Economic Quarterly—Volume 102, Number 1—First Quarter 2016—Pages 1–50

Models of Discount Window Lending: A Review

Huberto M. Ennis

> "*I have made a systematic analysis of discounting in my little book, A Program for Monetary Stability. Any answers that I might give to your questions now would be more offhand and less satisfactory than that statement.*" –Milton Friedman

In July 1968, the Federal Reserve released a report titled "Reappraisal of the Federal Reserve Discount Mechanism." The report contained the conclusions from a series of studies conducted by the Federal Reserve System during a period of three years. One of these studies was a compilation of the responses by a large group of academics to a set of eight questions about the subject. Milton Friedman's (complete) answer to the questions appears in the quote above. I interpret Friedman as saying that if one wants to assess the extent of knowledge on how a discount window institution should be structured and operated, the best way to proceed is to study the existing literature on the subject. The objective of this essay is to summarize some of the ideas that come from doing just that.

The academic literature on the lender of last resort (LLR) policy is extensive. There are (at least) two kinds of papers in that body of work: (1) papers that explain and formalize the barriers to perfect market functioning that the discount window is trying to address and then discuss how to best run a LLR policy given that situation; and (2) papers that start from the premise that there is a LLR and study in more detail particular aspects of its organization – such as, for example,

I would like to thank Beth Klee, Jeff Lacker, Allen Sirolly, Nico Trachter, Alex Wolman, John Weinberg, and Steve Williamson for comments. The views expressed in this article are my own and do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

DOI: http://doi.org/10.21144/eq1020101

whether the LLR should conduct supervisory activities, and how, or whether those should be outsourced to a separate agency. The focus here will be mainly on the first set of papers.

The government can conduct LLR activities in different ways. One classic approach is to delegate such authority to the central bank. But this is not the only way: the government can, in principle, put in place lending programs administered by the fiscal authority. Furthermore, not all of the central bank's LLR interventions have to be channeled through the regular discount window. Other specially designed lending facilities could be put in place instead. At the level that the subject is discussed in the particular literature I will be summarizing here, the distinction between all of these different forms of lending is often not very sharp. For this reason, and for the sake of concreteness, I will often refer to discount window lending as the generic LLR policy aimed at intervening in financial markets.

In general, to evaluate the optimality of a given discount window policy, one needs to determine the problem that such policy is trying to solve. Discount window lending may play a role in (at least) two different situations: (1) when only one or a few firms need to borrow short-term funding for idiosyncratic reasons; (2) when many firms in the system need to borrow in a situation that could be considered an economy-wide (systemic) event. In the first case, in principle, other firms have funds available that could satisfy the demands of the few borrowing firms at interest rates close to the prevailing (risk-adjusted) rates. The second case is a situation where only a significant change in interest rates would allow the system to equilibrate itself without any intervention, and a crisis may ensue. When only some firms are looking to borrow, the discount window may play a (meaningful) role only if there are some impediments to the functioning of markets limiting the ability of those firms to obtain funding from other firms. When the economy is experiencing a systemic event, the discount window may be one channel through which the central bank can adjust the aggregate quantity of liquidity in the market to avoid undesirable spikes in interest rates – with open market purchases of assets by the central bank, in exchange for bank reserves, being a natural alternative to that.

As is evident from this discussion, taking a general equilibrium approach is essential to evaluate the potential role of a discount window. When a set of firms (small or large) have borrowing needs, market forces will, in principle, produce the necessary price-and-quantity adjustments to accommodate those needs. The question then becomes: Are those prices and quantities desirable? Or, in other words: Is the allocation of credit in the economy efficient? To answer this question, one needs first to understand how the system adjusts in

general equilibrium. Additionally, one needs to determine the ideal (efficient) allocation to use as a benchmark in evaluating the equilibrium allocation. So, one needs a full description of the economic system (i.e., a general equilibrium model) and a notion of efficiency applicable to the set of feasible allocations of resources in that system.

Holmstrom and Tirole (1996) subscribe to this approach when they say: "Modeling aggregate liquidity shortages and analyzing liquidity premia require a general equilibrium model in which no spurious demand for liquidity is introduced through ad hoc restrictions on asset trades." In this quote, the ideas of "liquidity" and "liquidity shortages" are crucial elements. However, these are terms that are often used in different contexts, with different meanings. To be able to evaluate alternative policies we need to understand the specific phenomena that lie behind those terms.

One possibility is to broadly interpret the concept of "demand for liquidity" as the need of one firm, or a set of them, to borrow short-term funding. Relatedly, in some cases liquidity refers to the idea of "cash in the market" and the fact that under certain conditions there may be only a limited amount of nominal means of payments available to execute an appropriate amount of trade. These concepts intend to capture complex situations also best understood in the context of a well-specified explicit model. The intention in this review is to discuss in some detail several existing attempts in the literature to *formally* analyze the role of a LLR using such models.

The central bank discount window is generally considered an integral tool in monetary policy implementation frameworks (see, for example, Ennis and Keister [2008]). The idea there is that the interest rate charged at the discount window represents the costs of being short of reserves at the end of the trading day and hence determines the willingness of banks to pay for reserves during the day. Understanding the behavior of the daily demand for bank reserves is crucial to the implementation of monetary policy when the central bank follows the common practice of intervening in the market for reserves in order to target a level for some relevant interest rates. Furthermore, the discount window rate will also provide a virtual ceiling for the interest rate on interbank loans of reserves in such a situation. This more specific role of the discount window is not the main focus of this article (see Ennis and Weinberg [2016] for a brief discussion of this topic).

Separating the LLR function of the discount window from its role in monetary policy implementation has some theoretical backing. In a now classic paper, Goodfriend and King (1988) argued that unless there are significant barriers to the functioning of financial markets, central-bank open market operations – that is, buying and selling assets

in exchange for reserves in the open market – are sufficient to conduct monetary policy effectively.[1] In principle, this separation of functions can be analytically helpful. However, in this article we will review several possible obstacles to the perfect functioning of markets that will make the Goodfriend and King argument much less clear-cut (Flannery 1996). As a result, monetary policy issues will arise in the discussion even though these are not the main focus of the article.

The rest of the paper is divided in two sections. In Section 1, I review several general equilibrium models that have been used to address the question of how to conduct appropriate discount window policy. I try to follow (approximately) a chronological order in the presentation, and I make an effort in the discussion to identify features shared by some of the models. In Section 2, I summarize the main common themes that come out from reviewing the literature, and I provide some concluding remarks.

1. AN OVERVIEW OF THE MODELS

One of the first and most influential contributions to our understanding of the theoretical determinants of aggregate liquidity conditions is the model proposed by Diamond and Dybvig (1983). While the original model is not designed specifically to discuss the role of the discount window, many features of the Diamond-Dybvig framework have been used later in the literature to address the questions that interest us here. For this reason, I start this section with a brief discussion of this seminal contribution.

Diamond and Dybvig (1983) consider an abstract general equilibrium economy populated by a large number of agents facing idiosyncratic preference shocks that drive them to want to consume more or less, earlier or later. These shocks are private information. For simplicity, call the agents who want to consume early impatient. The rest of the agents are patient.

There is also available a productive technology that delivers positive returns but requires time to mature. The optimal arrangement is to provide insurance to consumers against their idiosyncratic preference shock, but since providing "liquidity" insurance is costly – as it reduces investment in the productive technology – insurance is only partial:

[1] One way to interpret Goodfriend and King's (1988) discussion is as recasting some of the most compelling arguments in Friedman's (1960) book, using a more modern perspective. Friedman, like Goodfriend and King, favored open market operations as the main monetary policy tool and went further in saying that, in the U.S., "rediscounting should be eliminated."

impatient agents are able to consume more than in autarky but less than patient agents.

Diamond and Dybvig show that in the absence of aggregate uncertainty about liquidity needs, there is an optimal arrangement where agents pool deposits in a bank-like institution and receive payments according to their preference shock in an incentive compatible way – that is, given those payments, agents do not want to pretend they have experienced a different shock than the one they actually received. In fact, that arrangement produces the first-best allocation. They then move on to study the case of aggregate uncertainty and show that, under the assumption that payouts are executed sequentially in a first come, first serve fashion (sequential service), the first-best allocation is no longer implementable.

Diamond and Dybvig also discuss the possibility of self-fulfilling runs in their model. An extensive literature has developed that refines the insights about financial fragility that come out from the model. A detailed discussion of that literature is beyond the scope of this article (see, for example, Ennis and Keister [2010] for a survey). As it turns out, the original Diamond-Dybvig framework does not produce clear-cut prescriptions about the value of having in place a discount window facility. The model, however, has been extended and modified in various ways to address such issues. Some of those contributions are discussed below.

Liquidity Risk, Moral Hazard, and the Interbank Market

Bhattacharya and Gale (1987) use the Diamond and Dybvig (1983) environment as a starting point for their analysis. They consider the case with no aggregate uncertainty – that is, there is a continuum of depositors and the law of large numbers applies, so that the proportion of impatient depositors in the economy is equal to the known probability of each individual depositor being impatient. Also, in their setup, there are two technologies: a liquid and an illiquid, more productive technology. The liquid technology can be liquidated early or late but there is no extra return from waiting. The illiquid technology, instead, produces higher returns when waiting but cannot be liquidated early.[2]

[2] Another way to think about this is that, for the illiquid technology, the liquidation costs are so high that there is effectively no benefit from trying to access the invested resources early. As it turns out, this feature of the illiquid technology makes the traditional Diamond-Dybvig self-fulfilling bank runs not possible in the environment.

Bhattacharya and Gale restrict (exogenously) the way intermediation can be organized in the economy. They assume that there is a continuum of intermediaries and that those intermediaries are divided into two groups (two types): in one group, intermediaries have a low proportion of impatient depositors and, in the other group, they have a high proportion.

This (industry) structure with multiple intermediaries is not explicitly justified in the model. In fact, centralizing the process of intermediation appears to be a simple solution to some of the problems that arise. Furthermore, proposing policies to address those problems while abstracting from the un-modeled reasons that justify the presence of multiple private intermediaries can be regarded as problematic: it is likely that those un-modeled reasons have important implications for the design of the optimal intervention policy. Ignoring such possibilities constitutes an obvious challenge to the robustness of the results. Keeping in mind this qualification, let us proceed to describe the main insights from this influential paper.

The timing of events in the model is important. Initially, at the time of the investment decision, intermediaries do not know which type they will be. Its type gets revealed to the intermediary at the time when impatient depositors wish to consume (no sequential service), but this information remains private (is not observable by the other intermediaries in the economy). Furthermore, the decision of how much liquid and illiquid investment to undertake is also private information of the intermediary. There is, hence, a combination of two private information problems: a hidden state problem associated with the possibility that an intermediary could misrepresent its proportion of impatient depositors after types are realized, and a hidden action (moral hazard) problem associated with the ability of intermediaries to choose the level of liquidity in their portfolios.[3]

Taking as given the assumed industrial organization of the intermediation industry – with multiple intermediaries – and the information structure imposed on the model, Bhattacharya and Gale solve a planning problem subject to incentive compatibility constraints. The planner receives reports from each intermediary about their type and designs payments to depositors so that each intermediary has

[3] Allen, Carletti, and Gale (2009) study a related model where banks face idiosyncratic liquidity shocks and can trade their long-term assets for liquidity in an interbank market. The authors also consider the possibility of aggregate liquidity shocks. Shocks and portfolios are observable, but deposit contracts are assumed incomplete (payoff cannot be contingent on the liquidity shocks). The model supports the Goodfriend and King (1988) insights in the sense that central-bank open market operations are the appropriate policy in such an environment.

incentives to: (1) truthfully reveal its type and (2) choose the recommended (optimal) amount of liquidity to hold in its portfolio. The specification of those payments to depositors (given type and liquidity) implicitly defines a transfer scheme across intermediaries (which are later interpreted in the decentralization as the result of a set of transactions in the interbank market).

Once intermediaries choose the level of investment and, hence, liquidity – all choose the same, since they are identical ex ante – the uncertainty about the proportion of impatient depositors constitutes a ("liquidity") risk to each intermediary and its depositors. Since there is no aggregate uncertainty, this risk could be pooled across intermediaries and, at least partially, could be insured. The incentive constraints, however, put a limit to the amount of insurance that can be effectively provided. An intermediary, when insured, has incentives to invest more in the high return illiquid technology and rely on the insurance provider (the planner) to make payments to impatient depositors. Also, if too much insurance is provided, then the intermediary has an incentive to misrepresent its type and claim a proportion of impatient depositors higher than the true one. The planner deals with this trade-off between insurance and incentives and strikes the optimal balance.[4]

Since the total proportion of impatient depositors in the economy is known and fixed ex ante, in terms of economy-wide resources, providing insurance to intermediaries is not costly. It only involves redistributing resources from one set of intermediaries to the other. On the other hand, just as in the Diamond-Dybvig setup, it is generally costly to provide insurance to *individuals* because making payments to impatient depositors requires investing less in the more productive (but illiquid) technology. For this reason, consumption of the impatient depositors of a given intermediary is always lower than the consumption of its patient depositors.

With only two types, the limits on the ability of the planner to fully insure intermediaries come from the nonobservability of liquidity decisions. That is, the moral hazard problem is the crucial friction in the model. In fact, if investment in liquidity were observable, then full insurance would be possible even when intermediaries' types remain private information.

The basic trade-offs determining the constrained-optimal allocation are the following. Liquidity is useful to pay to impatient depositors. An intermediary can underinvest in liquidity and then try to reduce the

[4] Ratnovski (2009) considers a model where banks may have incentives to underinvest in liquidity ex ante to exploit the central bank's tendency to serve as a LLR when trying to contain the damage from a systemic banking crisis.

associated cost of choosing low liquidity by falsely claiming a high proportion of impatient depositors. To control such behavior on the part of the intermediaries, it is optimal to limit the "net transfer" to those intermediaries that report a high proportion of impatient depositors. As a result, depositors in those banks ("illiquid" banks) consume less than depositors in banks with a low proportion of impatient depositors ("liquid" banks). In other words, only partial liquidity insurance is optimal, and both patient and impatient depositors in the "illiquid" banks (that is, not just the impatient depositors), by consuming less, share the cost associated with providing appropriate incentives to banks.[5]

In principle, one way to decentralize desirable allocations in this environment would be to have intermediaries borrowing and lending in an interbank market to "insure" their liquidity risk. However, an interbank market for loans operating under laissez-faire conditions would not implement the optimum – the solution to the planner's problem. Instead, careful inspection of the constrained-optimal allocation suggests that "illiquid" banks should get a loan of a given (limited) size at a "subsidized" rate. The authors argue that a discount window could play a role in providing these loans. Unfortunately, there is no detailed discussion of the way the system would work, short of completely substituting for the private interbank market and having the discount window provide all loans.

