]

[3] This is not quite true, as some individuals own two (or more) phones so as to take advantage of the different tariff policies of competing providers.

[4] Pesa is Kiswahili for "money"—hence M[obile]-Money. A second mobile banking service called ZAP has since been launched, operated by Zain, the second largest mobile phone operator in Kenya. ZAP's market share remains very small at this point in time.

[5] The marginal cost of depositing and sending money is very low. These fees cover the costs of maintaining and expanding the agent network and physical infrastructure, marketing, and profits.

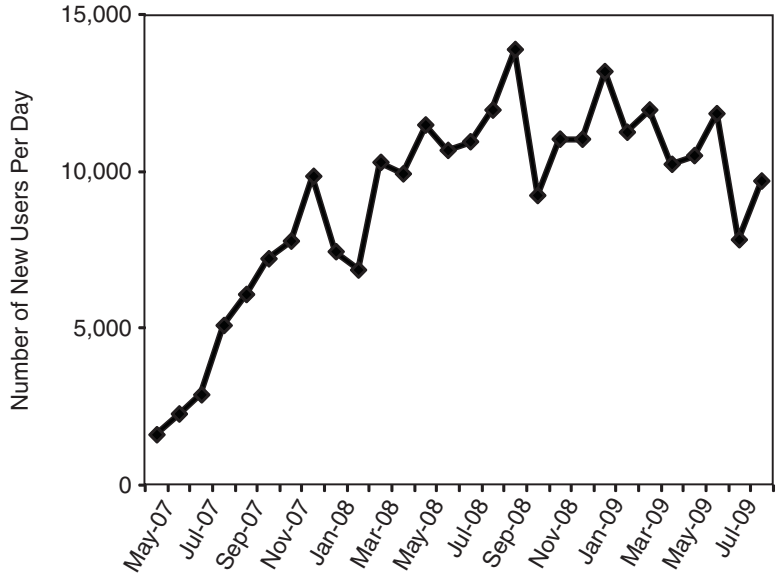
Figure 2 Total Net Tariff Rates for Depositing and Sending Money by the Post Office and by M-PESA to a Registered User and to a Nonregistered User



In particular, Safaricom accepts deposits of cash from customers with a Safaricom cell phone SIM (subscriber identity module) card and who have registered with Safaricom as M-PESA users. Registration is simple, requiring an official form of identification (typically the national ID card held by all Kenyans, or a passport) but none of the other validation documents that are typically necessary when a bank account is opened. Formally, in exchange for cash deposits, Safaricom issues a commodity known as "e-money," measured in the same units as money (denominated in shillings), which is held in an account under the user's name. This account is operated and managed by M-PESA and records the quantity of e-money owned by a customer at a given time. There is no charge to a customer for depositing funds into his account, but a sliding tariff is levied on withdrawals from M-PESA accounts (for example, the cost of withdrawing $100 is about $1).[6] An M-PESA user who sends e-money is charged a flat fee of about 40 U.S. cents if sending to another registered user, and a sliding fee if sending to a phone number that

[6] The complete tariff schedule is available at http://www.safaricom.co.ke/fileadmin/template/main/downloads/Mpesa_forms/14th%20Tariff%20Poster%20new.pdf

Figure 3 Average Daily Growth in M-PESA Registrations by Month

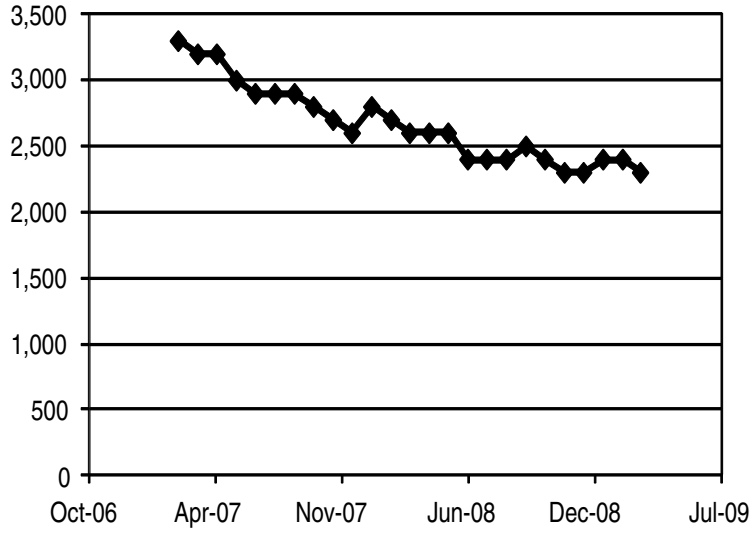


is not registered for M-PESA.[7] Figure 2 illustrates the schedules of total net tariffs for sending money by M-PESA, including the cost of withdrawing the funds incurred by the recipient, and compares these with two other money transfer services—Western Union and Postapay (operated by the Post Office). The M-PESA tariffs shown include both the sending and withdrawal fees and are differentiated according to receipt by registered and nonregistered user. Fees are charged to the user's account, from which e-money is deducted. Additional cash fees are officially not permitted, but there is evidence that they are sometimes charged on an informal basis by agents. We return to this issue below.

E-money can be transferred from one customer's M-PESA account to another's using SMS technology or sold back to Safaricom in exchange for money. Originally, transfers of e-money sent from one user to another were expected primarily to reflect unrequited, internal, within-country remittances, but nowadays, while remittances are still an important use of M-PESA,

[7] Nonregistered individuals can receive money sent by a registered user as long as they have a cell phone. The recipient receives a text message with a code that can be taken to an M-PESA agent who provides the cash less any fees. The fee schedule is designed so as to encourage recipients to register. Note that a nonregistered user cannot send e-money to a third individual from his phone.

Figure 4 Average Transaction Size (Kenyan Shillings): Moving Down-Market



Notes: The data come directly from Safaricom.

e-money transfers are often used to pay directly for goods and services, from school fees to the wages of domestic staff.[8]

The Growth of Mobile Money

M-PESA has spread quickly and has become one of the most successful mobile phone-based financial services in the world.[9] The average number of people opening up an M-PESA account (i.e., new registrations) per day exceeded 5,000 in August 2007 and reached nearly 10,000 in December that year (see Figure 3). By August 2009, 7.7 million M-PESA accounts had been registered. Ignoring multiple accounts and those held by foreigners, this means 38 percent

[8] Transactions are not always limited to innocent trades. For example, there are reports of people using M-PESA to pay bribes to traffic police. Even worse, rumors have circulated in Nairobi that kidnappers are requesting ransom to be paid by M-PESA, although these rumors have not been confirmed.

[9] Similar services in Tanzania and South Africa, for example, have penetrated the market much less. See Mas and Morawczynski (2009).

Table 1 What Do Individuals Use M-PESA For?

	Fraction of Sample (Based on Multiple Responses)
Receive Money	28.40%
Send Money	25.08%
Store/Save Money for Everyday Use	14.39%
Buy Airtime for Myself	13.58%
Buy Airtime for Someone Else	8.30%
Store/Save Money for Emergencies	6.69%
Store/Save Money for Unusually Large Purchases	0.27%
Pay Bills	1.35%
Receive Money for a Bill/Else Pay Bills	0.77%

Notes: Each entry is the share of registered M-PESA users in our sample who reported the corresponding function to be the most commonly used. The bill payment service had only just started at the time of the survey and has since become rather popular.

of the adult population of Kenya had gained access to M-PESA in just over two years.

Since the launch of M-PESA, wary of regulation by the Central Bank of Kenya, Safaricom has been at pains to stress that M-PESA is not a bank. However, the ubiquity of the cell phone across both urban and rural parts of the country, and the lack of penetration of regular banking services,[10] led to hopes that M-PESA accounts could substitute for bank accounts and reach the unbanked population. Data reported in Jack and Suri (2009) suggest this is partially true, although M-PESA has been adopted by both the banked and unbanked in roughly equal proportions.[11] In addition, more recently, M-PESA users have been able to withdraw funds from their M-PESA accounts at ATMs operated by one of the commercial banks (Equity Bank) and some banks have begun to use M-PESA as their mobile banking platform. However, deposits cannot (yet) be made at ATMs, and the network of ATMs and bank branches, while growing, remain limited: In the long run they could replace agents, but both capital costs and the costs of security, operation, and maintenance suggest agents will continue to play an important role for some time.[12]

[10] In 2006 it was estimated that 18.9 percent of Kenyan adults used a bank account or insurance product, and by 2009 this had increased to 22.6 percent (see Financial Sector Deepening, Finaccess I).

[11] These data are from a survey fielded in late 2008. Since then, there has been some growth in the number of individuals and households with a bank account because of the expansion of such institutions as Equity and Family Bank.

[12] In 2003 there were 230 ATMs in Kenya (see Central Bank of Kenya [2003] at http://www.centralbank.go.ke/downloads/nps/nps%20old/psk.pdf). Recent data suggest there are around 1,200.

Table 2 Daily Financial Transactions, Oct. 2007–Sept. 2008

	RTGS	ACH	ATM	Mobile
Value Per Day (billion KShs)	66.3	8.5	1.0	0.1
Transactions Per Day (thousands)	1.0	39.2	180.2	107.2
Value Per Transaction (million KShs)	64.67	0.216	0.006	0.003

Notes: KShs = Kenyan Shillings.

Source: Central Bank of Kenya (2009).

The average size of M-PESA transactions has fallen over time as it has reached more of the population and has been used more extensively, as shown in Figure 4. In the two years following its introduction, the average transaction size fell about 30 percent, having started at KShs 3,300 (about $50). Most of this decline has probably been because of the expansion of take-up among the poorer individuals and households.

Table 1 shows the various types of transactions for which M-PESA is used, which include not just sending and receiving money, but also storing or saving money, purchasing airtime (the prepaid credit used for voice and text communications), and paying bills.