Based on this logic, the results of Bhattacharya and Gale (1987) could be used as a justification for subsidized discount window lending. Under the optimal allocation, intermediaries would associate an interest rate R_b to the intertemporal trade-off between paying patient and impatient depositors. Implementing the constrained-optimal allocation would require providing loans to intermediaries with a high proportion of impatient depositors at a lower interest rate R_o.[6] In this sense, the rate R_o would involve a subsidy relative to R_b.

The origin of this apparent subsidy is the following. At the time that impatient depositors are being paid, the economy is just reallocating funds from those intermediaries that need funds to those that do not. From an economy-wide perspective, these transfers could be made one-for-one (as they do not create an extra resource cost for the economy). That suggests that R_o could be equal to unity. The reason

[5] Also working with a framework inspired in the Diamond and Dybvig model, Keister (2016) studies bailouts and financial fragility. In Keister's model, a similar moral hazard effect is present: when intermediaries expect a bailout, they become less liquid. Keister argues that the optimal way to handle this problem is to introduce a tax on short-term liabilities.

[6] Actually, depending on parameters, R_o could be greater than R_b in the solution to the planning problem. However, if the amount of liquidity redistribution necessary to implement the optimum is not too large, generally R_o will be lower than R_b.

that R_o is greater than unity is that, when R_o is relatively low, intermediaries do not have sufficient incentives to invest in liquidity *ex ante*. However, the value of R_o that accommodates the incentive problem may be lower than the intertemporal rate of transformation R_b (which directly depends on the return of the productive technology).

One way to summarize the main takeaway from this discussion is that if the necessary reallocation of liquidity across financial institutions does not involve an intertemporal reallocation of resources, then the interest rate on the implied loans does not have to reflect any intertemporal tradeoffs. As a result, the optimal interest rate on those loans can be different from the main intertemporal price prevalent in the economy and, for that reason, may appear to involve a subsidy.

Before closing the discussion, it is interesting to call attention to the remarks that Holmstrom and Tirole (1998, p. 35) make about Bhattacharya and Gale's article: "The paper characterizes the socially optimal mechanism for sharing risk across banks, noting that this mechanism cannot be implemented through an interbank lending market. Whether a market in bank shares or some other institution could implement the social optimum is left open. The authors suggest that a central bank might be the right institution for carrying out interbank risk sharing." The quote draws attention to the relative tenuousness of the proposed role for a discount window. It seems unclear, based on the analysis in the paper, whether other feasible trading activity could render discount window lending superfluous in such an environment.

Spatial Separation, Nominal Debt, and Illiquidity

Freeman (1996) studies a general equilibrium model where spatial separation among agents seeking to exploit gains from trade creates the need for the use of a specific type of debt instrument as a means of payment on some transactions. The debt instrument useful for transactions is one that promises to pay fiat money in the future. The particular way in which the meetings between agents occur makes promises of payment in fiat money the only meaningful ones. For this reason, in the equilibrium of the model, if fiat money has no value, those debt-financed transactions become not viable.

A basic description of the patterns of trade is the following. Initially, some agents (called debtors) meet with some other agents (called creditors). Debtors would like to acquire some goods from creditors but cannot engage in barter because they have goods that creditors do not like. At that point in the timeline of events, debtors also do not have money. They will, however, be able to sell their goods to another set

of agents later in time in exchange for money. In anticipation of this future transaction, debtors buy from creditors paying with an IOU that is a promise to pay cash in the future. Debt repayment happens at a central location where all agents who are present at the location can transact with each other.[7]

The problem is that arrivals and departures to and from the central location are not coordinated. In particular, some of the debtors arrive to the location only after some of the creditors have left, and who arrives and leaves when is unknown at the time that the IOUs are being issued. As a result, some debtors arrive to the central location after the creditors whom they need to make a payment to have left. This lack of coordination creates a motive for buying and selling debt claims – in other words, private debt "circulates" in the equilibrium of Freeman's model.

Creditors leaving early and in possession of an IOU from a debtor who has not yet arrived will want to trade the IOU with a creditor who is not yet leaving. The issue then becomes whether there is enough *cash in the market* to buy at par-value all the debt claims held by those creditors leaving early and still holding unredeemed IOUs. If the proportion of creditors leaving the central location early is high relative to the proportion of debtors arriving early, then cash in the market will be scarce and debt will sell at a discount. Freeman calls this equilibrium a liquidity-constrained equilibrium.[8]

Since creditors expect that with some probability they will have to sell their holdings of IOUs at a discount, they are less willing to initially exchange goods for IOUs, and this distorts the real allocation of goods in the economy. In other words, illiquidity matters for welfare.

Note that the reason for the illiquidity is that some agents who have money that could be used to purchase debt are not present in the market at the time when the sales need to take place. In that sense,

[7] Money has value in the economy because there are two-period-lived overlapping generations of agents and old creditors want to consume the goods possessed by young debtors. The only way that old creditors can purchase goods from young debtors is by using money. Promises to pay in the future do not work because old creditors will exit the economy in the following period. Debt arrangements are only feasible among the young agents: young creditors make loans to young debtors, and those loans are repaid the following period when the corresponding counterparties are old. Old debtors repay in cash to old creditors, who then use it to buy goods from the new young debtors – who keep the cash to use it next period in repaying their debts, which they contracted in the same period but before being able to transact with old creditors.

[8] There are other formal treatments in the literature of this type of cash-in-the-market effect. For example, Acharya and Yorulmazer (2008) discuss how cash-in-the-market pricing may interact with bank failures to produce high asset-price discounts and an inefficient allocation of resources. These authors argue that providing liquidity to the appropriate potential buyers of assets is a better policy than trying to contain asset sales with a bailout to failing banks.

there is market segmentation in the model. Alternatively, one could think that this is a version of the "slow moving capital" idea discussed, for example, by Duffie (2010). Essentially, all of these are possible justifications for cash-in-the-market pricing.

The central bank could try to address this "cash shortage" by issuing extra cash at the time that asset sales need to happen and channeling it somehow to the market. To that end, Freeman proposes a discount window policy that would solve the illiquidity problem with the additional benefit of not creating inflation – as money is injected in and out within the period. Basically, creditors can rediscount debt with the central bank and get cash, which they then can use to buy more debt from the market and again rediscount that debt at the window. It is easy to see that this will completely undo the illiquidity problem. All the rediscounting of debt is done by creditors leaving the central location late. Once the debtors arriving late finally arrive, they repay the debt to the creditors still at the central location and with that cash the creditors at the central location pay back the discount window loans. Hence, the quantity of money relevant for the determination of inflation does not change with the rediscounting, and the central bank can deal with the liquidity problem without creating inflation.

Obviously, a discount window is just one possible arrangement to deal with the illiquidity problem in this situation. In principle, the central bank could just buy the debt in the market, wait for debtors to arrive at the central location, when they would pay back the debt directly to the central bank.

In the last section of the paper (before the conclusion), Freeman extends the model to introduce default risk. If creditors know the (average) default risk of debt but the central bank does not, then the central bank can rely on creditors to deal with the default risk. Creditors are assumed to be able to fully diversify across debtors, so they only care about the average default risk. The central bank makes loans to creditors who are responsible for repaying the loans in full to the central bank. Then, creditors would be willing to buy debt in the market as long as the price of debt reflects the default risk of that debt (and not any liquidity premium). This creates a distinction between discount window lending (to trustworthy parties) and outright purchase of debt in the market.

In Freeman's model money plays two roles. It is used to trade goods between old creditors and young debtors at the end of the period, and it is used for the clearing and settlement of debt within the period. The real value of money is determined by the need to pay for goods, but the resulting amount of real balances may not be enough to permit the clearing of debt at par within a given period. Similar sources of

"illiquidity" appear in other setups in the literature, such as in the model by Holmstrom and Tirole (1998), which we discuss in the next subsection. In Freeman's model, the way to deal with this intra-period illiquidity is to have an intra-period elastic supply of currency provided by the central bank through the discount window. The central bank's lending facility adds enough flexibility to the quantity of money at each point in time, so as to allow money to appropriately play all its roles in the economy.

Freeman (1999) extends the model to include *aggregate* default shocks. The idea is that in some states of the world only a proportion of those debtors who have issued an IOU are able to travel to the central location, where repayment of debt happens. Initially, when debtors are trading with creditors, they do not know whether all or only some debtors will be able to travel to the central location later in the period, nor do they know which debtors will travel and which ones will not. The rest of the model is basically the same as in Freeman (1996).

In the model with aggregate default risk, how the central bank structures its liquidity provision can matter for welfare. In particular, open market operations result in price level fluctuations that can produce better risk sharing than the fluctuations that arise when the central bank intervenes through the discount window. The reason why price level fluctuations happen is that the central bank, due to the default shock, is not able to recover at the end of the period all the liquidity injected intraperiod to ameliorate the impact of cash-in-the-market pricing on transactions.

Note that by buying assets outright through open market operations, the central bank assumes the default risk directly and transfers it more evenly to all agents holding money through the resulting fluctuations in the price level. Instead, when the central bank provides loans to creditors in order for them to buy IOUs of debtors, in many situations, those creditors retain most of the default risk. The more uneven distribution of risk under discount window interventions can be detrimental to welfare.

Freeman's analysis abstracts from what is potentially an important problem associated with liquidity provision by a central bank: moral hazard. Two follow-up papers investigate the issue in Freeman-like frameworks: Martin (2006) and Mills (2006). Martin (2006) considers a model where agents can choose between a safe and a risky investment. Provision of central bank credit can distort the investment decision of agents, and Martin shows that a collateral requirement on central-bank loans can help to minimize the ensuing moral hazard problem.

In Martin's model, whenever agents have the resources to repay the loans, they do so. In other words, there is no strategic default. Mills (2006), on the other hand, allows agents to control their loan-repayment decisions. That is, agents who take a loan from the central bank will repay only if they have the appropriate incentives to do so. At the same time, Mills allows the central bank to engage in costly confiscation of property if an agent does not repay a loan. Because enforcement is costly, moral hazard can compromise the ability of the central bank to fully resolve the liquidity problems that may arise. Mills also considers an opportunity cost of providing collateral, absent in Martin's model, and shows that costly collateral also can limit the ability of the central bank to costlessly address liquidity issues.

Spatial Relocation, Lending, and the Limits to Diversification

Williamson (1998) (following Champ, Smith, and Williamson [1996]) writes a general equilibrium model that combines some features of the Diamond and Dybvig model with some features of the Freeman model and studies discount window lending in such a setup. Williamson also studies the role of deposit insurance in his proposed environment and how it interacts with the discount window. To begin the discussion, I will describe a simplified version of Williamson's model and consider only the effects of having a discount window in place. Later in the discussion, the role of deposit insurance in the model will be briefly considered.

There are two groups of agents in Williamson's economy. One group of agents has some resources and the other group has productive investment projects. The members of the former group will be called lenders, and the members of the latter group will be called borrowers. Investment projects have positive expected returns and take time to mature. The returns of each project have an idiosyncratic and an aggregate component, both random.

Demand for liquidity is motivated as follows. After investment takes place (and before it matures), a fraction of the lenders discover that they need to relocate. Investment cannot be moved across locations. Only liquid, low-return assets can be transported. Given this situation, it is optimal for agents to form a banking coalition (similar to those in Diamond and Dybvig [1983]), which allows lenders to pool their (in this case) relocation risk with other members of the coalition.

If it is possible to set up a bank that can be present in all locations (with branches, say), then the relocation risk can be fully insured.[9] That is, all agents, those relocating and those not relocating, consume the same amount. In other words, those agents with an immediate need of liquidity do not suffer a cost in the optimal arrangement. This is actually different from what happens in the Diamond-Dybvig model, where the impatient agents consume less than the patient agents in the optimal allocation.

The reason for this difference is that in Williamson's model, agents need the liquidity to take on their relocation trip but not to consume it immediately. As a result, when a bank can be in all locations, it does not need to liquidate investment to supply consumption to agents who relocate – each location loses and gains some agents due to the relocation process and symmetry implies that the total consumption needs in each location, after agents have relocated, remain balanced. In the optimal arrangement, agents do not take liquidity with them as they relocate. They instead have a claim on the bank associated with their deposit, and they can withdraw resources, according to that claim, in the locations that become their final destinations.[10]

Things are different, however, when banks can be present only in one location. Williamson motivates this limitation by resorting to the long-standing restrictions on bank branching throughout U.S. history.[11] In that case, when an agent relocates, she needs to take with her low-return liquid assets. For this reason, insurance becomes costly and the best implementable allocation does not provide full insurance to relocating agents.

The timing of events in the model is such that some agents need to relocate before the return on investment gets realized. For this reason, payments to relocating agents cannot be made contingent on aggregate productivity, and only agents not relocating can bear that risk. This implies that when productivity is high, agents not relocating

[9] Champ, Smith, and Williamson (1996) pursue a different interpretation by assuming that banks can issue notes that are transportable and that banks from different locations can exchange notes among themselves (or in a central market) on a regular basis. The results are essentially the same under either interpretation: branching or note issuance.

[10] While productive investment produces random returns, liquid low-return investment is riskless. For this reason, the optimal allocation still assigns some resources to the riskless (liquid) technology. This pattern is the result of pursuing the optimal portfolio allocation under uncertain returns and not due to optimal liquidity provision.

[11] Under the alternative interpretation of note issuance, the assumption would be that banks are constrained in their ability to freely issue notes. Champ, Smith, and Williamson (1996) motivate such restrictions on the system prevailing during the national banking era in the U.S. In the model, restrictions on branch banking or note issuance are not explicitly motivated.

consume more than agents relocating, but when productivity is low, agents not relocating consume less than agents relocating. This pattern of consumption is also a generalization over the one that takes place in the Diamond-Dybvig environment, where the agents experiencing the liquidity shock always consume less than the agents who do not have a liquidity shock (since providing insurance is costly).

Williamson then introduces a discount window in the model. The central bank has a discount window office in each location. The way the discount window works is that the central bank issues its own claims in exchange for productive bank assets and those claims can be transported by relocating agents to their new location (where they can be redeemed at the local discount window office). Under this arrangement, if there are enough assets suitable for rediscounting, then all agents again receive the same consumption levels, regardless of whether they are or are not relocating.[12]

When the level of productive assets pledgable at the discount window is not large enough, there may be a role for a deposit insurance system in the economy. Deposit insurance enhances the ability of the economy to insure agents against the aggregate productivity shock. The details of the interaction in the model between discount window lending and deposit insurance are complicated and not essential for the discussion here.

There is no fiat money in Williamson's model. Instead, the central bank issues IOUs that can be redeemed the following period at the local office of the central bank, with real assets as backing. In a closely related paper, Smith (2002) embeds a version of the Williamson model in a dynamic general equilibrium overlapping-generations economy, which allows him to introduce fiat money and discuss the implications of having the discount window provide loans in fiat money.