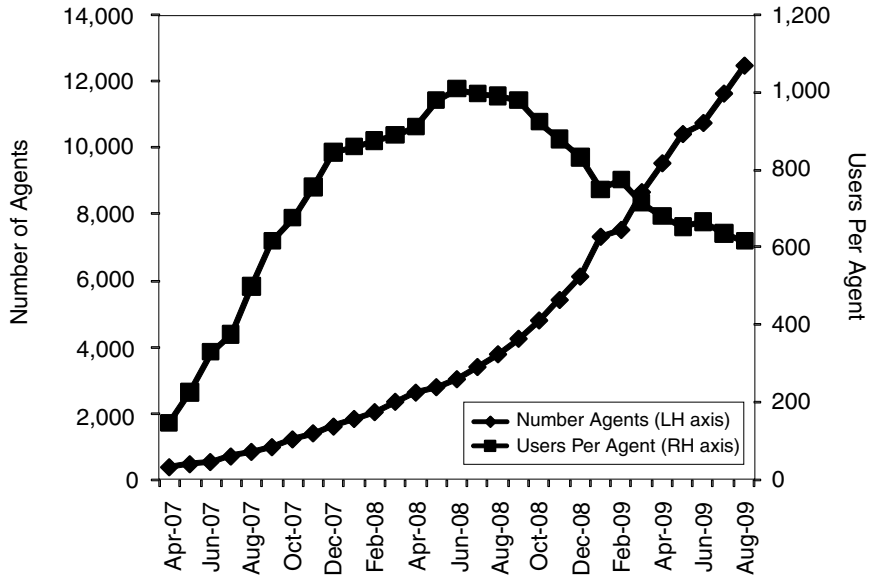
While the sustained growth in M-PESA registrations is notable, the volume of financial transactions mediated through M-PESA should not be exaggerated. Table 2 reports that the volume of transactions effected between banks under the RTGS (Real Time Gross Settlement) method is nearly 700 times the daily value transacted through M-PESA; and, maybe more relevant, the daily value transacted through the check system (automated clearinghouse, or ACH) is about 85 times the daily value transacted through M-PESA. Related, the average mobile transaction is about 100 times smaller than the average check transaction (ACH) and just half the size of the average ATM transaction.[13] M-PESA is not designed to replace all payment mechanisms, but has effectively filled a niche in the market.

The Agent Network

To facilitate purchases and sales of e-money, and in light of low rates of bank account coverage among a widely dispersed population, M-PESA maintains and operates an extensive network of more than 16,000 agents across Kenya. These agents are like small bank branches, often manned by a single person. As can be seen in Figure 5, the growth of this network lagged behind that of

[13] These data refer to a period before M-PESA could be used at ATMs.

Figure 5 Expansion of the Agent Network



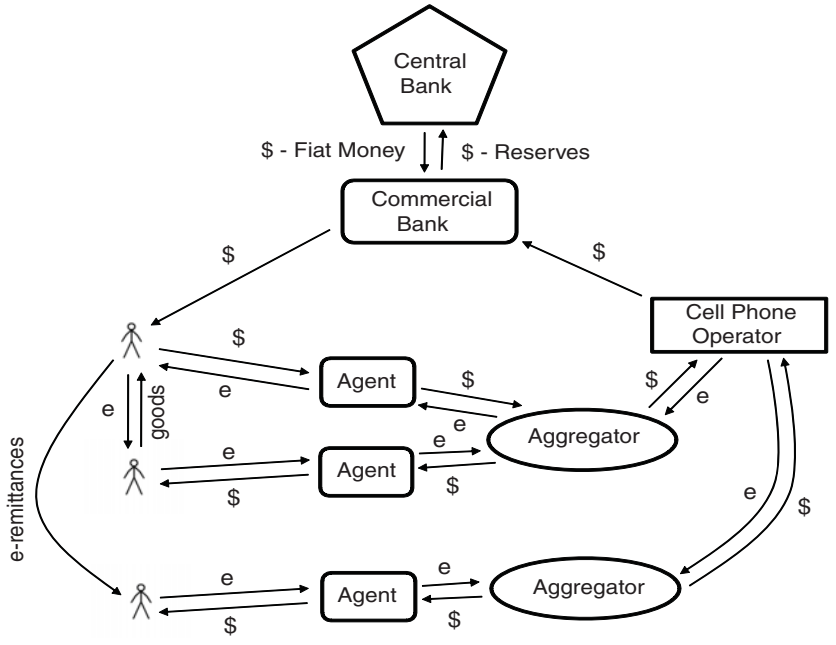
Source: Safaricom.

the customer base for the first year of M-PESA's operation, during which time the number of users per agent increased five-fold from a low of 200 to a high of 1,000. But since mid-2008, agent growth has accelerated and the number of users per agent has fallen back to about 600.

Registered M-PESA users can make deposits and withdrawals of cash (i.e., make purchases and sales of e-money) with the agents, who receive a commission on a sliding scale for both deposits and withdrawals.[14] Clearly, withdrawals of cash can only be effected if the agent has sufficient funds. But symmetrically, cash deposits can only be made if the agent has sufficient e-money balances on his/her phone. Agents face a nontrivial inventory management problem, having to predict the time profile of net e-money needs. Figure 6 shows a representation of the flow of money and e-money among individuals, Safaricom, commercial banks, and the central bank, and illustrates the core workings of M-PESA. The role of what we call the "coordinator,"

[14] The commission amounts are nonlinear (and concave) to the size of the transaction. Some reports suggest that in response to this, agents may encourage customers to split their transactions into multiple pieces, thereby increasing the overall commission.

Figure 6 Flows of Fiat Money and E-Money



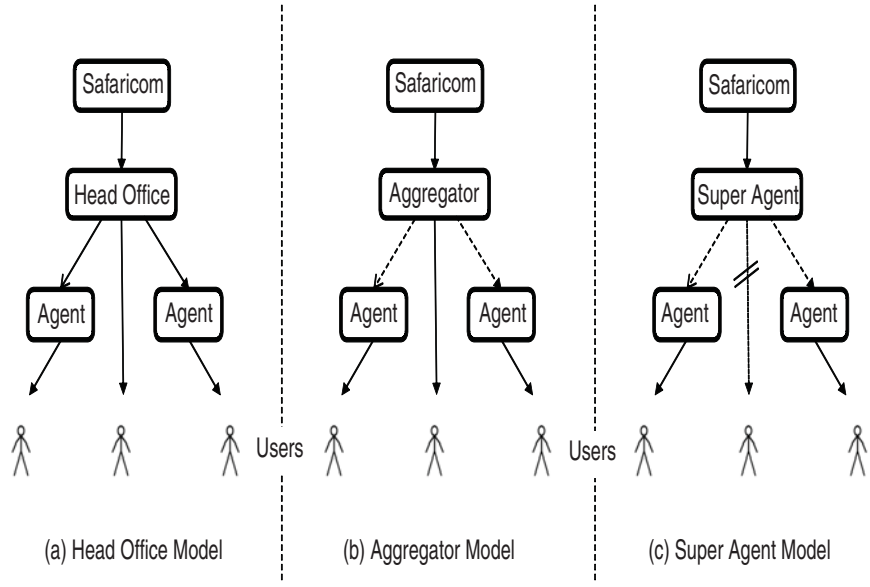
which in practice is a head office, an aggregator, or a super agent, is described in more detail below.

The network of commercial bank branches across Kenya, while growing, remains much smaller. As of November 2009, the Central Bank of Kenya[15] reports that 44 commercial banks had 849 branches across Kenya (about one branch for every 40,000 Kenyans), with 50 percent of branches concentrated in the largest (by size of branch network) four banks. As of 2008 in the United States, there were 7,086 institutions with 82,547 branches that came under Federal Deposit Insurance Corporation protection, yielding a density of bank branches about 10 times that in Kenya (whose population is about 10 percent of the United States).

In practice, M-PESA agents are organized into groups. Originally, M-PESA required that agent groups operate in at least three different locations, so that the probability of cash or e-money shortfalls could be minimized. This diversification within the group would only be effective, of course, if the inventories of money and e-money were efficiently re-allocated across agents

[15] http://www.centralbank.go.ke/financialsystem/banks/Register.aspx

Figure 7 M-PESA Agent Models



Notes: (a) The coordinating body is the "head office," which owns agents and can transact directly with customers. (b) The coordinating body is referred to as an "aggregator" and has arm's length contractual relationships with agents. (c) The coordinating body is a bank branch and is called a "super agent," but neither owns the agents nor transacts directly with customers.

in the group accordingly. There are now three agent models in operation, in which there is a central body that manages and coordinates the operations of a group of subsidiary agents. These models are differentiated with regard to the formal status of the coordinating body and the ownership structure of the group, and whether the central body conducts direct transactions with individual users, as shown in Figure 7.

In the first model, one member of the agent group is designated as the "head office," which deals directly with Safaricom, while subsidiary agents that are owned by the head office manage cash and e-money balances through transactions with the head office.[16] Both the head office and the agents can transact directly with M-PESA users. The second model is the aggregator

[16] M-PESA requires that each coordinating body has a bank account so that funds can be transferred easily between them. In order to open an M-PESA business, the coordinating body must have a minimum balance in a bank account, which is used to purchase initial holdings of e-money.

model, with the aggregator acting as a head office, dealing directly with Safaricom, and managing the cash and e-money balances of agents. However, the agents can be independently owned entities with which the aggregator has a contractual relationship. A final and much more recent model[17] allows a bank branch, referred to as a "super agent," to make cash and e-money transactions with agents on an ad hoc basis. However, the bank does not trade e-money with M-PESA customers. The super agent model is one example of the integration of M-PESA services into the banking system. Other developments in this vein include the ability to transfer funds, often via ATMs, between a user's M-PESA account and accounts at certain commercial banks with which M-PESA has forged partnerships. But even as M-PESA has facilitated transactions for the approximately 72 percent of user households in Jack and Suri's sample with bank accounts, it remains popular with the unbanked, of whom more than half (54 percent) used M-PESA.[18]

The cash collected by M-PESA agents in exchange for sales of e-money is either kept on the premises or deposited in the agent's (or head office's) bank account. When they wish to replenish their e-money balances, agents transfer money via the banking system to one of two bank accounts held by Safaricom. Safaricom is required to limit the quantity of e-money it issues to the amount of money it receives from agents—that is, e-money is 100 percent backed by deposits in commercial banks. However, these deposits are subject only to the regular 6 percent Kenyan Central Bank reserve requirement.

2. MODELS OF MONEY AND MEANS OF PAYMENT WITH SPATIAL SEPARATION

M-PESA's rapid expansion means that a large share of the Kenyan population now conducts at least some of their financial transactions by phone. In this section we discuss the implications of this new kind of payment system for the management of the financial system as a whole and of central bank regulatory and monetary policies in particular. To address these questions, we describe in some detail a number of models of money, the payment system, and clearing and settlement. The purpose is to focus on the features of the models that can provide insights into the operational design of mobile banking and inform policy choices facing regulators and monetary authorities. Therefore, we follow the summary of each model with a discussion of its implications for mobile banking. We proceed incrementally, beginning with simple but surprisingly

[17] This model started after the first round of the Jack and Suri (2009) survey.

[18] About 50 percent of households had at least one member with a bank account. Of banked households in the survey, about 60 percent used M-PESA, compared with the 54 percent of unbanked households reported above.

rich models of money, then progressively review more complex models that we believe reflect particular features of the Kenyan financial environment.

Townsend Model of Financial Deepening and Growth

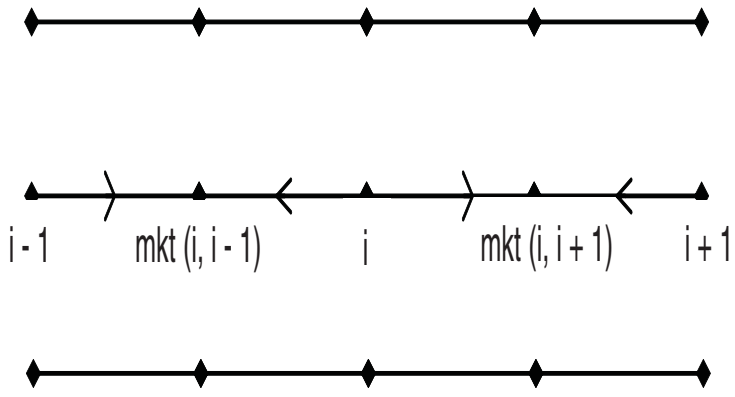
This model focuses directly on improvements in the technology of communication and links the degree of financial interconnectedness of agents with the level of economic development in a cross section and also over time. The idea is that as connectedness increases, with electronic payments connecting otherwise spatially separated agents, there is an increase in the specialization of labor, an increase in the consumption of market-produced goods, and a shift toward e-money relative to fiat money. This is the story of how financial deepening and growth are intertwined and how M-PESA could help Kenya increase gross domestic product over time at the same time as it increases monetized exchange.

Each household of type i can produce (by supplying labor) only good i, and each has a utility function over its own consumption of good i and a good it cannot produce, $i + 1$, as well as leisure. When households are in autarky, without physical or electronic contact, no trade is possible, so each household consumes all its production of good i only. In this situation, there is no need for a means of payment. In contrast, with some travel, as in the Cass and Yaari (1966) or Lucas (1980) versions of the Wicksell (1935) triangle applied many times, household i can only trade either with household $i + 1$ (whose good i values) or with household $i - 1$ (who values good i). But because of the structure of preferences (e.g., because household $i + 1$ does not value good i and instead wants goods $i + 1$ and $i + 2$), narrow bilateral *exchange* between i and $i + 1$ is in no one's self interest. This is the key lack of double coincidence of wants. Decentralized trade would give rise to autarky if it were not for valued fiat money.[19]

The timing-location is shown in Figure 8 where, in any given period, household i has two members, a shopper and a seller, who can only move horizontally to trade with households $i + 1$ and $i - 1$, respectively. Between time periods, members of household i, i even, shift down one line, and households i, i odd, stay put. Thus, debt issued by a household of type i, i even, can only be passed along to a household vertically above the issuer and so has no value. Only fiat money is used and it can have value. Specifically, one member of each household i travels to the market with $i + 1$ and purchases some of good $i + 1$ at price p_{i+1} with fiat money acquired previously; a second member travels to the market with $i - 1$ and sells some good i for money at price p_i. Note that it takes one period for goods produced and sold to come back via

[19] These models rule out private debt and future contracts in fiat money by assuming there are no pairings such that debt can be redeemed by the issuer. See below.

Figure 8 Trading Scheme for Paired Households



money holding in the interim as goods purchased. With constant prices across time and space and with a positive intertemporal discount rate, this makes it less beneficial to supply labor. This is a crucial aspect of this and other related models below.

In a Walrasian, centralized exchange regime with electronic debits and credits, households can now hold intraperiod debt for within-period purchases and, at the same time, send and receive electronic credits. At the end of the period, accounts are cleared. Intuitively, when one member of household i travels to market $(i, i+1)$ to buy good $i+1$ from household $i+1$, it is as if that member were using a credit card (or phone) linked electronically to a central account, which will not be paid until the end of the period. The second member of household i who travels to market $(i, i-1)$ and sells good i is paid with a credit card from household $i-1$. At the end of the period, these electronic debits and credits are cleared and accounts must balance (we return to interperiod debt in the Lacker model below). Note that goods produced and sold can be transformed in this way to goods purchased *within the same period*, so there is no inefficiency associated with holding idle money balances. In fact, in the equilibrium of this electronic accounting system, fiat money plays no role and its price is zero. The prices of goods themselves are in some (arbitrary) unit of account. Related, though households remain separated in space, it is as if they are transacting with one another in a centralized market that ignores the spatial segmentation as far as prices and values are concerned. However, this system works only if households are allowed to overdraft their electronic accounts and there is enough commitment or punishment to make sure they honor debts accrued within the period.

In summary, if we then assume that substitution effects dominate income effects and focus on prices, the cost of consumption of the nonproduced good in terms of labor is infinite in autarky and high in the fiat money regime relative to the centralized Walrasian electronic clearing e-regime. Moving from autarky to the decentralized money regime and then to the centralized Walrasian regime, the model predicts that labor supply increases, output of the produced commodity rises, consumption of the nonproduced good rises, consumption of the produced good drops, trade volume increases, and welfare increases. If an economy has a mix of decentralized and centralized regimes, as with some fraction of "lines" (see Figure 8) using fiat money and other "lines" using Walrasian credit, and these fractions vary across countries, then per capita national income rises as financial interconnectedness increases, fiat money decreases, and per capita private debt increases, but the ratio of fiat money to income decreases and the ratio of credit to income increases. This pattern tends to be what we see in cross-sectional data. Similar comparisons are valid for an economy that is becoming more financially integrated over time, like Kenya, where forward-looking households in the fiat currency part of the economy treat financial integration into the Walrasian e-system as an exogenous random event that happens with positive probability (essentially changing the discount rate). Note, however, that thus far, in this particular model, no household needs to use multiple means of payment.

Implications for Mobile Banking

What are the implications of this kind of model of financial deepening for a system like M-PESA? It is clear that M-PESA will change the financial connectedness of the individuals in the economy, which in the model above will cause higher economic development. Therefore, the main takeaway from this model is that M-PESA can be viewed as a technological innovation that lowers trading costs or, better put, allows financial transfers (credits and debits) across agents who are still separated in space. This improves welfare, at least in the model economy without government and no vested interests in the current intermediation system (and without other heterogeneity). Fiat money and electronic payments can co-exist if some households have access to M-PESA and some do not. However, in the model, but perhaps not in the M-PESA system, the household buying goods in effect creates a net increase in e-money within the period. If e-money were essentially only a debit card, then an initial deposit of currency would have to underlie the debit transaction, undercutting this key advantage. In other words, the theory argues that we might see features of net credit creation in the functioning of the actual M-PESA system, though perhaps at an aggregated or agent level and not necessarily at the level of individual households. However, for this feature to exist there must be a (harsh) means of preventing reneging or default so that accounts actually clear at the end. Even that requires foresight of the overall equilibrium, e.g., here

the shopper knows the prices at which the seller is receiving credits. Again, we come back to this mismatch and interperiod carryovers in the other models below.

Manuelli and Sargent Turnpike Model with Currency and Debt

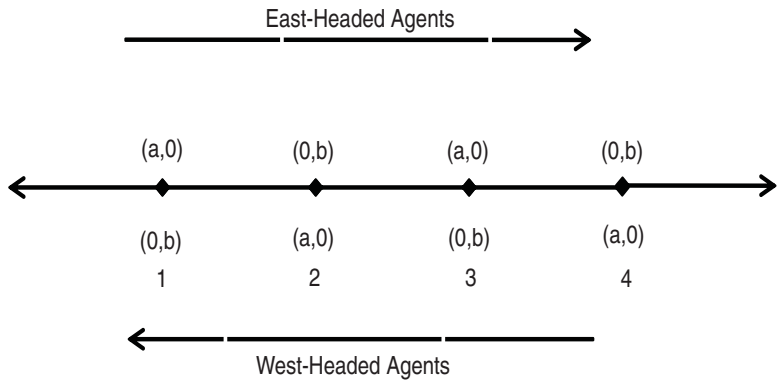
A closely related model of Manuelli and Sargent (2009) rationalizes the coexistence of fiat money and private credit. As in Townsend's turnpike models, agents meet in pairs and, while they have long enough relationships to undertake some efficiency-enhancing intertemporal trades via the extension of private credit, they do not stay together long enough to effect fully Pareto-efficient allocations. More specifically, time is divided into periods (think of these as "years"), each composed of four subintervals (e.g., "seasons"). Individuals meet for just half a year only, i.e., two consecutive subintervals, and then move on—some to the east, some to the west (see Figure 9). In the first subinterval of a half-year, one person in a given pair has a positive endowment of the single perishable consumption good and the other has none, and in the second subinterval these roles are reversed, giving rise to short-term (two-subinterval) private credit arrangements. However, the positive endowments in each subinterval can be either high or low (for example, $a > 0$, $b > 0$, and $a/b > 1$), while aggregate output in each half-year $(a + b)$ is constant, and each individual's annual aggregate endowment is constant, also equal to $(a + b)$. Because agents remain together for only two subintervals (one half-year), they cannot implement trades *across* half-years—that is, they cannot issue long-term debt. Fiat money plays a role in facilitating the trades that such debt would effect. Manuelli and Sargent generalize this to include labor supply, so that output is endogenous.