In Smith's model, there is only one investment technology (with nonstochastic returns), which cannot be moved across locations unless it is liquidated at a cost. However, in the equilibrium with valued fiat money, banks can still diversify their portfolio between liquidity and productive assets – just as in the Williamson model. Here, though, holding liquidity amounts to holding money and is only useful to deal with the relocation activities of agents (since, in contrast with Williamson's setup, there are no aggregate technology shocks). Money

[12] In the model, there are two types of productive assets, and the optimal contract requires that one type of asset be monitored after the loan is granted. Those assets are, then, deemed not suitable for rediscounting as this reduces the incentives of banks to monitor them.

is a recognizable asset that can be transported across locations and used in transactions.

Agents in the model form banks that cannot communicate across locations. As in Freeman's (1996) model, money plays two roles in the economy: it allows the banks to make payments to relocating depositors and it allows the old generation to buy goods from young agents and banks in the typical overlapping-generations pattern. Because transactions between the old, the young, and banks happen before the relocation shocks are realized, the value of money is not contingent on the size of the relocation shock (i.e., the proportion of agents relocating). In the absence of a discount window, then, full insurance is not always optimal: when the relocation shock is large enough, agents relocating may receive less consumption than agents not relocating.

When a discount window is introduced in the environment, outcomes are sensitive to the interest rate used for rediscounting. Smith (2002) studies the case when the discount window provides loans at a rate that is equal to the nominal interest rate in the economy. Because banks are indifferent between holding money and taking loans at the discount window, price level indeterminacy becomes a feature of the equilibrium. However, Antinolfi and Keister (2006) provide a more general analysis of discount window policies and show that when the discount window rate is a "penalty rate" there is a unique steady state equilibrium in the economy.

Antinolfi and Keister (2006) also show that it is optimal to make the penalty on the discount window interest rate as small as possible. The logic follows from ideas discussed already by Williamson (1998). Basically, banks take loans at the discount window to provide better insurance to depositors. If those loans are provided at a penalty rate, then banks economize on them and reduce insurance. It is optimal to minimize the resulting misallocation of risk by reducing the cost of insurance as much as possible (without going as far as to make banks indifferent between holding liquid assets and taking discount window loans, which would result in equilibrium indeterminacy).

Smith's setup is also suitable for studying the interaction between monetary policy and discount window policy. Here, monetary policy is understood as the rate of growth of money, which in steady state translates directly into the inflation rate. Interestingly, by changing the real return on money, monetary policy can induce banks to invest more or less in the productive technology – as money competes with productive investment in the bank's portfolio allocation problem.

Smith shows that in this economy, in the absence of a discount window, it is not optimal to follow the Friedman rule (a policy of targeting the nominal interest rate to be zero), since such a policy tends to drive

productive investment to levels that are suboptimally low. As a result, in the optimum, the rate of return on productive investment is higher than the return on money, and banks do not fully insure the relocation shock. This situation, then, opens the door to operating a discount window.

Once the discount window is in place, banks have less reason to hold liquidity, and instead they choose to invest more. In fact, Antinolfi and Keister (2006) show that by combining a discount window with a monetary policy that closely approximates the Friedman rule, the economy can get arbitrarily close to the first-best allocation. In the resulting equilibrium, banks invest most of the proceeds from deposits in productive investment and borrow from the discount window to deal with the relocation (liquidity) shock.

Limited Commitment and Aggregate Liquidity

Holmstrom and Tirole (1998) set out to specify a micro-founded general equilibrium model where shortages of aggregate liquidity can happen. By comparison, in Freeman's (1996) model shortages of liquidity happen because some of the liquidity available to the economy is held, at certain crucial points in time, in the wrong hands. In Holmstrom and Tirole (1998), instead, *aggregate liquidity* (all considered) is insufficient to allow the economy to reach the optimal allocation of resources.

Generating an aggregate shortage of liquidity is not an easy task – particularly when the objective is to keep the argument, and all its details, fully specified. A natural reaction to informal descriptions of situations where there is a shortage of liquidity (or collateral) is to ask: Why wouldn't the price of the liquid assets adjust to resolve the shortage? Holmstrom and Tirole address this and other related issues explicitly and are able to provide a formal equilibrium model that delivers a shortage of aggregate liquidity. Equipped with such a laboratory, then, Holmstrom and Tirole address the question of government intervention without the drawback of having ruled out, for unexplained reasons, alternative arrangements that could, in principle, improve the situation. Given the nature of Holmstrom and Tirole's contribution, a meaningful discussion requires a relatively detailed description of the specifics of their model. We turn to this description next.

Consider a setup in which there are a large number of firms that invest in a productive technology that produces random returns. In an interim period, between the time when the initial investment happens and when the returns are realized, each firm requires an extra investment to keep production running. The amount of extra investment is

also a random variable, interpreted as a liquidity shock. After liquidity shocks are observed, the firms' managers may undertake costly effort in order to improve the probability of success of the investment (i.e., increase the probability that returns are high). This effort is not observable, so the manager of the firm must receive part of the return as compensation to provide him with incentives to make the appropriate amount of effort. The direct implication of this moral hazard problem is the *limited pledgability* of future cash flows.

At the time of the liquidity shock, firms need to obtain funding from external sources. If a firm does not obtain extra funding, production is discontinued and the (potential) future return is lost. Given this, the firm's borrowing capacity is given by the total expected future return on investment, net of the compensation to the manager.[13]

In the first-best allocation (that is, when manager effort is observable), all firms with a liquidity shock lower than the expected value of future returns receive funding. In the second-best optimum (that is, when manager effort is not observable), not all those firms may receive funding. Yet, it is the case that in the second-best optimum some firms with liquidity shocks higher than the expected value of future returns *net of manager compensation* do receive funding. This outcome cannot be supported in a laissez-faire market arrangement – in such a case, only firms that have *net* future expected returns higher than the liquidity shock will receive funding to continue the project.[14]

In the model, an assumption of lack of commitment limits the ability of the private market to provide adequate liquidity to firms. In particular, deep-pocketed investors cannot sell uncollateralized liquidity insurance to firms because those investors can default with impunity ex post – when the time to make good on insurance claims comes. The only way to provide insurance credibly is to use productive assets as backing for the resulting promises. Alternatively, firms could hold claims to those assets directly and sell them to investors when the liquidity needs arise.

Note, however, that the total available value of claims on productive assets depends on the value of future returns, which is (to the extent that funding is provided) independent of interim liquidity needs of

[13] Firms are able to obtain external financing at the time of the initial investment because *expected* returns are positive. In many contingencies, when the liquidity shock is not too large, the initial investment delivers a high return. In other states of the world, though, when the liquidity shock is high, the ex-post return on that investment becomes very low, diluted by the issuance of new claims used to deal with the liquidity shock. While both contingencies are possible, expected returns at the initial investment stage are assumed to always be positive.

[14] Holmstrom and Tirole consider partial liquidation, but it does not solve the problem because constant returns to scale make partial liquidation ineffective.

firms. Hence, those claims may not be sufficient to back the amount of insurance required to achieve the optimum. In that sense, the economy may experience an aggregate shortage of liquidity. To quote Holmstrom and Tirole (1998, p. 15): "Consumers cannot sell claims on (or borrow against) their future endowments because they can default with impunity. Only promises that are backed up by marketable assets (claims on firms) can be made. This is a key assumption. Without it, there would be no shortage of liquid instruments, nor any role for government intervention."

Holmstrom and Tirole consider two cases: one where the liquidity shocks are independent across firms and the other when the liquidity shocks are correlated, creating aggregate shocks. When the liquidity shocks are independent, there is a private arrangement that can achieve the second-best optimum. Basically, firms form coalitions that resemble financial intermediaries. These intermediaries give each firm a committed line of credit that can be used if the firm cannot obtain enough funds in the market to accommodate those liquidity shocks that deserve funding according to the second-best allocation. In other words, financial intermediaries provide sufficient liquidity insurance consistent with the optimum and that insurance amounts to ex-post (i.e., after the liquidity shocks are realized) cross-subsidization across firms.

At this point, it is worth discussing briefly why intermediaries are needed. Recall that due to the lack of commitment, all claims from intermediaries need to be backed by claims on productive assets. Now suppose that instead of forming an intermediary, firms only trade claims on the future return of their investment. One possibility would be that firms, aside from making the initial productive investment, also dedicate some of their initial resources to acquire assets that they can later sell if necessary when the liquidity shocks are realized. Since the only store of value in the economy is the stock of claims issued by productive firms, the question becomes whether the total value of those claims is enough to implement the efficient (second-best) allocation. Holmstrom and Tirole show that this is not the case.

This shortage of liquidity is purely a matter of misallocation. In fact, at the level of the aggregate economy there are enough claims to potentially fund all the needed liquidity. However, since firms buy those claims before knowing their liquidity shocks, some of the claims end up in the hands of firms that do not need them – that is, those firms with low liquidity shocks. More succinctly, the fact that, ex post, liquidity is inappropriately distributed across firms is what makes it insufficient (as in Freeman [1996]). Forming an intermediary allows for a better ex-post allocation of liquidity and, in this way, improves outcomes. In practice, the way this happens is that some of the credit

lines are not fully drawn upon, permitting the liquidity to be more exclusively dedicated to satisfy the demands of firms that need to draw heavily on their credit lines.

With independent liquidity shocks, introducing intermediaries is "enough" and there is no role for government intervention. The case when the liquidity shocks are correlated, instead, provides a (potential) justification for government-provided liquidity. The basic idea is that the government can issue (noncontingent) claims on future tax revenue that firms then can hold (as a store of value) and potentially sell if and when they experience a large enough liquidity shock. If taxation is distortionary, the optimum may require that the bonds issued by the government sell at a premium (a liquidity premium). Then, given that premium, firms will adjust their demand for liquidity, which in turn determines the size of the liquidity shocks that they can withstand. Essentially, since holding government bonds is expensive, firms adjust their decisions in order to economize their reliance on those bonds.

With *noncontingent* government bonds, the best-attainable allocation (appropriately defined) may involve partial liquidation of investment (or the liquidation of some, but not all, firms with a given value of the liquidity shock). The implementation of this optimum is nontrivial: as Holmstrom and Tirole explain, one possibility would be to have some of the firms issuing both equity and short-term debt (with a specific, and somewhat unrealistic, covenant).

The government, though, can actually improve the allocation by issuing *state-contingent* bonds that pay a positive amount only when extra aggregate liquidity is needed. The rationale for this is simple: noncontingent bonds will provide excess liquidity in most states of the world. Since this liquidity is expensive to create – as it involves distortionary taxation – it is optimal to minimize the production of excess liquidity. Holmstrom and Tirole use this result to motivate possible state-contingent policies (monetary and fiscal) that are aimed at managing the provision of aggregate liquidity.

In a companion discussion of aggregate liquidity shortages, Holmstrom and Tirole (1996) explicitly consider discount window lending as a possible (state-contingent) policy that could be used to achieve the socially optimal allocation without resorting to the more uncommon state-contingent bonds. In principle, one interpretation would be that the counterpart of the necessary premium on government bonds is to have a penalty rate at the discount window. This interpretation, then, provides a possible justification for an *optimal penalty* rate at the discount window on the basis that government production of

liquidity involves distortionary taxation.[15] Note finally that the discount window would be used only in situations when there is an aggregate liquidity shock and "insufficient" private claims. This implies that the discount window would be particularly active at times that are often considered crisis-like situations.

The model by Holmstrom and Tirole (1996) highlights the close connection between monetary and fiscal policy when liquidity demand refers to the need to access riskless claims issued by the government. It suggests that in many cases the LLR function could be handled directly by the fiscal authorities, in particular if government bonds and reserves serve an equivalent role for solving the issue at hand.

Deposit Insurance, Bank-Failure Resolution, and Bankers' Incentives

Freixas, Parigi, and Rochet (2004) present a model where bank depositors are fully covered by deposit insurance and it is the government (i.e., the party providing insurance to depositors) that needs to design the appropriate framework to manage bankers' incentives. The optimality of deposit insurance is not addressed in the paper. In fact, depositors are assumed to be deep-pocketed, risk-neutral individuals. The paper can be seen, then, as an effort toward understanding how to organize the banking system, and the relevant government interventions, given that a decision has been made to provide deposit insurance.[16]

There are three relevant periods in the model. In the initial period, bankers take deposits and complement those funds with their own capital, which is assumed to be a fixed amount. With those funds, bankers make risky investments. In the interim period, bankers find out the state of their finances. Finally, in the last period, payoffs are realized.

Three situations are possible in the interim period: the banker may be solvent or insolvent, and if the banker is solvent, it may or may not experience a liquidity shock in the form of deposit withdrawals. Furthermore, if the banker is insolvent, it could in principle engage in "gambling for resurrection" by borrowing some funds and investing

[15] The traditional Bagehot doctrine on discount window lending involves a penalty rate as well. This penalty is often motivated as a way to control moral hazard. However, the implications of a penalty rate on incentives can be subtle. Castiglionesi and Wagner (2012), for example, demonstrate that under some conditions a penalty rate may actually increase moral hazard.

[16] Repullo (2000) studies the conflict of interest between a deposit insurance agency and a central bank confronting the decision to lend to a bank in need of liquidity. See also Kahn and Santos (2005).

them in a way that gives the banker a small probability of recovering from insolvency.

When the bank is solvent, investment has a positive probability of paying out in the last period. Investment by insolvent banks, instead, is sure to not produce any payoff in the future, unless the bank gambles for resurrection and succeeds. Bankers' incentives play a role in the initial and interim periods.

First, in the initial period, bankers can exert some costly effort to increase the probability that the bank will be solvent. In turn, if the bank is solvent, then the banker can exert some effort in the interim period to increase the probability that investment will be successful. In both cases, exerting effort is socially desirable, and the job of the government is to design a system that compensates bankers so as to induce them to do what is best for society.

Increasing bankers' compensation reduces the resources left to pay depositors (since bank capital is assumed fixed). Hence, the only way to accommodate higher banker compensation is to reduce total deposits and, hence, total investment. Since investment is productive, reducing the level of investment is costly for society. In other words, there is a trade-off between compensating bankers and the level of total investment undertaken by banks. The incentive design problem then involves compensating bankers with the minimal amount that would still induce them to exert effort. This incentive problem is similar to the one analyzed by Holmstrom and Tirole (1998).

In some situations, providing incentives to bankers in the interim period involves paying them enough that it is optimal for them to also exert effort in reducing the probability of insolvency in the initial period. In those cases, illiquid banks can resort to the interbank market for funds without compromising the optimality of the allocation.[17] The more interesting cases occur when extra incentives are needed to reduce the bank's probability of insolvency.