One interpretation of Manuelli and Sargent's model is as a generalization of Townsend's original turnpike in which endowments fluctuated with a periodicity of two and meetings lasted only one period. Instead of meeting for two periods, we can interpret Manuelli and Sargent as a model where households continue their travels after one period but remain linked electronically for two periods (though we ignore the requisite costly shipping of goods in the second period—some of the models below are more complicated so as to eliminate this flaw in our attempted interpretation). As the time and spatial limitations of communication fall (e.g., with the expansion of the network of M-PESA agents, accounts, and the use of cell phones), debts of increasingly long maturity can, in principle, be issued and repaid.

Implications for Mobile Banking and Monetary Policy

To the extent that mobile banking facilitates the operation of the private (often informal) credit market, a model that accommodates such products with

Figure 9 Turnpike Setup for Manuelli and Sargent (2009)



nontrivial implications for policy can be informative. To start, as in the Townsend models, the laissez faire, non-interventionist monetary equilibrium (without debt) is not Pareto optimal. Essentially, the wedge that we discussed in the earlier Cass-Yaari model, where money is earned through production and held without interest for one period, can be eliminated with intervention by paying interest on cash balances. This equates intertemporal substitution in consumption to the natural rate of time discount and ensures that no household hits a binding corner, running out of cash.

But the impact of monetary policy interventions in the form of changes to base money depends on whether private credit is allowed or, under the interpretation here, whether e-money that allows borrowing and lending is in the system. An increase in the growth rate of the money supply has ambiguous effects on the average level of output but increases the volatility of output when there are no restrictions on private borrowing and lending. However, in economies where individuals do not have access to private loan markets, say because they move on without cell phones, the results are quite different: An increase in the rate of money growth decreases mean output and has no effect on volatility of output (which remains zero). Likewise, if the economy is liberalized, or otherwise experiences a surprise innovation that allows private borrowing and lending, then prices increase and output becomes more volatile. Financial innovation is welfare-improving but intimately connected with the impact variables that central banks typically monitor or attempt to control.

As Manuelli and Sargent (2009) emphasize, the potential destabilizing effects of actual financial liberalizations are highlighted in both the academic and policy literatures. More generally, the effects of monetary policy depend on the way private credit markets are operating, even if in the process

of borrowing and lending there is no net creation of e-money. Thus, when formulating monetary policy, the central bank will need to take into account the effective change in financial regimes that M-PESA has brought with it. Indeed, in the above class of models, optimal monetary policy in terms of control over fiat base money is still relatively straightforward but not without interest. Specifically, the allocation achieved under optimal policy differs from the one associated with the corresponding economy with no locational restrictions and centralized trades permitted at time zero. While both allocations are Pareto optimal, they are not the same, implying that efficient monetary policy has redistributive consequences. Further, optimal government-issued currency continues to play an essential role even when interest is optimally paid on holdings of such currency. And, the interest-on-currency policy does not work in a way that can be replicated by free banking in a Walrasian world. Related, moving from a suboptimal policy to one with interest on currency may redistribute income and not be Pareto improving. In this model, unlike the first, e-money does not drive out fiat money nor the need for an optimal monetary policy. This is reminiscent of a class of related models of monetary management in which implementation of policy depends on the ability of agents to trade in asset markets.[20] Financial market segmentation relies on costs that may be arguably decreasing.

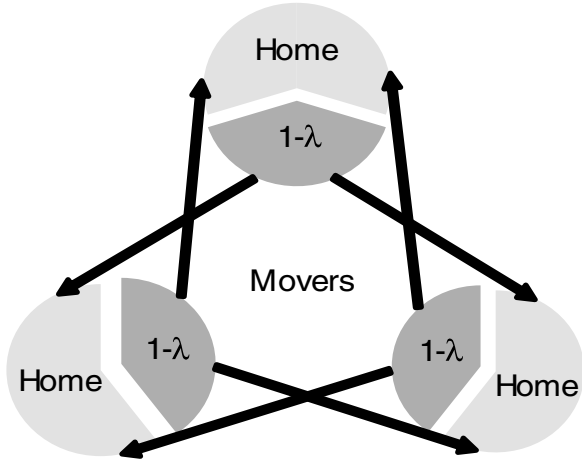
Townsend's Models of Activist Monetary Policy and Money as a Communication Device

A generalization of Manuelli and Sargent would allow credit arrangements to be used to implement trade among individuals who remain in their home location and deal with each other repeatedly over time. This is also similar to the Walrasian accounting system of the first model above, except that here again the trade is intertemporal, with borrowing and lending over time, so that any individual's balance does not have to net to zero at the end of each period. In the next set of models, credit is identifiable as direct communication and promises. Fiat money then co-exists with credit and serves as a communication device for dealing with strangers across locations.[21] But the models here feature Diamond and Dybvig (1983)-style preference shocks with patient and urgent households generating the desired intertemporal trade. Moreover, the models here deliver welfare gains from an activist monetary authority responding to shocks and managing liquidity needs. More generally, the quantity and

[20] For example, see Grossman and Weiss (1983), Rotemberg (1984), Romer (1987), Lucas (1990), Fuerst (1992), and Perez-Verdia (2000).

[21] See also Ireland (1994), Kocherlakota and Wallace (1998), Cavalcanti and Wallace (1999), Kocherlakota (2005), and Wallace (2005), and the review in Wallace (2000) in which outside money and inside money issued by banks with known trading histories co-exist.

Figure 10 On Each Island, a Share of Each Population Leaves to Other Islands in Period 2



kinds of money in the system are determined optimally in an effort to compensate for missing credit and insurance markets. In this way, one can build on the platform of e-transfers to create a highly effective recordkeeping system in which electronic accounts allow for a rich variety of financial instruments.

Townsend's (1989) model envisions a scenario where there are N islands each with N inhabitants (the case of $N = 3$ is shown in Figure 10 but, more generally, N is a large number).[22] Preference shocks that are correlated among a segment of each island's residents occur in the first period. That is, some fraction of the residents are patient, in principle willing to lend, and the residual fraction are urgent, wanting to borrow. However, a share $(1 - \lambda)$ of the population of each island moves, spreading out across all the other islands in such a way that no mover encounters anyone from his home island at his new destination. This creates a problem if recordkeeping is limited to locations, that is, if there is no cross-island communication or accounting system so that only nonmovers can borrow and lend: Promises involving movers (either among themselves but going to different locations or between them and nonmovers),

[22] In this model, the agents all pre-commit to arbitrary tax and transfer schemes over time and to all institutions and resource allocation rules. In the language of the models, they commit to an economy-wide credit arrangement that specifies consumption and transfers to agents conditional on aggregate states and on individual specific location shifters (that are public) and individual announcements of preference shocks (private). Apart from these plans, there is no government and no distinction between private and public.

on the other hand, are not credible as they cannot be consummated at a later date.

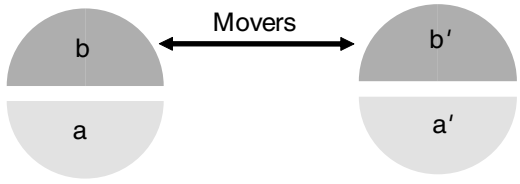
As movers are effectively excluded from the credit market, a social planner could attempt to implement efficient intertemporal consumption profiles by asking movers at each date to report their preferences, allocating consumption accordingly. But if the information reported cannot be credibly transmitted to other islands without a recordkeeping device, then the only incentive-compatible mechanism is one that gives all movers the same level of consumption in both periods, independent of their preferences. Portable fiat money allocated to movers, and monotonically related to their first period announcements, can facilitate the transmission of information across time and space to the strangers they meet at their destinations. In this interpretation, fiat money is a portable token. By allowing side trades between individuals, monotonicity can be strengthened to linearity, delivering a price of fiat money or tokens for goods. Of course, the initial nominal price level remains arbitrary, as that is simply a matter of the denomination of the unit of account.

However, if additional periods are added to the model (e.g., another round of movers), future movers must also be allocated fiat money in order to engage in intertemporal trade. The purchasing power of each unit of money allocated to second-round movers must, for efficiency reasons, be the same as that offered to first-round movers, but the quantity will be increasing in the number of movers and the proportion who are patient.[23] As the preferences of new generations of consumers are revealed, planned consumption levels supported by allocations of fiat money in early periods may be revised. Since the purchasing power as previously explained is constant across early movers in a given period, the associated adjustment to consumption levels is effected through changes in the price level. That is, inflation eats up the purchasing power of first-round movers if it is judged that they should be getting less given what new information tells the monetary authority about the way they and second-round movers should be treated. Note that this activist policy is quite different from, say, a Friedman rule, as described in the earlier class of models above in which a constant rate of deflation can remove the distorting wedge. This is not enough here. Optimal policy is state contingent (Manuelli and Sargent [2009] anticipate such results in the concluding section of their article).

Note, however, that fiat money as tokens conveys only the information that a household has been patient in the past, not that the household has been a first- or second-round mover. If even more information, such as the dates and the nature of past transactions, was encoded in the system, then the distinction between local and inter-island accounts could disappear. That is, one

[23] Individuals who are patient consume more later and therefore need more units of money to confirm this to future strangers.

Figure 11 Islands with Stayers (a) and Movers (b)



can imagine one kind of fiat money—e.g., red tokens for first-round movers, green tokens for second-round movers, and electronic accounts for those who stay home. Indeed, accounts that distinguish all these space/time transactions could be accomplished with the electronic recordkeeping that mobile technologies and markets allow, at least in principle. Indeed, with all of that, we could in theory go further and here again completely mimic the outcome of a perfect Walrasian accounting system in which changing locations per se has no consequences. The fraction of agents leaving an island would be exactly the same as the fraction arriving and, financially speaking, there would be no strangers.

Townsend (1987) generalizes this idea of multiple monies (or differentiated e-accounts) in a similar framework with four agents, spatial separation, and private information on preference shocks. In particular, suppose there are two islands with two individuals each, as illustrated in Figure 11. In period 1, agents a and b live on the left island and agents a' and b' live on the right island. In period 2, b and b' switch places, while a and a' (who are subject to shocks) remain on their home islands. Agents b and b' are risk neutral and in principle can provide insurance to agents a and a', who are risk averse. With one good, preference shocks determine not only the degree of risk aversion but also relative patience. With two goods, there can be preference shocks for each good over time (e.g., patient for good one and urgent for good two) and an overall intertemporal shock determining utility in period 1 versus period 2.

Townsend then examines the properties of trade facilitated by alternative communication devices in this environment, both for the cases of a single good as well as for multiple goods. First, oral communication can take place only between agents in the same location and so cannot be used to convey credible information across time to strangers (if agents cannot carry tokens, commodities, or messages). The equilibrium is thus Pareto inefficient. On the other hand, tokens (money) that are appropriately distributed in period 1 can be used to verify information in period 2, helping with incentives to reveal information correctly and acting again as a technology for storing that information. The previous model provides intuition for the case of one good.

However, with two goods, one type of token may not be enough. Intuitively, one wants to convey the full history of shocks for each good in the first period, yet ensure incentive compatibility in the second when agents can turn out to be very desirous of consumption overall. For example, one type of token, say green, is handed out in period 1 given a certain realization of preference shocks, while the red token is handed out given another realization of these shocks again in period 1 (alternatively, these are different "credits" in different cell accounts). Then in the second period, the agent is required to show not just the correct number of tokens, but also the correct colored token (or have the requisite balances in a specific cell account). Indeed, much can be done even with n-commodities and m-combinations of shocks using combinations of red and green tokens (two types of e-money) as an encryption system. The point more generally is that multiple monies are used to convey the history of trade, borrowing and lending, and insurance, not simply a means of payment or transfer system.

Implications for Mobile Banking

The bottom line of these models of money as a communication device is that the better the communication of past shocks or transactions, the more efficient can be the allocation of consumption; however (with initial heterogeneity), this may be wealth redistributing. The model features tokens or fiat money but, again, portable cell devices linked to some of the account history of earlier transactions would provide similar features. To achieve an efficient allocation there can arise, as in these models, the need for active liquidity management. We can see that in a scenario where M-PESA emerges as the entity behind a large fraction of transactions, e-money could substitute for fiat money or tokens. This would not necessarily replace the need for an activist monetary policy, but it would alter that policy so that the level of tokens created on net by the financial system ideally responds to mobility and the state of demand, as would electronic credits if allowed optimally to function that way. Here a distinction between private credit and public money becomes blurred as we consider questions about optimal market design. The social good is served by having mutually agreed upon and collectively enforced rules.

Another lesson from these models is that electronic records of past transactions allow new financial instruments, in this case better borrowing/lending and insurance over space and time. Tying fiat money to e-money and thinking of both as solely facilitating payments may lead one to miss otherwise beneficial arrangements that have to do with insurance against spatial and intertemporal idiosyncratic and aggregate shocks. Indeed, under the current M-PESA system, the prices at which money trades for e-money are supposed to be fixed over time and across space; e-money and cash trade for each other one for one (as described above, however, there are nonlinearities in the transactions costs by amount traded)—yet these fees can be seen as allowing in

principle a trading price between cash and e-money that is different from one. Whether or not one wants to allow money prices and the rate of exchange of money for e-money to move with the state of local demand and inventory of the actors again begs the question of what e-money is supposed to be: a means of payment only, if it facilitates an expansion of the monetary base, or a partial substitute for missing, more centralized economy-wide insurance and credit markets.

Townsend and Wallace—Circulating Private Debt and a Coordination Problem

There is yet another way to think of money, namely as an object that, even if privately issued, appears frequently in exchange, i.e., with a high velocity. We can understand this by simply extending the model environment in the previous section to four periods with households b and b' continuing to switch locations from one period to another, back and forth, and with households a and a' remaining in a single location. Townsend and Wallace (1982) replace preference shocks with time-varying endowments of a single good, but with different profiles for the different agents, to induce the desire for intertemporal trade. They also assume there are many agents of each type in any given location to justify price-taking behavior. In one of the equilibria, the first period household b makes a deposit of goods (but could be money) to (that is, lends to) agent a, as if agent a were a bank issuing long-term debt (or at least debt payable on demand). However, household b does not hold this debt but rather moves in the second period to a different location inhabited by agent a'. At this new location, neither party is physically connected to bank a. Subsequently, in the third period, agent a' will pass the debt to b', who in turn redeems it in the last period since b' meets up with the original issuer, agent a. Note that long-term debt is also the debt that circulates, that has a high velocity. In that sense circulating debt has something to do with maturity transformation. Short-term debt (e.g., two-period debt) between agents a and b (or a' and b') not only extinguishes sooner but it also does not circulate.

With private debt transferable electronically, one has the same equilibria but with the more realistic interpretation of agent a as an M-PESA agent who issues debt (in this case in exchange for goods, not fiat money, but see below). That is, household b uses the e-money account to trade with subsequent households and, again, the e-money is netted out back to zero when a third party comes to agent a to redeem it. In this model, agent a' can also play this role as banker, or M-PESA agent, instead of a. However, without coordination, another problem emerges. The amount of e-money issued by agents a and a' has to be coordinated so as to be consistent with the overall equilibrium. If M-PESA agents a and a' are not communicating across space, then it is hard to imagine how this would happen. Townsend and Wallace refer to various

historical episodes such as the crash of markets in bills of exchange as evidence that the model with coordination problems is picking up problems that may occur in practice.

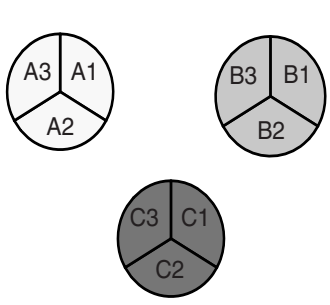
Implications for Mobile Banking

Electronic debits can be transferred across agents in spatially separated locations and have a high velocity. This seems to capture a big part of the Kenyan M-PESA reality. This comes, however, with potential coordination issues that need to be thought through. In the current model with two locations, four agent types, and four periods, one achieves the first-best with the right combination of circulating private debt and other short-term noncirculating debt. In that equilibrium, prices/interest rates are moving around over time and space and all markets in goods and financial instruments are clear. Again, fixing the price at which one object trades against another would seem to create additional problems. But even if prices were flexible, it appears that agents need to coordinate on the overall credit issue. Lack of initial coordination could show up as an over-issue or under-issue of the correct financial instrument, or of the combination of instruments that is supposed to give the correct maturity structure, showing up in turn later on as sharp movements in prices. This could even lead to doubts about the commitment or ability of agents to achieve the requisite transfers of purchasing power necessary for liquidity in intermediate periods or to ensure redemption of debt at maturity. Manuelli and Sargent ponder whether fiat money can help solve this type of coordination problem.

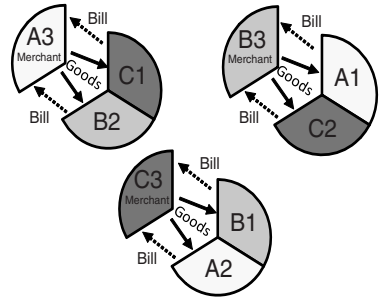
Lacker's Model of Clearing and Settlement and Inter-Agent Markets

Lacker (1997) focuses on clearing and settlement via a central bank and the impact of certain central bank policies such as reserve requirements and interest paid on reserves. Building on the earlier models of Townsend (1983, 1989), Lacker develops a model in which there is a large number, N, of islands, on each of which live N individuals. Each island produces a single perishable good that must be consumed on the island. This geography is illustrated for the case $N = 3$ in Panel A of Figure 12, in which the islands are labeled A, B, and C, and the individuals are 1, 2, and 3. In each period, all but one of the individuals who live on a given island travel to all the other islands at random, one to each, with one staying behind. In Panel B of Figure 12, individual 3 remains home. All individuals consume the good that is produced on the island they visit (so the one who remains consumes the good produced on his island). As in Townsend (1989), before they leave "home," travelers entrust their endowment of goods to the individual who stays behind (called the merchant banker) and is responsible for handing it out to arrivals from

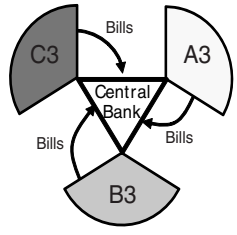
Figure 12 Lacker's Model of Settlement



Panel A: Three individuals (1, 2, 3) endowed with goods born on each of three home islands (A, B, C)



Panel B: Individual 3 (merchant banker) on each island stays home; 1 and 2 from each island (shoppers) travel to other islands and consume goods on visited islands in exchange for bills



Panel C: Merchant bankers from each island meet at the central bank to present bills and settle accounts

other islands. The record of the amount entrusted to the stay-at-home agent is an individual's "deposit" and the merchant banker is thought of as operating a bank that holds his island's deposits.

As illustrated in Figure 12, each island receives a fully diversified group of visitors each period, one from each other island. If preferences and endowments were suitably fixed (e.g., if each individual consumed $1/N$ units of the good of the island s/he visited), consumption and income would balance on a person-by-person basis. However, Lacker assumes, like some of the models above, that each period the islands are hit by Diamond-Dybvig idiosyncratic independent identically distributed preference shocks that affect the urgency of consumption. All individuals from a given island get the same shock. Since each island is visited by an individual from every other island and since shocks are independent across islands, there is no aggregate uncertainty about the demand for each island's good (as N goes to ∞). However, an island that suffers a run of large urgent shocks over time consumes more over time than an island that suffers a run of small shocks.

Because goods do not move between islands, there is no possibility to directly exchange one for another. Instead, an individual purchases consumption from the merchant on the island he visits by providing a bill or check drawn on his deposit held by his own merchant who stayed at home. (This could be an electronic charge to the e-account but, again, not one paid instantaneously.) In turn, each merchant collects bills or e-credits from all other islands, one for each visitor. In any given period, some islands will consume more than they produce (i.e., issue more bills than they collect or be left with negative e-balances), while the opposite will be true for others. Intertemporal trade between islands across periods, i.e., interbank borrowing and lending of e-balances, is thus efficient.