In principle, an unverified insolvent bank can pretend to be an illiquid bank, take a loan in the interim period (of the same amount as illiquid banks), and use those funds to gamble for resurrection. To avoid this situation, insolvent bankers need to be compensated to agree to identify themselves as insolvent (and not engage in socially wasteful gambling for resurrection). This may require that, upon failure, shareholder value is not fully wiped out. In this sense, bank-resolution policies are an important component of the incentive scheme.

[17] Strictly speaking, the optimal allocation is implementable only if interbank loans are not subject to repayment risk. This requires that the size of the loan be small relative to the lower bound on the return of investment.

Given that insolvent bankers receive a positive compensation after declaring bankruptcy, the efficient allocation requires that banks borrowing funds in the interim period pay a premium for those funds. There are two reasons for this: first, by charging a penalty rate, the government reduces total shareholder compensation and manages the trade-off between incentive provision and total investment. Second, the penalty rate is a way to induce sorting: once funds are offered at a premium, insolvent banks have no incentives to borrow, while illiquid banks still do.[18]

One way to implement the optimal allocation is to have interbank loans be junior to claims of the deposit insurance fund and have discount window loans be senior to both. Under such a situation, the central bank has the ability to fine-tune the pricing of discount window loans, establishing the appropriate penalty rate consistent with optimality, which can still be below the alternative rate that banks would need to pay in the interbank market (where loans are uncollateralized, junior claims).

To close this discussion, it is worth pointing out that in the Freixas-Parigi-Rochet model this kind of arrangement where discount window loans (at a penalty rate) can be part of the optimal way to organize the banking system in the presence of deposit insurance depends on several particular conditions on parameters. In many other situations, the discount window has no clear role in the sense that it cannot improve on what can be achieved with only an interbank market and, in those cases, it may not be possible to decentralize the optimal allocation.

Unique-Equilibrium Coordination Failures

Rochet and Vives (2004) study a banking problem where the assumed banking arrangement may create a coordination failure. While the model has the flavor of the Diamond-Dybvig model, there are some significant differences. To start, contrary to Diamond and Dybvig (1983), this paper does not focus on understanding the constrained-optimal allocations without any exogenously imposed institutional constraints. In fact, some features of the banking arrangement are taken as given without providing explicit micro-foundations. The emphasis, instead, is in understanding the implications of those features once they are in

[18] This is an alternative theoretical justification from the one provided by Holmstrom and Tirole (1996) for charging a penalty interest rate on loans at the discount window. To the extent that this penalty rate is more about managing bankers' incentives, it is closer in interpretation to Bagehot's doctrine.

place.[19] The authors discuss throughout the paper possible avenues to approach the micro-foundations question. Their overriding objective, however, is to try and keep the framework as simple as possible to be able to employ the global-games methodology (Morris and Shin 2003) that pins down equilibrium even in the presence of coordination failures.

Banks in the model have some capital and receive some deposits from investors. With those resources, the bank can invest in risky assets or in reserves. Investment takes two periods to mature, at which time it delivers some returns. In an interim period, before investment matures, depositors are entitled to withdraw their deposits from the bank. There are no liquidity shocks (there are no impatient agents of the Diamond-Dybvig type). Instead, depositors receive idiosyncratic signals about the future return of investment and, for each depositor, if her signal is (sufficiently) bad, then she decides to withdraw her money from the bank early (in the interim period).

When some depositors withdraw early, the bank can use the reserves to pay those depositors. If withdrawals are higher than reserves, the bank sells the investment in the market at a discount. The discount is exogenously assumed, and it stands for possible fire sales (or other sources of liquidity premia). The authors discuss how adverse selection could motivate fire sales, but this aspect is not explicitly modeled. Finally, if the bank cannot repay the promised amount to all depositors (early and late withdrawers), it fails. The bank is not allowed to adjust payments to depositors; that is, payouts are noncontingent, unless of course the bank fails.

Depositors' preferences are not explicitly spelled out. Rather, depositors follow what the authors call a *behavioral rule*: each depositor wants to withdraw if her individual assessment of the probability that the bank will fail is high enough. The authors postulate that this is a reasonable rule to capture the behavior of fund managers investing in, say, jumbo CDs at banks. The authors argue that this interpretation is more in line with "modern" versions of bank runs (where withdrawals by wholesale-funding sources play a prominent role).

There are situations in the model when the bank is solvent but may still fail because of the need to accommodate early withdrawals by liquidating assets at a discount. The idea is that when there are fire

[19] Broadly speaking, this is the approach also taken by Freixas, Parigi, and Rochet (2004) and, to a lesser extent, by Bhattacharya and Gale (1987) and Williamson (1998). As we discussed before, this approach is open to the criticism that any attempts at endogenizing the institutional arrangements may require new assumptions that could have important implications for (and potentially undo) the results discussed by these authors.

sales, if enough agents withdraw early, then the bank liquidates assets (i.e., invested resources) at a discount and there are fewer resources available to pay other depositors. This makes failure more likely and feeds back to the number of withdrawals, increasing it as more depositors conclude that the bank will fail given their private signal of the value of the return on investment.[20]

The discount window is assumed to have an informational advantage originated in the central bank's supervisory powers. This information allows the discount window to recognize the "true" value of the assets of the banks, not the one implied by the fire sales. In this way, using the discount window, the central bank can avoid any early liquidation of assets and, hence, the failure of solvent banks.

The central bank is assumed to also have access to funding at no extra cost. In other words, the central bank can access resources without having to increase distortionary taxation. These extra resources available to the central bank are not explicitly modeled, and it is not clear what would be optimal if these resources were explicitly taken into account from the start.

In the model, discount window credit should not be provided at a penalty rate. Rochet and Vives discuss many of the factors not present in the model that would suggest that a penalty rate may be appropriate – for example, if the central bank has better information relative to the private sector, but not perfect information. There is also a discussion of the possibility that the private sector could provide lines of credit to banks and then actively monitor them (as suggested by Goodfriend and Lacker [1999]). The authors point out that this could be an appropriate approach when there are no central-bank advantages in supervisory knowledge and financial capacity.

Late in the paper (in Section 7), the authors sketch a justification for the deposit contract that allows agents to withdraw in the intermediate period. As in Freixas, Parigi, and Rochet (2004), the idea is that the bank manager needs to exert effort to improve the distribution of possible investment returns, but that effort is costly and not verifiable. A way to give incentives to bank managers is to allow depositors to withdraw early. The bank manager is assumed to specially dislike bank failure in the intermediate period, and depositors can "discipline" the manager by threatening to withdraw early (Calomiris and Kahn 1991).

[20] Depositors follow a threshold rule: if the signal is below a threshold, then the depositor withdraws early. The threshold gets determined in equilibrium and depends on the strategy of other depositors (it is a fixed point) because if more depositors withdraw, the bank is, for a given return on investment, more likely to fail (as more withdrawals mean more early liquidation at discounted values).

A couple of interesting situations may arise in this case. First, it is no longer efficient for the central bank to intervene in a way that rules out all possible early bank failures. Some early bank failures are necessary to provide incentives to bank managers. Still, the equilibrium without intervention may result in too many early bank failures. Even if the bank is insolvent, there are cases where it is efficient for the central bank to provide credit in the intermediate period to avoid early liquidation of assets at discounted (fire sale) values. The bank will still fail in the second (final) period, but losses would be lower (even after the loan from the central bank is fully paid back).

Perhaps more interesting is the fact that there are some cases where the central bank needs to intervene and close down a bank in the intermediate period even though the bank is solvent. The authors interpret this as a "prompt corrective action" rule. The idea is that sometimes, to give bank managers incentives to exert effort, the bank needs to be liquidated early even if the bank would be solvent (assuming that the central bank provides appropriate discount window liquidity). Hence, a central bank that has a discount window open to all banks needs to complement that policy with a "prompt corrective action" policy when the incentive problem of bank managers is severe enough.

Inalienable Human Capital and Banking

In a series of papers published in the early 2000s, Diamond and Rajan developed a comprehensive theory of banking. Initially, they studied what banks do and the optimal structure of banking contracts and of banks' balance sheets. In a second stage, they extended the model to address systemic banking crisis (Diamond and Rajan 2005). This extended model is the one used here as a basis for the discussion.

There are three types of agents in the economy: investors, bankers, and entrepreneurs. Entrepreneurs have projects that need funding. Investors have funds that could be used to fund those projects. Funding is in short supply, though, so only a portion of the projects can actually get funding. All projects pay the same return after some period of time. While ex ante all projects are identical, ex post some projects take longer to mature. In other words, a proportion of the projects pay their return early ("early projects"), and the rest ("late projects") pay some time later.

Investors need to consume early (that is, at the time when the early projects pay out their return). Bankers and entrepreneurs, on the other hand, can consume late.

There is an information friction that complicates the funding of projects. Entrepreneurs cannot commit ex ante to run their project

after receiving funding, and their human capital is essential for running the project successfully. One way to think about this lack of commitment problem is that the courts system cannot force entrepreneurs to dedicate their *inalienable* human capital to the process of running a project (Hart and Moore 1994). As a result, after a project receives funding, the entrepreneur can threaten to withdraw his human capital and induce a renegotiation of the terms of the loan – an instance of the well-known hold-up problem.

Bankers have a technological advantage over investors. In particular, bankers are able to learn about the project and run it if necessary. The return that the project delivers when run by a banker is a fraction of what the entrepreneur can get, but it is not zero. Hence, the banker can fund the entrepreneur up to the amount that the banker would be able to get when running the project himself. If the entrepreneur tries to renegotiate, the banker takes over the project and runs it himself. Knowing this, the entrepreneur does not attempt to renegotiate. Here, it is important that the loans to entrepreneurs are callable (i.e., that the bank can ask for repayment at any time and take over the project if the repayment does not happen).[21]

Since only investors, not bankers, have the resources to fund the projects, channeling funds to entrepreneurs requires that investors make deposits at the banks and those banks make loans to the entrepreneurs. As a result, a second hold-up problem arises: the banker, after receiving deposits from investors, could try to renegotiate the contract by threatening to not collect from the entrepreneurs.[22] One way to solve this second hold-up problem is to create a collective action problem among the bank's depositors. In particular, the bank can offer uninsured demand deposits that are paid out on a first come, first serve basis and obtain deposits from a large number of investors. Courts can enforce deposit contracts as long as the bank has funds. Under these conditions, there is an equilibrium in which investors/depositors run whenever the bank attempts to renegotiate down the payments associated with the deposit contract. If a depositor thinks that other depositors will run when threatened with renegotiation, then it is in her best interest to run as well, even if in the end depositors, as a group, obtain less than if the run would not have occurred.

---

[21] Diamond and Rajan (2005) have a discussion of the empirical relevance of callable bonds in their paper and further point out that a significant proportion of outstanding commercial and industrial loans in the U.S. are of very short maturity (which makes them essentially callable).

[22] No one else but the bank that initially funded an entrepreneur has the ability to collect from the entrepreneur, so the loans are illiquid from the perspective of the bank. That is, the bank cannot sell its loans in the market.

Similar to Rochet and Vives (2004), here runs on the bank are a way to provide appropriate incentives to bankers. Entrepreneurs cannot receive funding directly from investors due to an extreme hold-up problem. The possibility of runs, in turn, controls the hold-up problem between bankers and investors (depositors). This allows funds to flow from investors to entrepreneurs, through bankers, in a way that otherwise would not be possible.

Bankers can also restructure projects at any time before the projects mature. A restructured project yields some resources immediately and some resources in the future. The payouts from a restructured project can be collected by anyone (there is no hold-up problem in that case). However, restructuring projects is costly in the sense that a restructured project yields fewer resources than the initial investment.

Each bank is subject to an idiosyncratic shock that determines the fraction of projects in its portfolio that are maturing early. This shock is crucial for bank solvency. Projects that mature early provide resources to pay initial investors – all of them needing to consume early. So, a high fraction of projects maturing early makes the bank more likely to be solvent. The bank can also access a market for liquidity and try to obtain resources by borrowing against the return from the late projects. How much liquidity the bank can obtain depends on the market interest rate. If the bank cannot raise enough liquidity to pay all initial investors, then it is deemed insolvent.

The market for liquidity at the time when initial investors need to consume is a key market in the model. The interest rate in that market plays a role in determining bank solvency and project restructuring. In turn, demand and supply of liquidity in that market depends on depositors' and banks' decisions. Let us now briefly discuss how demand and supply of liquidity in that market get determined and how they depend on the interest rate.

Project restructuring impacts both supply and demand of liquidity. Both solvent and insolvent banks may engage in project restructuring. The decision of solvent banks to restructure late projects depends on the market interest rate. If the interest rate is low, then a solvent bank will choose to continue all projects. For intermediate values of the interest rate, a solvent bank will choose to restructure only enough projects to pay back initial investors. Finally, if the interest rate is high, a solvent bank will choose to restructure all of its late projects. The reason behind this pattern of behavior is simple. To continue funding late projects, a bank needs to attract new deposits. If the interest rate is high, deposits are costly. If, instead, the bank just restructures late projects, it obtains immediate liquidity.

Bank solvency also depends on the market interest rate. In principle, a bank can become insolvent just because the interest rate in the market is too high. Since borrowing is backed by the future discounted value of late-project returns, when the market interest rate is very high the bank cannot borrow as much. As a result, the bank has access to less liquidity and may not be able to pay initial investors in full.

The demand and supply of liquidity in the market also depend on bank solvency. One basic source of liquidity in the market is the income of entrepreneurs with projects that mature early and have a loan from a solvent bank. Due to limited pledgability and lack of commitment, these entrepreneurs' income is higher than what they need to repay the bank. Since entrepreneurs do not need to consume early, they can reinvest these extra resources by lending to solvent banks in need of funding.

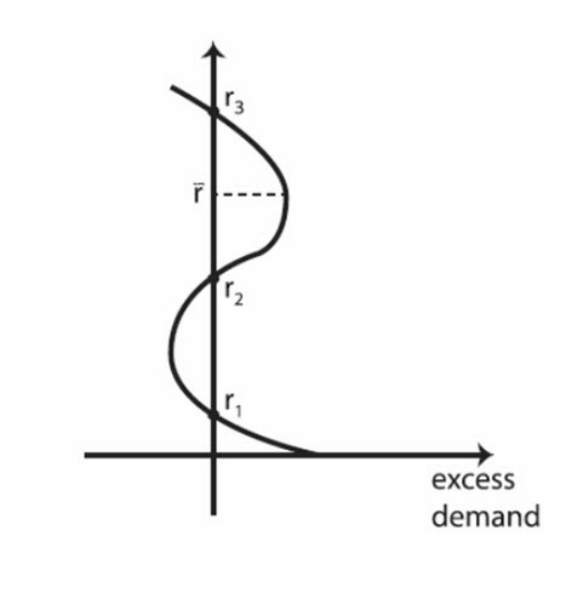
Solvent banks with late projects in their portfolios need to obtain extra liquidity in the market to pay back initial investors without having to restructure those projects. The banks can, then, use the future return on those late projects to guarantee their ability to repay new loans when they become due.