In the final stage of the period, all the merchant bankers travel with their bills to a central location and submit them (to each other) for payment (Panel C of Figure 12). With cell technologies, physical meetings would not be necessary. Payment is effected through an accounting mechanism, with each island's account being credited and debited according to the bills or e-money presented to and by it. The residual that does not clear is carried over, in surplus or deficit.

Implications for Mobile Banking

Lacker refers to the central institution that keeps the accounts of each island as the central bank and these accounts are thought of as *reserve* accounts. However, this could equally be a private clearinghouse run by Safaricom or some other independent entity, as the model focuses on the account-keeping and clearing functions of the institution, not the issuing of money per se. Positive account balances with the institution are the liabilities of that institution, while negative balances represent overdrafts. In the model, bills are *cleared* (i.e., accepted by the clearinghouse/central bank) at the end of the period and *settled* (i.e., deposits transferred by the institution from one island's reserve account to another's) at the end of the period for across-period borrowing and lending.

Beyond Lacker's model, limits on within-period bank overdrafts with the clearinghouse/central bank can induce some banks (ones with a positive preference shock) to borrow from others. Likewise, limits on overnight overdrafts can induce residents of islands that have had a string of positive, urgent shocks to constrain their consumption below the efficient level, as they are unable to borrow enough. Lacker's model is a useful motivation for thinking about another aspect of the M-PESA system, in particular overall clearing and the related inventory management problem faced by agents. The kind of contractual conditions Safaricom might want to specify would be crucial given the reality of the actual economy in which the distinction between within-period and across-period clearing and borrowing/lending is hard to maintain.

We identify each M-PESA agent with a merchant banker in Lacker's model, although individuals are not bound to agents like residents are to islands. An M-PESA agent's trading account at Safaricom corresponds directly to the reserve account held by each bank with the central bank. To parallel the model, individuals deposit their endowments (of cash) with an agent each period, which requires that the agent hold sufficient e-money. An agent would take out an overdraft loan from Safaricom if he were required to issue e-money to a customer before having presented the equivalent amount of cash to Safaricom. Because transferring a bank note or cash is slower than transferring e-money, it seems likely that there could be demand for such overdrafts.

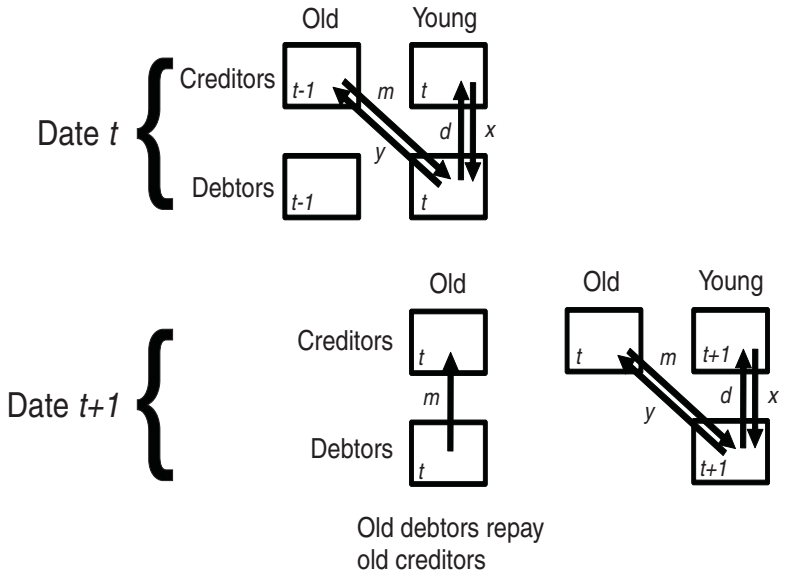
E-money is sent between individuals (i.e., checks are exchanged) and recipients present their e-money (i.e., checks) to agents. This happens at the end of the period in Lacker's model. If agents have enough cash to purchase the e-money from customers, their trading accounts are credited with the relevant amounts. In reality, as individuals visit the agent over the course of the day, his net demand for e-money will fluctuate and he might require short-term overdrafts from Safaricom or need to acquire cash in some other way. In the absence of such a facility, he will need to trade off the costs of holding "zero-interest reserves" (e-money balances on his trading account) against the costs of reduced trade (and commissions). Alternatively, agents could lend e-money to one another, creating the equivalent of an interbank market as envisioned in Lacker's model. Again, this might be organized by another institution (like a clearinghouse) that itself purchased e-money from Safaricom and lent it out to agents at some interest rate. Likewise, the head offices or super agents could perform this role, though neither appears to charge interest. Each head office or super agent would face a similar inventory management problem of course, having to hold enough e-money and/or cash to lend out during the day/period.

Freeman and Green's Models of Liquidity—Optimal Base-Money Management

Freeman's (1996) model and Green's (1999) reformulation, related to Townsend (1989) as exposited above, focus on getting money and circulating debt in the same setting simultaneously because of imperfect meetings between creditors and debtors. This, again, has implications for liquidity and monetary policy.

In Green's overlapping generations model, there are two types of individuals (creditors and debtors) who live for two periods each. A creditor is someone who in equilibrium will be willing to defer consumption, while a debtor will wish to borrow. We follow tradition and refer to young and old agents, simply to imply the first and second periods of the two-period, dynamic transactions profiles of the households. When young, creditors and debtors are endowed with perishable goods x and y, respectively. In the first

Figure 13 Trade with Money and Credit



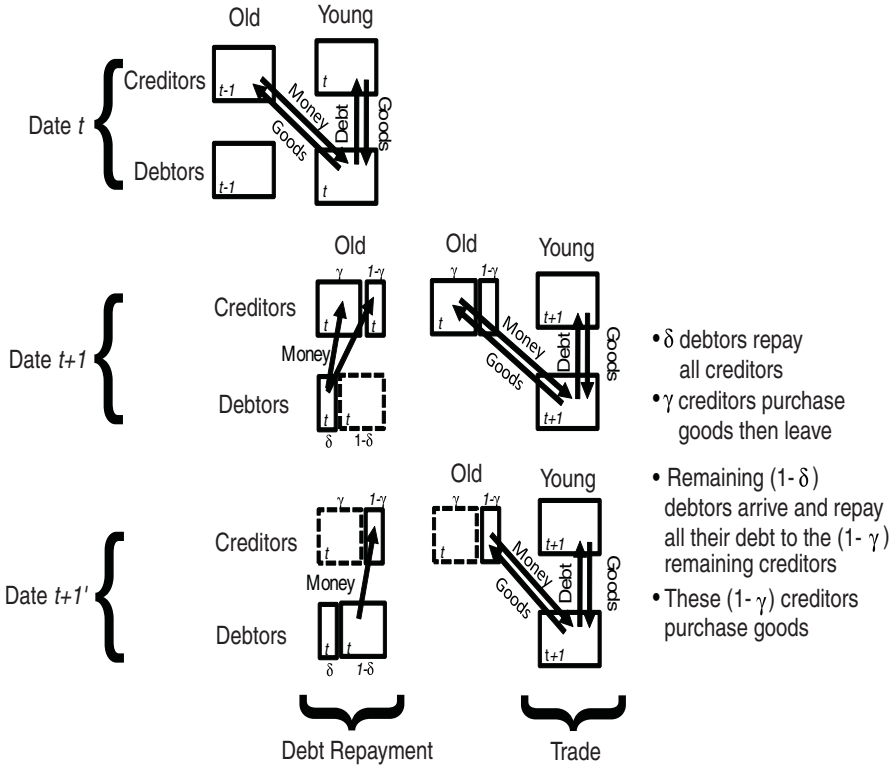
Notes: The dates within each box refer to the period of birth.

period, old creditors are endowed with fiat money and old debtors have nothing. Creditors and debtors also differ in their preferences: Creditors wish to consume when they are young and old, while debtors wish to consume only while young. Both types prefer to consume a mix of goods x and y instead of just their own.

Suppose young debtors meet young creditors first and only then go on to meet old creditors. Young debtors purchase x in return for debt d that they issue to young creditors, as illustrated in the first panel of Figure 13. Subsequently, young debtors sell their own good y to old creditors in exchange for money. At the beginning of the next period (the second panel of Figure 13), now-old (previously young) debtors settle their debts using money with now-old (previously young) creditors. Once the debt is settled the process repeats, with the now-old creditors holding money and the new young cohorts endowed with goods.

Nontrivial monetary dynamics can arise when creditors and debtors do not necessarily meet at the "right" time. With various waves of movers, old agents either arrive late or leave early: In particular, a fraction $(1 - \delta)$ of debtors arrive late and a fraction γ of creditors leave early. This naturally complicates the debt settlement process and can lead to inefficiencies. In

Figure 14 Late-Arriving Debtors and Early-Leaving Creditors

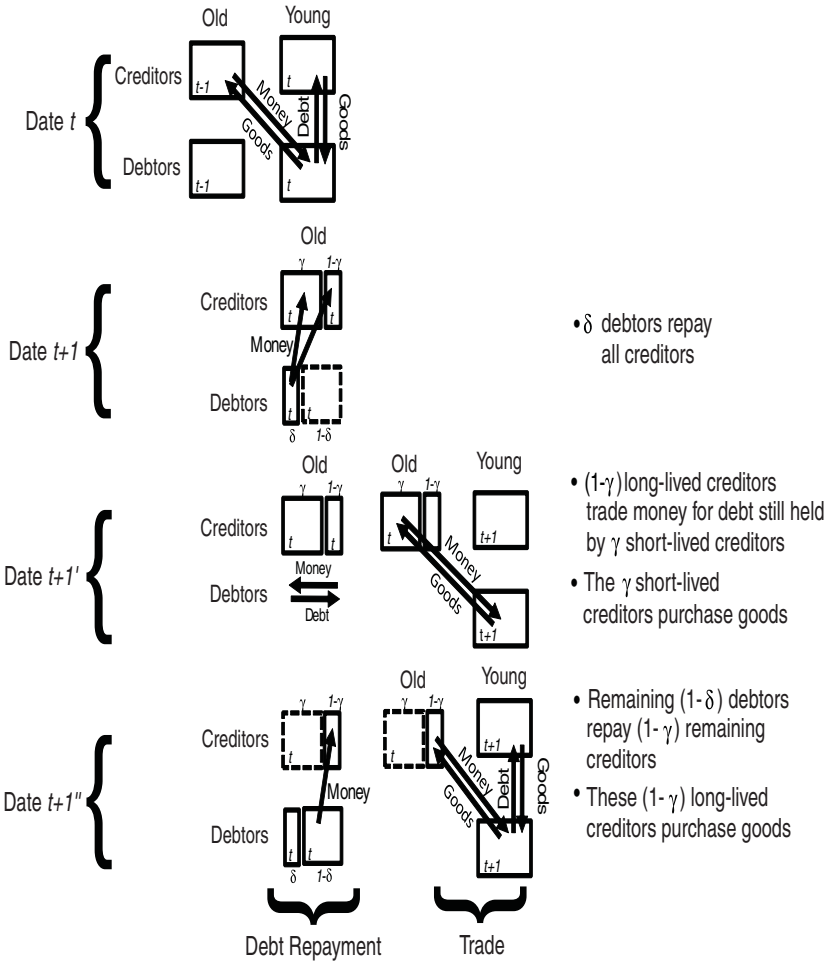


Notes: In the figure, debt settlement and trade between the young and old occur in the same period, $t + 1$.

particular, while efficiency requires that all creditors consume the same quantity of goods when old (purchased from young debtors), those who leave early may in equilibrium consume *less* than this amount while those who leave later consume more. Thus, early-leaving creditors can end up facing liquidity shortages that constrain trade.