Bank liquidation impacts the demand and supply of liquidity in complex ways. Restructured projects tend to increase the demand for liquidity in the market. The reason is that restructured projects generate pledgable future income that banks would want to use in order to borrow extra liquidity. Bank liquidations, by forcing project restructuring, also increase the demand for liquidity in the market. Furthermore, bank runs trigger restructuring of early projects, which reduces the supply of liquidity because entrepreneurs receive less income to reinvest.

The multiple effects and interactions between the market interest rate and banks' decisions imply that the excess demand function for liquidity may be nonmonotonic. Conditional on a given number of bank failures, increases in the interest rate tend to lower the excess demand for liquidity. However, changes in the number of bank failures can change this relationship. Both solvent and insolvent banks demand liquidity from the market. When movements in the interest rate push banks out of the solvent group and into the insolvent group, the impact on total demand (and supply) of liquidity can result in a segment of the excess demand function with positive slope (see Figure 1).[23]

[23] An increase in the interest rate can increase the excess demand for liquidity when the proportion of early projects is high in banks that switch from solvent to insolvent due to the interest rate increase. When banks with a high proportion of early projects become insolvent, the resulting restructuring of those early projects creates an extra demand for liquidity in the market.

Figure 1 Excess Demand for Liquidity



We have discussed so far how projects get funded and why banks are needed in that process. We also discussed the determination of demand, supply, and interest rates in the market for liquidity. A crucial factor driving policy interventions in the model is the possibility of bank insolvency. The timing of the arrival of information is important for this issue. In particular, information about banks' idiosyncratic shocks arrives *before* early projects pay their return. Initial investors observe these shocks and have rational expectations about market interest rates. At that time, then, they can calculate whether a bank will be solvent given the expected interest rates. If the bank is insolvent, initial depositors run on the bank. In response, the bank restructures all projects, even those that are expected to mature early. This surge to generate liquidity is socially very inefficient. In fact, note that liquidity is not really needed at that point. Depositors are demanding liquidity even before anyone needs to consume. This nonfundamental demand for liquidity is a direct consequence of the self-fulfilling run on demand deposits.[24]

[24] Because the bank is insolvent, depositors expect a (necessary) renegotiation of their deposit contracts in the near future. Since depositors also expect that other de-

The payoffs associated with bank deposits are not allowed to be contingent on the realization of the idiosyncratic shocks. This contract incompleteness is particularly consequential when a high proportion of late projects put a bank in an insolvency position. In equilibrium, depositors at such a bank anticipate renegotiation and, hence, decide to run. This run is not essential for disciplining the banker. The banker, regardless of behavior, just does not have enough resources to pay every initial investor in full. Adjusting down payments would be more efficient than liquidating the bank and restructuring all projects. This is an important source of inefficiency in the model.

Diamond and Rajan are forthright about the importance of this assumption. For example, they tell us that they "do not allow contracts to be directly contingent on the state," which they assume "is observable but not verifiable" and explain that their "analysis is positive – to show what happens when there is an ex-post solvency problem or an aggregate liquidity shortage given the use of demand deposit contracts." To be sure, they confirm that "if there was no uncertainty about the ex-post state of nature, or if there were complete markets, no such conditions would arise."[25]

Diamond and Rajan are particularly interested in periods of crisis. One way to think about banking crises in this environment is to consider the case when the average fraction of early projects across banks in the economy is a random variable. When most banks have lots of early projects, those banks are solvent and bank runs do not occur. However, when the proportion of early projects is low for many banks, some of these banks become insolvent and experience runs. Depending on how bank failures and the consequent restructuring of projects impact market liquidity, situations that can be regarded as inefficient banking crises can develop in equilibrium.

Diamond and Rajan discuss examples of this kind of inefficient crisis. In particular, they describe a situation where the banking system, in principle, could satisfy the liquidity needs of investors by

---

positors will react to the threat of renegotiation by withdrawing their deposits, the expectation of that renegotiation immediately triggers the run.

[25] The deposit contracts considered by Diamond and Rajan are *demand* deposits in the sense that withdrawal can happen at will and at par, at any time. The bank does not offer, for example, term deposits for which withdrawal cannot happen until the early projects have matured. In principle, the "callable" feature of deposit contracts can be formally justified within the structure of the model. Since the banker can try to renegotiate the payment to initial investors at any time (by threatening to not collect from entrepreneurs) the deposit contracts have to be "runnable" at all times (to discipline the banker). If deposits were term deposits, then, before the deposit matures, the bank could try to renegotiate the claims of initial investors and these investors would not be able to run on the bank because the courts would not enforce repayment until the term of the deposit expires.

restructuring only late projects. However, in equilibrium all banks experience a run and all projects, including early maturing ones, are restructured. Diamond and Rajan call this situation a *systemic meltdown*.

To gain some intuition over why a meltdown can happen, notice that solvent banks do not restructure late projects when the interest rate is low, and hence the supply of liquidity in the market is also low. With an excess demand for liquidity, the interest rate tends to increase, which can make more banks insolvent. As insolvency produces more restructuring of early projects, this reduces even further the supply of liquidity in the market – increasing excess demand, which then can be increasing in interest rates. If the interest rate keeps rising, eventually all banks are subject to a run and all projects get restructured. This process generates a form of financial "contagion" that spreads via changes in market interest rates.[26]

As Diamond and Rajan explain, the timing of arrival of information is crucial for this "contagion" outcome. In the model, information about the aggregate state (the distribution of shocks to banks) arrives earlier than the time when liquidity is produced and consumed, and depositors redeem their claims immediately after the arrival of this information, in anticipation of future liquidity shortages. The resulting wave of withdrawals forces banks to restructure projects – even early projects that would otherwise be a source of liquidity in the market.

The inefficiency of equilibrium outcomes opens the door to potentially beneficial policy interventions. Diamond and Rajan define two benchmark policy interventions: a *pure liquidity infusion* and a *pure recapitalization*. A pure liquidity infusion involves providing loans to banks at the prevailing interest rate. In a pure recapitalization, banks receive a transfer (a subsidy) in the form of claims that can be traded in the market to obtain liquidity and avoid failure. The policy authority has the ability to tax some agents in order to fund the interventions. Diamond and Rajan argue that, in their model, any financial market intervention can be viewed as a combination of these two pure forms of intervention.

Of particular interest for the subject of this article – i.e., discount window policy – is the case of a pure liquidity infusion. To fund the loans, the policy authority taxes investors after they had a chance to withdraw from their bank. Under certain conditions, this intervention can increase market liquidity and lower the interest rate so that fewer

[26] Freixas, Parigi, and Rochet (2000) and Allen and Gale (2000) are classic papers studying contagion originating in more standard types of spillovers in banking – such as when one bank's failure directly creates losses on other banks' loan portfolios.

banks are insolvent and fewer inefficient runs happen in equilibrium. This policy intervention is clearly not a Pareto improvement since some investors are taxed and end up consuming less. Still, the policy may produce a welfare improvement by reducing inefficiencies associated with the unnecessary liquidation of banking assets.

The nonmonotonicity of the excess demand function in the market for liquidity implies that, under certain conditions, the model also has multiple equilibria. Figure 1 illustrates such a situation (see Lemma 3 in Diamond and Rajan [2005]). The three possible equilibrium interest rates are r_1, r_2, and r_3. When the market interest rate is above $\tilde{r}$, only self-sufficient banks (which do not need to obtain liquidity in the market to pay back investors) are solvent. In the equilibrium where the interest rate is r_3, the initial investors expect a high interest rate that makes all but the self-sufficient banks insolvent. As a result, all banks that are not self-sufficient suffer runs, and the liquidity in the market confirms the high expected interest rate r_3.

To avoid a highly inefficient situation like the one associated with interest rate r_3, a policy authority (a central bank) can put in place a discount window that offers loans to banks at an interest rate between r_1 and r_2. Under such policy, the only equilibrium in the market would be the more efficient situation associated with interest rate r_1. In fact, discount window activity would actually not be observed in equilibrium. The discount window acts just as a mechanism to coordinate agents' expectations. It is important to recognize here, though, that such policy only works if agents believe that the central bank would have access to sufficient tax revenues were actual discount window lending to become necessary. In other words, policy credibility is essential, even if never tested.

Widespread Pessimism and Flight-to-Quality Episodes

Caballero and Krishnamurthy (2008) study an economy with a continuum of agents who may experience a liquidity shock (an urgent need to consume). The shocks are correlated across agents: in particular, the economy may experience one or two "waves" of shocks. If the economy experiences one wave of shocks, then half the agents in the economy receive the shock. Who in the population receives the shock is random. If the economy experiences two waves of shocks, then those agents who did not receive a shock in the first wave receive a shock in the second wave. Agents know the probabilities of the economy experiencing none, one, or two waves of shocks, but they do not know if they will receive a

shock in the first wave or in the second wave if the economy experiences two waves.

There is no asymmetric information in the economy, and there are complete markets. Hence, agents can enter insurance contracts (or buy and sell contingent claims) in order to insure the liquidity shocks. In the optimal allocation, agents buy more insurance for the eventuality of receiving a liquidity shock in the first wave than in the second wave. This is the case because the first wave is more likely (the economy can experience a second wave only if it has already experienced a first wave).

While in the Caballero and Krishnamurthy economy there is no sequential service constraint (Diamond and Dybvig 1983), there is sequentiality in the revelation of the state: first agents find out that there has been a first wave and, only after making/receiving the payments corresponding to that contingency, agents find out if there is a second wave.

Note that after entering an insurance agreement, if an agent gets a shock in the second wave, we could say that he is ex-post unlucky. That is, the eventuality that was less likely and the one for which the agent bought less insurance has actually happened. Caballero and Krishnamurthy consider the possibility that agents could become overly pessimistic about this contingency. If such is the case, this effectively increases the probability that the agents assign to the event of being affected by the second wave of shocks. As a result, then, agents increase the amount of insurance that they buy to cover that contingency. If this bias is large enough, agents basically insure the second-wave shock as much as they insure the first-wave shock (note that if they do, then there is no longer a sense in which an agent who receives the shock in the second wave is unlucky relative to an agent who receives a shock in the first wave – they both consume the same).

Caballero and Krishnamurthy use the model to think about a situation where there is a *flight-to-quality* event. In normal times, agents in the economy use an unbiased estimate of their likelihood of being in the second wave of liquidity shocks. However, when something unusual and unexpected happens (something that does not directly affect them but makes them pessimistic about their prospects), agents become overly pessimistic about their own situation (not the situation of the aggregate economy) and act as if they were more likely (than what they really are) to experience a second-wave shock. As a result, agents start buying more insurance for the second-wave shock and lower the amount of insurance they buy for the first-wave shock. The direct implication is that there ends up being much less liquidity during the (much more likely) first wave of shocks.

Caballero and Krishnamurthy study optimal intervention by a central bank in their environment. The central bank does not have more information than agents, but since the central bank cares primarily about "aggregates" and not about particular agents, it is not exposed to the pessimistic biases of individual agents.[27] During a flight-to-quality episode, the central bank could improve outcomes by (somehow) inducing agents to insure less against the second-wave shock and more against the first-wave shock.

Caballero and Krishnamurthy consider the case when the central bank has access to resources ("collateral") that private agents do not have. They argue this is similar to the assumption in Holmstrom and Tirole (1998) where the central bank can exploit the power it has to impose future taxes. The point they want to make is that the benefit of having and using those resources during a flight-to-quality episode is higher than the direct cost of obtaining those resources. The policy intervention calls for the central bank to promise agents that it will provide them with resources if the second-wave shock hits. Agents anticipate the central-bank contingent-transfer and reduce the amount of insurance they obtain against the second-wave shock. At the same time, they increase the amount of insurance they obtain for the first-wave shock (the more likely shock). Note that the intervention is only in the rare event that two waves of shocks hit the economy. In that sense, the paper provides justification for a "last resort" intervention policy. Note, however, that the policy influences outcomes not just when they happen but in all eventualities, because it changes agents' ex-ante decisions and improves the insurance arrangements (which would otherwise be inappropriate due to biases that agents have in assessing the probability of extreme, unfavorable individual events).

It is interesting to note (as the authors do) that the LLR policy is aimed at correcting decisions by private agents to over-insure some shocks and under-insure others. For this reason, moral hazard is not a big problem. It is the case that agents reduce private insurance for the shocks that are insured by the central bank, but agents also increase the amount of private insurance on the shocks that were previously under-insured. There is a certain degree of complementarity between the public insurance of the second-wave shock and the private insurance of the first-wave shock (when one increases, the other increases too). The provision of public insurance for some (unlikely)

[27] A key feature that allows the central bank to intervene efficiently without having any a priori advantage over agents (in information or perceptions) is that the source of the problem is not that agents are overly concerned about aggregate shocks. They are overly concerned only about the impact of those aggregate shocks on their individual outcomes. So, when the central bank aggregates outcomes, the biases disappear.

shocks helps to correct the under-insurance of other (more likely) shocks, which are not publicly insured. Broadly speaking, the central-bank backstop in this model does what central-bank backstops generally do – it reduces agents' incentives to insure themselves against the event that the central bank is backstopping them on. Usually, the weakening of incentives results in inefficiencies (due to moral hazard). Here, however, given that agents were initially over-insuring the risk in question, the effect of the backstop is to correct a distortion and improve efficiency.[28]

It seems likely that an agent with deep pockets and an unbiased assessment on the probability of different shocks to individuals (or with the ability to diversify/aggregate/pool risk across individuals) would be able to sell insurance at a profit. The paper is clear in explaining that if there are sufficient resources in the economy, then the optimal allocation involves full insurance of all shocks and misperceptions are irrelevant for allocations. Only when liquidity is limited (that is, when resources available in the short run are limited) agents' misperceptions create a problem. Still, the paper does not consider a situation with heterogeneous agents, some with plentiful resources and some with limited resources, such that some beneficial trade could occur. It seems likely that in those circumstances, the misperceptions of some agents could create extraordinary benefits for other agents (those with access to liquidity).

In the model, the central bank is just an agent who has access to resources via taxation and is able to redistribute those resources to correct inappropriate private insurance arrangements. But the central bank is benevolent, so it is not exploiting agents' misperceptions.[29] The extent to which a private deep-pocketed and self-interested agent could improve the overall situation is, hence, not clear.