This situation is illustrated in Figure 14, in which period $t + 1$ is divided into two subperiods ($t + 1$ in the second panel down, and $t + 1'$ in the third panel). Agents not in the market at the relevant moment in time are shown as boxes with broken lines. It is assumed that the creditors are fully diversified at $t + 1$, holding debt issued by each and every old debtor. At the beginning of $t + 1$, all the old creditors are present, as are a fraction, δ, of old debtors. The debtors settle their debts and each creditor receives a share, $\delta < 1$, of

Figure 15 Trade in Money and Credit with a Secondary Debt Market



the amount owed to him. At date $t + 1'$, the late-arriving old debtors are able to settle their debts in full with old creditors who remain. Early-leaving old creditors will consume less than their efficient level of consumption and late-leaving ones will consume more.

An alternative scenario illustrated in Figure 15 allows old creditors to exchange debt. Period $t + 1$ is now divided into $t + 1$, $t + 1'$, and $t + 1''$. At the beginning of $t + 1$ (the second panel), early-arriving debtors repay their debts to all creditors. As before, all creditors continue to hold outstanding debt issued by debtors who have not yet arrived. Next, at time $t + 1'$ (the third panel), early-leaving creditors sell their remaining holdings of debt to

long-lived creditors in exchange for money. If $\delta < \gamma$, with a relative scarcity of debtors and an abundance of early-leaving creditors, the creditors spend all of their money on debt, which has a price less than one. Early-leaving creditors then purchase goods from young debtors and quit the market. In the final panel, at time $t + 1''$, late-arriving debtors settle their debt with those creditors who remain, including the debt those creditors bought from early-leaving creditors. They then purchase goods from young debtors. The amount of consumption they enjoy is $(1 - \delta)/(1 - \gamma)$ times the efficient level. Thus, if not enough debtors show up in time, even if creditors trade their debt, the allocation is inefficient.

As Freeman observed, a central bank can remedy the inefficiency by issuing money to some or all creditors and then withdrawing it from circulation later, say by taxing young creditors as they enter the second period. The important issue is that new money is issued to creditors and is issued *before* the early-leaving creditors depart.

Implications for Mobile Banking

In Green's version of the model, debtors consume nothing in the second period of their lives, but creditors do. The only reason for old debtors to come to the market is to pay off their debts. So if e-money allows them, or some of them, to do this without coming to the market, then the share that are "late" is smaller. In the extreme case, there would be no late-arriving debtors and no liquidity problems for creditors. But if some old debtors still didn't pay off their debts in time (maybe because they couldn't find an agent with e-money), then it would be possible that early-leaving creditors wouldn't have enough money to finance the efficient level of purchases from young debtors.

These models can inform our thinking about mobile banking in a couple of ways. First, focusing on the reduction in transactions costs associated with transferring e-money, mobile banking might reduce the proportion $(1 - \delta)$ of debtors who "arrive late" and the proportion of creditors, γ, who "leave early." Debtors who previously had to physically meet their creditors in order to settle their debts can now settle them with e-money and no longer need to be present. On the other hand, even if not all debts are repaid at the beginning of the period (i.e., if there remain some late-arriving debtors), the existence of e-money could relax the liquidity constraint faced by early-leaving creditors and make central bank intervention less necessary.

However, it is overly simplistic to assume that mobile banking allows individuals to send *money* costlessly: it allows them to send *e*-money costlessly (or at least at low cost) but they must acquire it first. A more complete model would thus include individuals holding optimal mixes of money and e-money and would describe the production process whereby each is converted into the other. In practice, this conversion is effected by M-PESA agents who simply

Table 3 The Problems Consumers Have Had with Agents

	Most Used Agent	Closest Agent
Agent Gave Less Money/ E-Money Than I Was Owed	2.63%	2.80%
Agent Charged Me to Deposit	1.11%	1.68%
Agent Overcharged Me	1.17%	1.84%
Agent Undercharged Me	0.52%	0.78%
Agent Was Absent	0.74%	0.91%
Agent Refused to Perform the Transaction	0.80%	0.61%
Agent Was Unknowledgeable	1.04%	1.57%
Agent Was Rude	3.66%	6.02%
Agent Had No E-Money/Not Enough E-Money	43.60%	22.81%
Agent Had No Cash/Not Enough Cash	34.78%	51.33%
Other	9.95%	9.65%

perform the role of a technological black box—a black box that is sometimes out of service.

Although this feature is not part of the Freeman and Green set-up, if money and e-money are both used in equilibrium, then a "late-arriving debtor" might correspond to an individual who is otherwise "on time" and has sufficient financial resources (money and/or e-money) to repay his debts, but who is frustrated in not being able to find an M-PESA agent with sufficient e-money or money.[24] Similarly, an early-leaving creditor in this environment could be one who has in fact been repaid, say in e-money, but who must use money to purchase the consumption good.[25] If he cannot find an M-PESA agent with sufficient cash, then he could be liquidity constrained as above. First, if he is stuck with e-money but must trade with money, he will suffer a loss equal to his excess e-money holdings. On the other hand, even if he cannot find an M-PESA agent to trade with, he might trade his e-money with another creditor for cash, just as early-leaving creditors sell their debt to late-leaving creditors for cash in the third panel of Figure 15. But such trades must take place between locationally proximate agents, and if there is an excess supply of e-money locally, the allocation of consumption might remain inefficient.

[24] Whether the debtor needs to find an M-PESA agent with money or e-money will depend on what form of financial wealth the debtor has on hand and how the creditor wishes to be paid. This in turn will depend on the specific features of the two monies.

[25] Again, whether the creditor needs money or e-money depends on how the seller wants to be paid.

Table 4 Unable to Deposit Cash (No E-Money) or Unable to Withdraw Cash (No Cash)

	Most Used Agent	Closest Agent
Have You Ever Been Unable to Deposit Money at this Agent?		
Yes	6.63%	6.22%
No	93.37%	93.78%
Total	100%	100%
Have You Ever Been Unable to Withdraw Money from this Agent?		
Yes	6.63%	15.33%
No	93.37%	84.67%
Total	100%	100%

3. LINKING THEORY WITH DATA: RESULTS FROM HOUSEHOLD AND AGENT SURVEYS

In this section we present data that speaks to some of the issues raised by the models of money summarized above, especially as regards shortages of e-money and cash and whether there are indeed credits in the system because of the operational logistics of agents as described in Section 1. These data, some of which are reported in Jack and Suri (2009), derive from a survey of 3,000 households and 250 M-PESA agents in Kenya in late 2008.[26] We focus on issues related to M-PESA agents as reported by consumers and the agents themselves, as motivated by the models.

First, 10 percent of all consumers reported facing at least one problem with the agents they had visited. Of those who reported problems, Table 3 shows the breakdown of the problems they had. By far, the most common problems are agents' lack of cash and e-money. The first four rows in the table, in fact, suggest that in some cases agents have been able to increase the price of e-money by varying the fees they charge consumers. This is an important implication of the models discussed above—fixing prices for cash and e-money will require an accompanying policy stance. However, the penultimate two rows in the table confirm that this strategy is employed nowhere near enough to clear the market.

In addition, in the survey, consumers were specifically asked if they were either unable to deposit money or unable to withdraw money from the

[26] Part of these data form the basis of a confidential report issued by Financial Sector Deepening to the Central Bank of Kenya. In addition, Jack and Suri (2010) look at some of the microeconomic risk-sharing impacts of M-PESA. Other papers have also looked at the more microlevel impacts of e-money on currency demand (for example, see Fujiki and Tanaka [2009]).

Table 5 How Often Do Agents Run Out of E-Money?

	Fraction
More Than Once a Day	3.2%
Once a Day	6.4%
Once a Week	14.0%
Once a Month	5.6%
Once Every Three Months	1.2%
Once Every Six Months	0.4%
Less Often Than That	12.0%
Never	57.2%

M-PESA agent closest to them or from the agent they used the most. Table 4[27] shows that approximately 6 percent of M-PESA users are unable to deposit money with an agent, i.e., the agent does not have any e-money to give the consumer in return. Also, as many as 15 percent of consumers were unable to withdraw money from the closest agent, i.e., the agent had no cash to give the consumer in exchange for e-money.

In the survey of agents themselves, respondents were asked how often they run out of e-money and how often they run out of cash—these results are reported in Tables 5 and 6. On average, about 29 percent of agents run out of e-money once a month or more frequently and indeed a nontrivial fraction (14 percent) run out about once a week. Similarly, about 26 percent of agents run out of cash once a month or more frequently than that and about 10 percent run out once a week (and, in fact, about 8 percent run out on a daily basis). Clearly there are liquidity issues, both in terms of cash as well as in terms of e-money. This is anticipated from the discussion of Lacker (1997), Freeman (1996), and Green (1996)—models in which such liquidity constraints are evident.

Safaricom initially required all M-PESA agents to pre-purchase e-money before they could trade it for money to the public. If an agent runs out of e-money, he is required to purchase more, either from Safaricom or from the public when they redeem cash, before being able to take cash deposits from the public. This suggests that there are no credits or debits involved between agents and Safaricom and therefore no role for a formal settlement system. Indeed, even as the agent model has evolved, this feature has been maintained, at least with respect to the relationship between the "coordinating bodies" of Figure 7 and Safaricom. On the other hand, cash and e-money transactions between agents and their head offices or aggregators need not remain in continuous balance, and the parties can operate in a net credit or

[27] Note that the questions asked for Tables 3 and 4 are quite different. Table 3 asks about the main consumer-reported problems with agents while Table 4 asks about the incidence of two specific problems.

Table 6 How Often Do Agents Run Out of Cash?

	Fraction
More Than Once a Day	3.2%
Once a Day	8.4%
Once a Week	10.0%
Once a Month	4.8%
Once Every Three Months	1.2%
Once Every Six Months	0.4%
Less Often Than That	22.4%
Never	49.6%

debit position vis-à-vis each other. These imbalances are of little concern for the head office model (Panel A of Figure 7), to the extent that the agents are owned and controlled closely enough that internal financial arrangement of the group does not affect its viability. However, the more arm's length relationship between agents and an aggregator (Panel B) suggests that chronic imbalances with such a group could prove problematic.

We note that while the potential exists for persistent financial imbalances within a group under the aggregator model, in principle M-PESA users on the one hand and Safaricom on the other do not face any risk associated with the bankruptcy of any particular agent or agent group, as deposits of cash are matched at the level of the coordinating body by transfers of e-money and vice versa. If an individual user finds that all agents within a reasonable distance go out of business, he will likely face a liquidity constraint unless he is able to use his e-money to directly purchase goods and services.

The survey asked agents how they pay for e-money when they request it from their head office. In well over half the cases, agents receive transfers from their head offices without any immediate corresponding payments (see Table 7).

Similarly, agents were asked how they get cash for M-PESA transactions when they run out. As reported in Table 8, in more than half the cases agents do not immediately exchange e-money for cash received.

The statistics in Tables 7 and 8 suggest that credit arrangements, explicit or otherwise, between agents and their head offices or aggregators appear to be widespread. We do not know the maturity of these credits but, given the large number of agents reporting them, it is possible that at least some of them are longer than simple overnight positions. Indeed it seems inevitable that there may be a nontrivial amount of credit in these transactions as the "supply chain" of e-money involves exchanges in spatially and temporally separated markets, leading naturally to one-way transfers of either cash or e-money. Given the mechanics of M-PESA, these credits cannot be issued between agents and individual users or between the coordinating body and Safaricom. But, the

Table 7 How is E-Money Paid For When an Agent Requests It?

	Fraction
It is a Direct Purchase of E-Money From the Head Office (Involves a Cash Transfer)	36.2%
Receive a Direct Transfer From the Head Office with No Concurrent Payment	31.2%
Receive a Direct Transfer From the Head Office with No Payment Required	18.1%
Other	12.6%
Refused to Answer	2.0%
Total	100%

evidence indicates that such net credits/debits do exist between agents and their coordinating bodies. Again, all the models above allow for credit and debits, which are often welfare improving. In addition, some of the models above illustrate the welfare-improving nature of broader financial integration, which such credits and debits would encourage.

4. CONCLUSION AND FURTHER MODELING

The most successful version of mobile banking in Kenya (and perhaps the world), M-PESA, is—quite literally—everywhere. In many cases, the scenarios envisioned in existing monetary theory models appear to match the reality of M-PESA and, as such, these models promise to inform decisions taken by both Safaricom in managing M-PESA and the central bank in managing the Kenyan economy. The empirical evidence presented, from surveys of both M-PESA users and agents, further serves to illustrate the importance of these lessons.

For example, just as the central bank may intervene to relax liquidity constraints, it is arguable that Safaricom should actively manage "e-liquidity" by issuing e-money that is at times, in some locations, unbacked by money deposits, assuming that such activism would be costless and allowed by the central bank. In fact, the data suggest that some M-PESA agents are engaging in such e-liquidity management already (for example, when they receive e-money transfers from their head offices without a corresponding transfer back of cash). This has implications, of course, for the measurement and meaning of monetary and debt aggregates. Improved systems, however, require that the company have better information on net demands for e-money across agents than the agents themselves have, or at least be better able to act on this information without the space/time coordination problems that the models suggest. Not surprisingly, Safaricom has been changing their agent model over time to better deal with cash and e-money liquidity issues. Time series

Table 8 How Do Agents Get Cash When They Run Out of It?

	Fraction
Redeem From Head Office in Exchange for E-Money	17.6%
Redeem Direct Transfer of Cash From Head Office with No E-Money Exchange	20.4%
Use Own Savings	42.8%
Other*	18.0%
Don't Know	1.2%
Total	100%

Notes: *Of which 27 percent is from "sale of credit card," 27 percent is "wait for deposits," 21 percent is "borrow from management," and 11 percent is "from the other business."

and geographically disaggregated data on fluctuations in demand would be useful for further evaluating these issues and making improvements to their system.

On the modeling side, understanding the operations of the M-PESA agent network seems key to the development of an overall comprehensive model of e-money. For example, modeling the decisions and constraints of agents would potentially allow us to endogenize the timing patterns assumed in some of the existing models. Frictions that impede efficient and immediate reallocations of money and e-money balances across agents would thereby replace these timing assumptions as the fundamental source of liquidity constraints. Similarly, realistic heterogeneity across consumers—for example, in terms of phone ownership, access to mobile coverage, safety of the local environment, frequency of market access, access to M-PESA agents—could be modeled more explicitly.

REFERENCES

Cavalcanti, Ricardo, and Neil Wallace. 1999. "A Model of Private Bank-Note Issue." *Review of Economic Dynamics* 2 (January): 104–36.

Cass, David, and Menahem E. Yaari. 1966. "A Re-Examination of the Pure Consumption Loans Model." *Journal of Political Economy* 74 (August): 353–67.

Central Bank of Kenya. 2009. Presentation at Conference on Banking and Payment Technologies, East Africa, Nairobi (February).

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

Financial Sector Deepening. FinAccess Reports, various. Kenya: FSD.

Foster, Kevin, Erik Meijer, Scott Schuh, and Michael A. Zabek. 2010. "The 2008 Survey of Consumer Payment Choice." Federal Reserve Bank of Boston Public Policy Discussion Paper 09-10 (April).

Freeman, Scott. 1996. "The Payments System, Liquidity, and Rediscounting." *American Economic Review* 86 (December): 1,126–38.

Fuerst, Timothy S. 1992. "Liquidity, Loanable Funds, and Real Activity." *Journal of Monetary Economics* 29 (February): 3–24.

Fujiki, Hiroshi, and Migiwa Tanaka. 2009. "Currency Demand, New Technology and the Adoption of Electronic Money: Evidence Using Individual Household Data." IMES Discussion Paper Series (December).

Green, Edward J. 1999. "Money and Debt in the Structure of Payments." Federal Reserve Bank of Minneapolis *Quarterly Review* 23 (Spring): 13–29.

Grossman, Sanford J., and Laurence Weiss. 1983. "A Transactions Based Model of the Monetary Transmission Mechanism." *American Economic Review* 73 (December): 871–80.

Ireland, Peter N. 1994. "Money and the Gain From Enduring Relationships in the Turnpike Model." Federal Reserve Bank of Richmond Working Paper 94-07.

Jack, William, and Tavneet Suri. 2009. "Mobile Money: The Economics of M-PESA." Working Paper.

Jack, William, and Tavneet Suri. 2010. "The Risk Sharing Benefits of Mobile Money." Working Paper.

Kocherlakota, Narayana. 2005. "Discussion of 'From Private Banking to Central Banking: Ingredients of a Welfare Analysis.'" *International Economic Review* 46 (May): 633–6.

Kocherlakota, Narayana, and Neil Wallace. 1998. "Incomplete Record-Keeping and Optimal Payment Arrangements." *Journal of Economic Theory* 81 (August): 272–89.

Lacker, Jeffrey M. 1997. "Clearing, Settlement and Monetary Policy." *Journal of Monetary Economics* 40 (October): 347–81.

Lucas, Robert E., Jr. 1980. "Equilibrium in a Pure Currency Economy." *Economic Inquiry* 18 (April): 203–20.

Lucas, Robert E., Jr. 1990. "Liquidity and Interest Rates." *Journal of Economic Theory* 50 (April): 237–64.

Manuelli, Rodolfo, and Thomas J. Sargent. 2009. "Alternative Monetary Policies in a Turnpike Economy: Vintage Article." New York University Working Paper (June).

Mas, Ignacio, and Olga Morawczynski. 2009. "Designing Mobile Money Services: Lessons from M-PESA." *Innovations* 4 (April): 77–91.

Perez-Verdia, Carlos. 2000. "Transaction Costs and Liquidity." University of Chicago, PhD dissertation.

Romer, David. 1987. "The Monetary Transmission Mechanism in a General Equilibrium Version of the Baumol-Tobin Model." *Journal of Monetary Economics* 20 (July): 105–22.

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

Townsend, Robert. 1983. "Financial Structure and Economic Activity." *American Economic Review* 73 (December): 895–911.

Townsend, Robert. 1987. "Economic Organization with Limited Communication." *American Economic Review* 77 (December): 954–71.

Townsend, Robert. 1989. "Currency and Credit in a Private Information Economy." *Journal of Political Economy* 97 (December): 1,323–44.

Townsend, Robert, and Neil Wallace. 1982. "A Model of Circulating Private Debt." Federal Reserve Bank of Minneapolis Staff Report 83.

Wallace, Neil. 2000. "Knowledge of Individual Histories and Optimal Payment Arrangements." Federal Reserve Bank of Minnesota *Quarterly Review* 24 (Summer): 11–21.

Wallace, Neil. 2005. "From Private Banking to Central Banking: Ingredients of a Welfare Analysis." *International Economic Review* 46 (May): 619–31.

Wicksell, Knut. 1935. *Lectures on Political Economy*. London: MacMillan.