There is no formal explanation for what triggers the switch of agents to a pessimistic state. Caballero and Krishnamurthy provide an informal discussion of the situations that are most likely to trigger such perception shifts. New (unanticipated, unknown) events and

[28] Intervention could still create moral hazard if agents have to incur costs to become better informed about the economy and the nature of the shocks. This may be especially important if such information could make them less prone to misperceptions that make them overly pessimistic.

[29] Note that defining a benevolent central bank is not without complications. Basically, the central bank ends up being paternalistic in the sense that it is using probabilities to assess outcomes that are different from the priors used by the agents. If the agents take their priors as fixed features of their preferences, then using different probabilities is not consistent with choosing agents' most-preferred allocations. The authors discuss this at length in the paper and provide several interpretations that justify their use of a paternalistic central bank.

innovations play a prominent role in their discussion. Based on these informal discussions, they conclude that interventions that are aimed at dealing with new and unknown situations and that create and coordinate understanding of such situations (such as facilitating discussions among major market participants, in the spirit of the intervention by the New York Fed during the collapse of the hedge fund Long-Term Capital Management in 1998) could be beneficial.[30]

Limited Enforcement in the Interbank Market

Gertler and Kiyotaki (2010) explore a dynamic macroeconomic model with a financial intermediation sector and frictions in the ability of the financial sector to obtain external funding. In the model, there are a large number of firms and a large number of islands (locations, sectors). While labor can move freely across firms and islands, capital cannot. Furthermore, each period the firms in some islands (but not in all) receive an opportunity to invest. Firms that are able to invest generate extra demand for liquidity for the financial intermediaries (banks) of their island. In consequence, some banks in the economy value liquidity more than others, and an interbank market for funds can emerge to exploit the gains from trade.

Financial intermediaries are assumed to be necessary to channel funds from depositors (and potentially other banks) to productive firms. There are a large number of households with intertemporal preferences over consumption and leisure, and their behavior can be characterized using the problem of a representative household. Gertler and Kiyotaki make further technical assumptions to facilitate aggregation in the production side of the economy and keep the model tractable. Without any financial frictions, the economy actually reduces to a (relatively standard) Real Business Cycle model.

Financial intermediaries take deposits from households to then fund firms' capital and investment. The friction in the intermediation process is a version of the agency problem studied by Kiyotaki and Moore (1997), which endogenously imposes a limit on the ability of banks to obtain external funding. The basic assumption is that after a bank obtains funds and acquires claims on productive firms, the bank can divert a fraction of those claims for its private benefit and default on its creditors. To avoid default, creditors are willing to fund only a portion

[30] A topic that remains largely unexplored in the theoretical literature is the role of the discount window during system disruptions (Lacker 2004; and Ennis and Price 2015). A notable exception is Martin (2009).

of the total claims held by a bank. Bank net worth is needed for the rest.[31]

The evolution of bank net worth depends on past investment by the bank and the cost of its funding. Gertler and Kiyotaki assume an exogenous exit rate from banking so that the accumulation of net worth over time does not fully resolve the agency problem. Banks can attract deposits from any island in the economy, but they can only buy productive claims from firms in their island. Furthermore, banks access the deposit market before knowing which islands will receive the investment-opportunity shock. By the time the shocks are realized, the deposit market is closed. These assumptions induce a level of market segmentation that is crucial for the outcomes of the model.

The market for productive claims in an island has firms on the supply side and banks on the demand side. The demand for claims by banks is downward sloping since, for a given level of bank net worth (and hence funding), higher prices imply that banks can buy fewer claims. Increases in bank net worth, in turn, shift upward the demand curve for claims. In this way, the demand side of the market for productive claims is reminiscent of the "cash-in-the-market" mechanism in Freeman (1996).

The supply of claims also depends on the market price for claims. If claims sell at a high price, then investment is more profitable; firms that have an opportunity to invest, invest more; and hence there are more claims to be sold in the market. The equilibrium price of claims clears the market every period.

When bank net worth is not too high and the agency problem is operational, the marginal return from buying an extra productive claim is higher than the cost of the extra deposits necessary to fund that claim. Financial intermediaries are credit-constrained.

Since, in the deposit market, banks do not yet know if they are on an island where firms will have an opportunity to invest, they all raise the same amount of deposits. After the investment opportunities realize, however, banks can interact in an interbank market. Gertler and Kiyotaki consider the case when the same agency frictions that apply to deposits also apply to borrowings in the interbank market and the case when the frictions in the interbank market are less intense or not existent at all.

If the interbank market is frictionless, then the economy functions as if banks would not face any idiosyncratic liquidity shocks. Banks in

[31] There are no frictions in the relationship between banks and firms. Banks, instead of making loans to firms, buy claims on the future cash flow associated to capital investment.

investing islands borrow from banks in noninvesting islands. It is still the case that aggregate bank lending is constrained by aggregate bank net worth (the agency problem limits deposit funding), but there are no extra inefficiencies that arise from the combination of market segmentation and idiosyncratic shocks. In particular, the price of productive claims is the same in all islands.

When the interbank market is subject to frictions, on the other hand, the financial intermediation system cannot fully circumvent market segmentation via interbank trading. The supply of productive claims is higher in investing islands and, in consequence, the price of those claims is lower. This means that it is more profitable to acquire a claim in an investing island. But, funds from noninvesting islands may not flow to the investing islands if there is not enough net worth to support the extra funding (without compromising incentives). In equilibrium, then, the price of claims in investing islands is relatively low and, as a result, total investment is inefficiently low. The financial friction affects the real allocation of resources in the economy and its general macroeconomic performance.

Interbank borrowing impacts the incentive problem faced by banks in the same way that deposits do. Increasing interbank borrowing does not help increase the total amount of funds available to banks. Given borrowing constraints that are binding, more interbank borrowing has to be compensated with a decrease in deposits to keep bank leverage consistent with incentives.

Gertler and Kiyotaki discuss possible interventions that could be used to address the effects of interbank-market frictions on the real allocation of resources. They consider three types of policies: direct lending to firms, a discount window, and bank-equity injections. For this article, their discussion of the discount window is most germane.

If the central bank does not have an advantage over the private sector in its ability to enforce repayment, then discount window lending cannot improve outcomes. If, instead, the central bank has an enforcement advantage, then lending to banks in investing islands can increase total investment and improve economic outcomes in the economy. In fact, this enforcement advantage can make the discount window so attractive as to completely displace any private interbank trading. If both the discount window and the interbank market are to remain active, then the central bank needs to charge a penalty rate for loans at the discount window.

The discount window enforcement advantage makes it also a better alternative to deposits and it could displace deposits as a source of funding for banks, as well. To rule out this rather extreme (and unrealistic) outcome, Gertler and Kiyotaki impose a limit on the ability of the

central bank to efficiently evaluate borrowers and enforce repayment at the discount window. This limit translates formally into a capacity constraint that implies that the enforcement advantage of the central bank applies only up to a given maximum amount of discount window lending.

The financial frictions in the model are operative at all times, and the central bank intervention, to the extent that it can improve the situation, can do so also at all times. Gertler and Kiyotaki are particularly interested, however, in understanding the specific ways in which financial frictions may impact macroeconomic outcomes during crises. The way they model crises is by introducing a shock to the quality of capital. In effect, a negative shock to capital quality reduces the value of intermediaries' asset portfolios. Leverage amplifies the initial effect, reducing significantly the demand for productive claims in the economy. The fall in demand drives prices of claims down (a form of fire sales that arises due to the model's cash-in-the-market pricing), which feeds back into the balance sheet of banks, reducing their ability to finance capital even further. Moreover, the drop in current profits reduces the accumulation of bank net worth and tends to create more protracted crisis-like episodes. In other words, financial frictions can amplify and propagate the underlying shocks that drive crises in the model.

Using a quantitative example, Gertler and Kiyotaki discuss how policy can reduce the impact of a crisis shock in the economy. While they only consider direct lending as a policy response, they contend that discount window lending would have a similar power to dampen the macroeconomic implications of those shocks.

Adverse Selection

Philippon and Skreta (2012) study a model of financial contracting in the spirit of Myers and Majluf (1984), where private information and adverse selection generate suboptimally low levels of investment. The model has a large number of firms that need funding for a productive investment project. There is also a large set of risk-neutral investors with deep pockets. Firms also own "legacy" assets of different quality that influence the ability of firms to repay debt in the future. The quality of the legacy assets is private information, generating a distribution of different levels of repayment risks across firms.

In the absence of intervention, the interest rate on loans in the market reflects the average repayment risk of the set of firms asking for loans. Firms with low repayment risk end up facing a less attractive deal in the market and hence find investment less beneficial. Firms that decide not to invest do not seek funding in the market. In equilibrium,

only those firms with repayment risk above an endogenous threshold will undertake the investment projects and be active in the credit market. This is the case even though all firms' investment projects have a positive expected net cash flow. In other words, under perfect information, it would be optimal that they all invest.

Philippon and Skreta study optimal government interventions in this setup using a mechanism design approach. Interventions are optimal if they achieve a level of investment at minimum cost for the government. They show that, to increase investment and move the economy closer to efficiency, the government needs to make direct loans to firms at a lower rate than the one prevailing in a laissez-faire situation. The government program attracts firms with relatively low probability of repayment. As a result, the composition of the pool borrowing from private investors improves, allowing the private market interest rate to be lower and making the program consistent with an active private credit market.

The government lending program can be considered a version of the discount window. Because in some equilibrium situations there is selection in the participation decision of firms, with firms borrowing from the discount window having high repayment risk, the model can produce (equilibrium) discount window stigma (Courtois and Ennis 2010).[32]

An important contribution of Philippon and Skreta is to show that in their framework direct lending is the best way to design a government-intervention program – in the sense that it minimizes the cost of the intervention for a given level of targeted investment. In this way, the paper provides strong support for the idea that, in certain situations, using the discount window to make low-interest-rate loans to firms (banks) can enhance efficiency in the economy, particularly in periods when adverse selection seems to be the main friction thwarting the appropriate functioning of private credit markets.[33]

---

[32] Ennis (2017) studies in detail the implications for discount window stigma of the Philippon-Skreta model.

[33] In a very recent paper, Gorton and Ordoñez (2016) also study an economy with private information where a discount window can have an efficiency enhancing role. Interestingly, stigma plays a role in Gorton and Ordoñez's model as well, but instead of hampering the ability of the central bank to provide appropriate liquidity to banks, stigma gives incentives to banks not to reveal their borrowing activities and, in this way, increases the effectiveness of the government program.

Over-the-Counter Trading in the Interbank Market

Ashcraft and Duffie (2007) document that the intraday allocation and pricing of funds in the U.S. interbank market tend to reflect the decentralized nature of transactions in that market. Furthermore, their stylized facts are consistent with the predictions coming out from search-based theories of over-the-counter (OTC) financial markets, which have recently received significant attention in the literature (Duffie, Gârleanu and Pedersen 2005).

Afonso and Lagos (2015) study intraday interbank OTC borrowing and lending with a focus on fund intermediation – that is, situations where a bank borrows from another bank in anticipation of lending those funds to yet another bank during the same trading session. They compare the implications of the model with various indicators of activity in the U.S. federal funds market and conclude that the model does a good job of capturing those features. The discount window plays a relatively passive role in Afonso and Lagos' model with banks tapping the window at the end of the trading session if their balances are below a required value.

Ennis and Weinberg (2013) also study a model with bilateral bargaining and search frictions in the interbank market.[34] In the model, though, banks only get one chance to interact in the OTC interbank market, and hence no intermediation of the type highlighted by Afonso and Lagos takes place in equilibrium. Ennis and Weinberg (2013) focus on the issue of stigma at the discount window. They assume that banks can transact frictionlessly with the central bank in a way that is reminiscent of Williamson's (1998) assumption that the central bank (and only the central bank) can trade in all locations and, in that way, circumvent the assumed market segmentation.

In the Ennis-Weinberg model, banks own assets of heterogeneous quality that determine their loan-repayment risk. In effect, banks sell assets to investors in order to repay interbank loans. An investor may not be able to observe the quality of the asset held by banks but can use information on the activities of banks in the interbank market to try to infer the quality of assets. When a bank with a low-quality asset is trying to borrow in the interbank market, it may not be able to obtain a loan if its counterparty can evaluate the asset and determine that it is

[34] See Bech and Klee (2011) and Acharya, Gromb, and Yorulmazer (2012) for two other models of the interbank market where bargaining plays an important role. Acharya, Gromb, and Yorulmazer highlight how the discount window influences the outside option of the borrowing side of the bargaining game and, in that way, the outcome of the negotiations.

low quality. Banks that do not obtain funding in the interbank market may access the discount window. Under some conditions, banks with low-quality assets are more likely to be in such a situation. As a result, the pool of banks borrowing at the window is biased toward banks with low-quality assets, and borrowing at the discount window becomes an endogenous negative signal of the quality of assets held by banks. In equilibrium, some banks can become "reluctant" to borrow from the discount window and may prefer to borrow from the interbank market at interest rates higher than the discount window rate just to avoid being stigmatized in the asset market.

Ennis and Weinberg are mainly concerned with the positive implications of the model and particularly with respect to discount window stigma. There is less work done on the normative aspects of discount window lending in this type of model. However, there is a very active literature addressing the general issues related to OTC trading in financial markets. The lessons for discount window policy that could come out from that body of work have not yet been fully developed, but based on some recent contributions such as, for example, Lagos, Rocheteau, and Weill (2011), it seems to be a promising avenue for further research.

2. CONCLUSIONS

In this essay, I have reviewed a strand of the economic literature dedicated to gaining a better understanding of the role of the discount window as the instrument of a LLR policy. My main focus has been on general equilibrium rationalizations of the policy using explicit, formal economic models. While this covers an important part of the existing literature, it is by no means comprehensive – I covered only papers where the discount window was explicitly discussed. In general, the discussion in this theoretical literature is held at a relatively abstract level, relying on simplified formal descriptions of financial interactions, without capturing the peculiarities so often present in practice.

There is a parallel literature discussing more practical considerations related to discount window policy without resorting to formal economic models for framing the main arguments (see, for example, Carlson, Duygan-Bump, and Nelson [2015] and the classic "little" book by Friedman [1960]). The relevance of this more applied literature can be easily recognized. I contend that, even for the practitioner, there are valuable insights emerging from the more theoretical literature described in this article. It seems likely that familiarity with this theoretical literature is also much less common in policy circles. By minimizing the focus on technical issues, one objective of this essay was to try

and bring down barriers between the practice and the theory behind central-bank liquidity provision.

Formal arguments, if well-structured, are either complete explanations of ideas or explicit about the areas of incompleteness (that is, where ad-hoc assumptions are being employed for lack of a good explanation or just as a shortcut). By reviewing the formal models available in the literature, one is able to get a better sense of the issues that are well-understood and the issues that are still largely unexplained or plainly unexplored.

To close the article, let me provide a brief summary of the main ideas addressed in the existing models. At their core, the models need to formalize a concept of liquidity. Different models do this in different ways. In many cases, agents (households and/or firms) in the economy confront an urgent need to access extra resources. The Diamond and Dybvig model is, of course, a canonical example of this approach. Other examples include the case when firms need interim extra funding to continue running their project, or when a subset of agents in the economy are moving to a different location and only some commodities (assets or goods) are transportable.

Idiosyncratic shocks across agents often motivate the formation of banks that act as coalitions to pool the risk associated with those shocks. In other models, banks are useful just because they have a technological advantage (in monitory loans, for example) relative to individual investors. Some models have no explicit institution resembling a bank – individual agents directly interact with the discount window.

In general, to have the discount window playing a valuable role in an economy with banks, it must be the case that those banks are organized in a way that keeps them exposed to residual uncertainty, even after pooling individual agents' exposures, and, furthermore, that there are barriers impeding the reallocation of resources through markets. In many cases, the limits to market functioning originate in segmentation and the resulting impossibility for certain agents to engage in potentially beneficial trade. In other cases, private information or limited commitment undermines some agents' capacity to trade. Legal and institutional constraints also play a role in some of the models.

The combination of bank-level liquidity shocks and market frictions creates liquidity shortfalls that result in a misallocation of resources. Sometimes the misallocation has to do with uneven consumption across households, and in other cases it is caused by the early liquidation of productive investment. Another source of misallocation is the possibility of having positive net present value projects that go unexploited.

It is interesting to highlight that, in some of the models, liquidity rationing results from the fact that the price of the liquid assets is

pinned down by a different set of factors than those associated with liquidity demand. So, for example, if the price of liquid assets is given by the future discounted value of its associated cash flows, but those liquid assets are needed in an interim period for liquidity purposes, then rationing and "scarcity" may happen. A similar situation arises when money simultaneously plays the role of the available liquid asset, aside from its usual role as a store of value. Holmstrom and Tirole's (1998) and Freeman's (1996) models, respectively, are good illustrations of this general but rather subtle idea.

Market frictions are often not enough to make discount window interventions beneficial – in many of the models, the government also has an advantage over private agents in its ability to overcome physical impediments to trade (spatial or otherwise) or in its ability to tax individuals in the future. Indeed, in some of the models, the government (via the discount window or otherwise) can generate the needed extra liquidity by issuing claims on future taxes and committing to fulfill them in the future. In other cases, the discount window is simply assumed to be better able (than private agents) to redistribute liquidity across agents at a given point in time due to, for example, its relative ubiquity.

To counterbalance the advantage attributed in the models to the discount window as a channel to allocate liquidity, some of the models contemplate the threat of moral hazard that comes with interventions and the provision of liquidity insurance by the central bank. While the moral hazard implications of central-bank lending are well-recognized in policy circles, the subject is (perhaps surprisingly) not very thoroughly studied in the more formal and technical literature reviewed here.

As should be clear even from this brief closing summary of the main ideas, there are a lot of elements that need to be present to create the conditions for the discount window to be a valuable institution in an economy. For this reason, in general, the models so far developed are relatively abstract and, at the same time, complex. Despite that, my contention was that many practical insights can come out from a detailed study of those models. This review hopefully serves as a concise introduction and potentially a useful guide to those interested in pursuing such an undertaking.

REFERENCES

Acharya, Viral V., Denis Gromb, and Tanju Yorulmazer. 2012. "Imperfect Competition in the Interbank Market for Liquidity as a Rationale for Central Banking." *American Economic Journal: Macroeconomics* 4 (April): 184–217.

Acharya, Viral V., and Tanju Yorulmazer. 2008. "Cash-in-the-Market Pricing and Optimal Resolution of Bank Failures." *Review of Financial Studies* 21 (November): 2705–42.

Afonso, Gara, and Ricardo Lagos. 2015. "Trade Dynamics in the Market for Federal Funds." *Econometrica* 83 (January): 263–313.

Allen, Franklin, Elena Carletti, and Douglas Gale. 2009. "Interbank Market Liquidity and Central Bank Intervention." *Journal of Monetary Economics* 56 (July): 639–52.

Allen, Franklin, and Douglas Gale. 2000. "Financial Contagion." *Journal of Political Economy* 108 (February): 1–33.

Antinolfi, Gaetano, Elisabeth Huybens, and Todd Keister. 2001. "Monetary Stability and Liquidity Crises: The Role of the Lender of Last Resort." *Journal of Economic Theory* 99 (July): 187–219.

Antinolfi, Gaetano, and Todd Keister. 2006. "Discount Window Policy, Banking Crises, and Indeterminacy of Equilibrium." *Macroeconomic Dynamics* 10 (February): 1–19.

Ashcraft, Adam B., and Darrell Duffie. 2007. "Systemic Illiquidity in the Federal Funds Market." *American Economic Review Papers and Proceedings* 97 (May): 221–5.

Bech, Morten L., and Elizabeth Klee. 2011. "The Mechanics of a Graceful Exit: Interest on Reserves and Segmentation in the Federal Funds Market." *Journal of Monetary Economics* 58 (July): 415–31.

Bhattacharya, Sudipto, and Douglas Gale. 1987. "Preference Shocks, Liquidity and Central Bank Policy." In *New Approaches to Monetary Economics*, edited by W.A. Barnett and K. J. Singleton. Cambridge: Cambridge University Press, 69–88.

Board of Governors of the Federal Reserve System. 1968. "Reappraisal of the Federal Reserve Discount Mechanism: Report of a System Committee." Steering Committee for the Fundamental Reappraisal of the Discount Mechanism (July 15).

Caballero, Ricardo J., and Arvind Krishnamurthy. 2008. "Collective Risk Management in a Flight to Quality Episode." *Journal of Finance* 63 (October): 2195–2230.

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

Carlson, Mark A., Burcu Duygan-Bump, and William R. Nelson. 2015. "Why Do We Need Both Liquidity Regulations and a Lender of Last Resort? A Perspective from Federal Reserve Lending During the 2007-09 US Financial Crisis." Bank for International Settlements Working Paper 493 (March).

Castiglionesi, Fabio, and Wolf Wagner. 2012. "Turning Bagehot on his Head: Lending at Penalty Rates when Banks Can Become Insolvent." *Journal of Money, Credit and Banking* 44 (February): 201–19.

Champ, Bruce, Bruce D. Smith, and Stephen D. Williamson. 1996. "Currency Elasticity and Banking Panics: Theory and Evidence." *Canadian Journal of Economics* 29 (November): 828–64.

Courtois, Renee, and Huberto M. Ennis. 2010. "Is There Stigma Associated with Discount Window Borrowing?" Federal Reserve Bank of Richmond *Economic Brief* 10-05.

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

Diamond, Douglas W., and Raghuram G. Rajan. 2005. "Liquidity Shortages and Banking Crises." *Journal of Finance* 60 (April): 615–47.

Duffie, Darrell. 2010. "Presidential Address: Asset Price Dynamics with Slow-Moving Capital." *Journal of Finance* 65 (August): 1237–67.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen. 2005. "Over-the-Counter Markets." *Econometrica* 73 (November): 1815–47.

Ennis, Huberto M. 2017. "Interventions in Markets with Adverse Selection: Implications for Discount Window Stigma." Federal Reserve Bank of Richmond Working Paper 17-01 (January).

Ennis, Huberto M., and Todd Keister. 2008. "Understanding Monetary Policy Implementation." Federal Reserve Bank of Richmond *Economic Quarterly* 94 (Summer): 235–63.

Ennis, Huberto M., and Todd Keister. 2010. "On the Fundamental Reasons for Bank Fragility." Federal Reserve Bank of Richmond *Economic Quarterly* 96 (First Quarter): 33–58.

Ennis, Huberto M., and David A. Price. 2015. "Discount Window Lending: Policy Trade-offs and the 1985 BoNY Computer Failure." Federal Reserve Bank of Richmond *Economic Brief* 15-05.

Ennis, Huberto M., and John A. Weinberg. 2013. "Over-the-Counter Loans, Adverse Selection, and Stigma in the Interbank Market." *Review of Economic Dynamics* 16 (October): 601–16.

Ennis, Huberto M., and John A. Weinberg. 2016. "The Role of Central Bank Lending in the Conduct of Monetary Policy." Federal Reserve Bank of Richmond *Economic Brief* 16-12.

Flannery, Mark J. 1996. "Financial Crises, Payment System Problems, and Discount Window Lending." *Journal of Money, Credit and Banking* 28 (November, part 2): 804–24.

Freeman, Scott. 1996. "The Payments System, Liquidity, and Rediscounting." *American Economic Review* 86 (December): 1126–38.

Freeman, Scott. 1999. "Rediscounting Under Aggregate Risk." *Journal of Monetary Economics* 43 (February): 197–216.

Freixas, Xavier, Bruno M. Parigi, and Jean-Charles Rochet. 2000. "Systemic Risk, Interbank Relations, and Liquidity Provision by the Central Bank." *Journal of Money, Credit and Banking* 32 (August, part 2): 611–38.

Freixas, Xavier, Bruno M. Parigi, Jean-Charles Rochet. 2004. "The Lender of Last Resort: A Twenty-First Century Approach." *Journal of the European Economic Association* 2 (December): 1085–1115.

Friedman, Milton. 1960. *A Program for Monetary Stability.* New York: Fordham University Press.

Gertler, Mark, and Nobuhiro Kiyotaki. 2010. "Financial Intermediation and Credit Policy in Business Cycle Analysis." In *Handbook of Monetary Economics 3A,* edited by B.M. Friedman and M. Woodford. Amsterdam: Elsevier, 547–99.

Goodfriend, Marvin, and Robert G. King. 1988. "Financial Deregulation, Monetary Policy, and Central Banking." Federal Reserve Bank of Richmond *Economic Review* 74 (May/June): 3–33.

Goodfriend, Marvin, and Jeffrey M. Lacker. 1999. "Limited Commitment and Central Bank Lending." Federal Reserve Bank of Richmond *Economic Quarterly* 85 (Fall): 1–27.

Gorton, Gary, and Guillermo Ordoñez. 2016. "Fighting Crises." National Bureau of Economic Research Working Paper 22787 (October).

Hart, Oliver, and John Moore. 1994. "A Theory of Debt Based on the Inalienability of Human Capital." *Quarterly Journal of Economics* 109 (November): 841–79.

Holmstrom, Bengt R., and Jean Tirole. 1996. "Modeling Aggregate Liquidity." *American Economic Review Papers and Proceedings* 86 (May): 187–91.

Holmstrom, Bengt R., and Jean Tirole. 1998. "Private and Public Supply of Liquidity." *Journal of Political Economy* 106 (February): 1–40.

Kahn, Charles, and Joao Santos. 2005. "Allocating Bank Regulatory Powers: Lender of Last Resort, Deposit Insurance and Supervision." *European Economic Review* 49 (November): 2107–36.

Keister, Todd. 2016. "Bailouts and Financial Fragility." *Review of Economic Studies* 83 (April): 704–36.

Kiyotaki, Nobuhiro, and John Moore. 1997. "Credit Cycles." *Journal of Political Economy* 105 (April): 211–48.

Lacker, Jeffrey M. 2004. "Payment System Disruptions and the Federal Reserve Following September 11, 2001." *Journal of Monetary Economics* 51 (July): 935–65.

Lagos, Ricardo, Guillaume Rocheteau, and Pierre-Olivier Weill. 2011. "Crises and Liquidity in Over-the-Counter Markets." *Journal of Economic Theory* 146 (November): 2169–205.

Martin, Antoine. 2006. "Liquidity Provision vs. Deposit Insurance: Preventing Bank Panics without Moral Hazard." *Economic Theory* 28 (May): 197–211.

Martin, Antoine. 2009. "Reconciling Bagehot and the Fed's Response to September 11." *Journal of Money, Credit and Banking* 41 (March/April): 397–415.

Mills, David C. Jr. 2006. "Alternative Central Bank Credit Policies for Liquidity Provision in a Model of Payments." *Journal of Monetary Economics* 53 (October): 1593–1611.

Morris, Stephen, and Hyun Song Shin. 2003. "Global Games: Theory and Applications." In *Advances in Economics and Econometrics. Proceedings of the Eighth World Congress of the Econometric Society*, edited by M. Dewatripont, L. Hansen, and S. Turnovsky. Cambridge: Cambridge University Press, 56–114.

Myers, Stewart C., and Nicholas S. Majluf. 1984. "Corporate Financing and Investment Decisions When Firms Have Information that Investors Do Not Have." *Journal of Financial Economics* 13 (June): 187–221.

Philippon, Thomas, and Vasiliki Skreta. 2012. "Optimal Interventions in Markets with Adverse Selection." *American Economic Review* 102 (February): 1–28.

Ratnovski, Lev. 2009. "Bank Liquidity Regulation and the Lender of Last Resort." *Journal of Financial Intermediation* 18 (October): 541–58.

Repullo, Rafael. 2000. "Who Should Act as Lender of Last Resort? An Incomplete Contracts Model." *Journal of Money, Credit and Banking* 32 (August): 580–605.

Rochet, Jean-Charles, and Xavier Vives. 2004. "Coordination Failures and the Lender of Last Resort: Was Bagehot Right After All?" *Journal of the European Economic Association* 2 (December): 1116–47.

Smith, Bruce D. 2002. "Monetary Policy, Banking Crises, and the Friedman Rule." *American Economic Review Papers and Proceedings* 92 (May): 128–34.

Williamson, Stephen D. 1998. "Discount Window Lending and Deposit Insurance." *Review of Economic Dynamics* 1 (January): 246–75.

Economic Quarterly—Volume 102, Number 1—First Quarter 2016—Pages 51–78

Consumer Payment Choice in the Fifth District: Learning from a Retail Chain

Zhu Wang and Alexander L. Wolman

The U.S. payments system has undergone fundamental changes over the past several decades. Perhaps the most significant trend is the shift from paper payment instruments, namely cash and check, to electronic ones such as credit and debit cards. Understanding this shift is important, as it affects billions of transactions worth trillions of dollars each year.[1] For many years, experts on payments systems have forecast the arrival of a completely electronic, paperless payments system, but it has not yet happened. Cash and check still play a large role in the economy, particularly in some sectors.

In this context, a sizable body of empirical literature has developed to study consumer payment choice. Most of the studies rely on data from consumer surveys.[2] While this research has improved our understanding of how consumers choose to pay, consumer survey data have their limitations, including small sample size and imperfect reporting.

Our paper reports and analyzes new evidence on consumer payment choice in retail transactions, including the use of cash, credit card, debit card, and check, based on a comprehensive dataset comprising

We thank Joseph Johnson for excellent research assistance. The views expressed in this article are those of the authors and do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System. All errors are our own.

DOI: http://doi.org/10.21144/eq1020102

[1] According to the latest Federal Reserve Payments Study (2014), the estimated number of noncash payments alone, excluding wire transfers, was 122.8 billion in 2012, with a value of $79.0 trillion.

[2] For example, Borzekowski et al. (2008), Borzekowski and Kiser (2008), Zinman (2009), Ching and Hayashi (2010), Arango et al. (2011), Cohen and Rysman (2012), Schuh and Stavins (2012), and Koulayev et al. (2016).

merchant transaction records. The data, provided by a large retail chain, cover every transaction in each of its stores over the five-year period from April 2010 through March 2015. We focus on hundreds of stores located across the Fifth Federal Reserve District. The purpose of our study is to provide a better understanding of payment variation for retail transactions in this region.[3]

Our study has several important findings. First, the fraction of cash transactions decreases in transaction size and is affected by location-specific variables that reflect consumers' preferences and the opportunity costs of using cash relative to noncash means of payments.

Second, based on the estimation results, we evaluate the relative importance of different groups of variables in explaining the payment variation across locations in our sample. We find that median transaction size, demographics, education levels, and state fixed effects are the top factors related to consumer payment choice. Taking these into consideration, we project the payment variation across the entire Fifth District for retail outlets similar to those in our sample.

Finally, we identify interesting time patterns of payment variation. In particular, the shares of cash and check transactions decline steadily over our five-year sample period, while debit and credit's shares rise. The overall cash fraction of transactions is estimated to have declined by 2.46 percentage points per year, largely replaced by debit. We show that the decline in cash at this particular retailer was likely not driven by transitory factors, and only a relatively small fraction could be explained by changes in median transaction size and zip-code-level variables. This leaves a large fraction of the time trend to be explained, with prime candidates being technological progress in debit and changing consumer perceptions of debit relative to cash.

The structure of the paper is as follows. Section 1 describes the data used in our analysis as well as the empirical approach. Section 2 introduces the regression model and presents an overview of the estimation results. Section 3 evaluates the relative importance of different variables in explaining payment variation across locations in our sample, and projects payment variation across the entire Fifth District. Section 4 discusses the longer-run decline of cash. Finally, Section 5 concludes.

---

[3] See Wang and Wolman (2016) for a study covering the entire chain's thousands of stores across the country between April 2010 and March 2013. That study mainly explores payment variation across transaction sizes and time frequencies. In contrast, this paper focuses on decomposing the relative importance of different local variables and projecting cross-sectional payment patterns in the Fifth District.

1. DATA AND EMPIRICAL APPROACH

The transactions data used in our study is from a large discount retailer with hundreds of stores across the Fifth Federal Reserve District, which covers Maryland, North Carolina, South Carolina, Virginia, Washington, DC, and West Virginia. The stores sell a wide variety of goods in various price ranges, with household consumables such as food and health and beauty aids accounting for a majority of sales. The unit of observation is a transaction, and the time period is April 1, 2010, through March 30, 2015. For each transaction, the data include means of payment, time, location, and amount. We include only transactions that consist of a sale of goods, with one payment type used, where the payment type is cash, credit card, debit card, or check – the four general-purpose means of payment.[4] The retailer also provides cash-back services, and the purchase components of cash-back transactions are included in our analysis. In contrast, transactions made with special-purpose means of payment such as electronic benefit transfer (EBT), coupons, and store return cards are excluded. All told, our analysis covers 86 percent of the total transactions in the sample period. Our summary of the data in this section will refer to all stores located in the Fifth District; the zip-code-level data introduced below and used in the empirical analysis covers most of those stores' zip codes, but we will need to omit a small fraction of retail outlets from that analysis because the zip-code-level data are unavailable.[5]

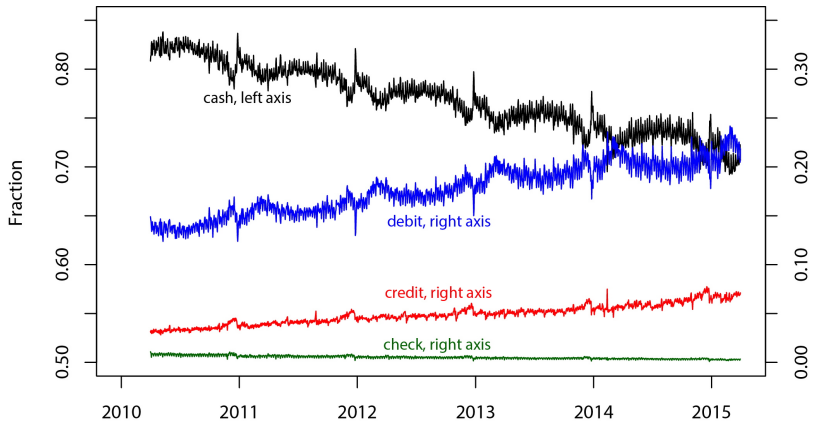
Payment Variation

The purpose of our paper is to explain payment variation across locations and time in the Fifth District. Figure 1 presents payment variation across time in our sample. The data are plotted at the daily level, displaying the fraction of all the transactions accounted for by each payment type. Note that while cash is measured on the left axis, and debit, credit, and check are all measured on the right axis, both axes vary by 0.35 from bottom to top, so fluctuations for each payment type are displayed comparably. The figure shows that cash is the dominant payment instrument at this retailer, followed by debit, credit,

[4] Data limitations prevent us from distinguishing credit cards from signature debit and prepaid cards. However, our estimates reveal variation in what we report as "credit cards" that is significantly different from the variation in PIN debit. Because signature debit and prepaid cards are close substitutes for PIN debit, in that they rely on consumers' account balances rather than borrowed funds, we can reasonably assume the estimated patterns are primarily driven by the true credit cards.

[5] We omit Washington, DC, from the regression analysis due to lack of zip-code-level crime data.

Figure 1 Payment Variation Across Time

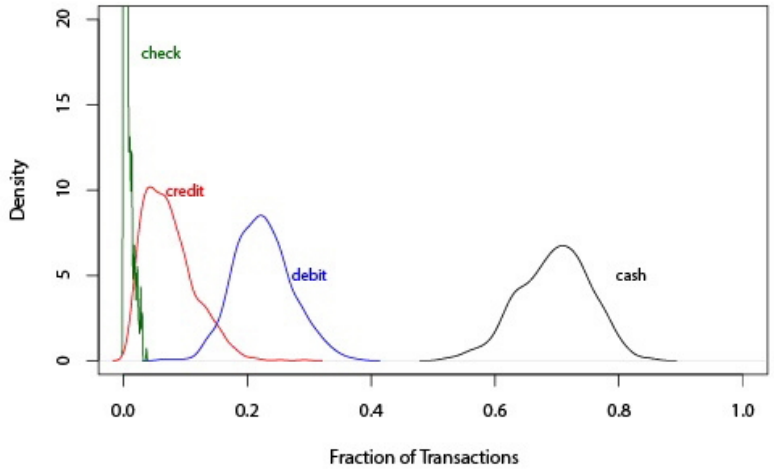


and check. Over the sample period, the fractions of cash and check are trending down, with debit and credit trending up. There are higher frequency patterns as well, with cash and debit again moving in opposite directions. We will allow for these patterns in the econometric model by including day-of-week, day-of-month, and month-of-sample dummies.

Figure 2 presents payment variation across locations, restricting attention to the last full month of the sample, March 2015. We aggregate the data by zip code and display smoothed estimates of the density functions for the fraction of transactions conducted with cash, debit, credit, and check.[6] We use only one month because of the time trend evident in Figure 1. The ranking from Figure 1 is also apparent in Figure 2: cash is the dominant form of payment, followed by debit, credit, and check. Moreover, Figure 2 shows significant variation across zip-code locations in cash, debit, and credit use. This variation highlights the need for including location-specific variables in our econometric model.

[6] Note that the estimated kernel density for checks is truncated in Figure 2. The check fractions are concentrated near zero, so the figure would be uninformative about the other payment instruments if we extended the y-scale to include the entire check density.

Figure 2 Payment Variation Across Locations



Explanatory Variables

The payment variation identified in Figures 1 and 2 suggests the quantitative importance of including location- and time-specific variables in an econometric model of payment choice. In Wang and Wolman (2016), we discuss how theories of money demand and payment choice motivate the choice of particular variables. Here, we simply list the variables we use and explain informally why they may be associated with variation in the shares of different payment types across locations and time. Note that our data identify transactions but not customers, so we treat the characteristics of the zip code in which each store is located as representative of the characteristics of the store's customers and the economic environment in which they live. Table 1 lists summary statistics for the zip-code-level explanatory variables used in the regressions, fixed at their 2011 values.[7]

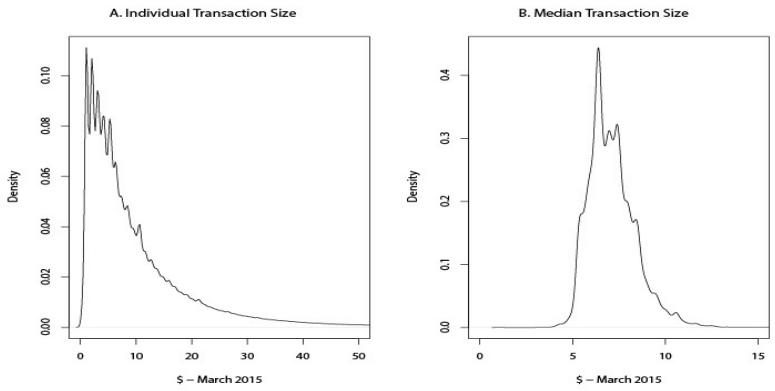
[7] Most of our zip-code-level variables come from the U.S. Census' American Community Survey (ACS) and the FDIC's Summary of Deposits. The robbery data are from the FBI's Uniform Crime Report. We fix zip-code-level explanatory variables at their 2011 values (five-year estimates), because the ACS provides only five-year estimates for areas with fewer than 20,000 residents. In Section 4, where we study longer-run payment variation, we will discuss the effects of time variation in zip-code-level explanatory variables.

Table 1 Summary Statistics of Zip-Code Variables

Variable (unit)		Mean	Std. dev.	1%	99%
Banking condition					
HHI metro		0.211	0.172	0.070	0.735
HHI rural		0.273	0.110	0.125	0.561
Branches per capita ($1/10^3$)		0.44	0.44	0.06	2.59
Socioeconomic condition					
Robbery rate ($1/10^5$)		29.76	40.31	0.00	235.07
Median household income ($)		41015	12666	22214	90078
Population density (per mile^2)		579	1024	20	5017
Family households (%)		67.06	5.84	43.44	79.05
Housing (%): Renter occupied		27.73	9.25	10.49	54.56
Owner occupied		59.08	9.85	28.05	80.18
Vacant		13.20	7.80	4.38	43.89
Demographics (%)					
Female		51.08	2.37	40.35	55.01
Age	<15	18.74	2.79	11.20	24.80
	15-34	25.27	5.06	16.25	46.17
	35-54	27.08	2.45	19.43	32.43
	55-69	18.69	3.49	10.65	29.08
	>70	10.23	3.04	4.02	19.12
Race	White	68.23	21.99	8.66	98.86
	Black	24.22	20.16	0.25	82.09
	Hispanic	6.20	6.34	0.36	32.16
	Native	1.11	5.24	0.05	26.93
	Asian	1.24	1.88	0.04	8.17
	Pacific Islander	0.05	0.07	0.00	0.35
	Other	3.22	3.93	0.04	19.38
	Multiple	1.93	0.94	0.42	5.32
Education level (%)					
Below high school		19.26	6.90	5.70	36.90
High school		33.85	7.11	15.90	51.70
Some college		20.28	3.88	11.50	29.20
College		26.61	9.97	10.70	56.10

(I) Median Transaction Size We use the median transaction size for each zip-code day to capture the transaction size distribution. The theory outlined in Wang and Wolman (2016) suggests that higher transaction sizes will be associated, all else equal, with less cash use. Figure 3 provides information about the size distribution of transactions in March 2015 without regard to means of payment. Figure 3A displays a smoothed density function, by transaction size, for all transactions in the month. Figure 3B plots the distribution of median transaction sizes across zip-code days. Figure 3B complements Figure 2 in showing that there is substantial heterogeneity across locations with respect to size of transaction, as well as payment mix.

Figure 3 Kernel Densities of Transaction Size in March 2015



(II) Banking Variables Local banking condition matters for payment choice, but the effects are subtle. Cash use may be expected to decrease in banking-sector competition (which results in lower banking fees and/or better deposit terms that increase consumers' opportunity costs of using cash) but increase in bank branches per capita (which reduces consumers' costs of replenishing cash balances). Following the banking literature and antitrust tradition, we measure banking-sector concentration by the Herfindahl-Hirschman Index (HHI) in each Metropolitan Statistical Area (MSA) or rural county.[8] Bank branches per capita are measured at the zip-code level.

(III) Socioeconomic Variables We include the robbery rate, median household income, population density, fraction of family households, and fraction of homeownership as socioeconomic variables. The robbery rate is measured at the county level while other variables are measured at the zip-code level.

A higher robbery rate increases the cost of holding cash, which we would expect to reduce cash use. The other variables are likely to correlate with consumers' access to bank accounts or ownership of credit or debit cards. Note that population density is relevant for adoption

[8] Both the theoretical literature and antitrust practice typically assume that the relevant geographic banking market is a local area where banks compete to offer financial services to households and small businesses. That market area is often approximated by an MSA in urban areas and by a county in rural areas. The most commonly used measure of market concentration is the HHI, calculated by squaring each bank's share of deposits in a market and then summing these squared shares.

because, as McAndrews and Wang (2012) point out, replacing traditional paper payments with electronic payments requires merchants and consumers to each pay a fixed cost but reduces marginal costs for doing transactions. Their work suggests adoption and usage of electronic payment instruments should be higher in areas with a high population density or more business activity.

(IV) Demographic Variables Gender, age/cohort group, and race are included to reflect the fact that payment behavior may vary systematically with demographic characteristics. These variables are each measured as a fraction of the population at the zip-code level.

(V) Education Variables We specify four education levels: below high school, high school, some college, and college and above. Higher education is often associated with better financial literacy and higher opportunity time cost of using cash, so it may be associated with a higher adoption and usage of noncash payments. The education variables are each measured as a fraction of the population at the zip-code level.

(VI) and (VII) State and Time Dummies We also include state dummies as well as day-of-week, day-of-month, and month-of-sample dummies.

2. ESTIMATION RESULTS

We turn now to an empirical model aimed at explaining the variation in payment shares through the behavior of the explanatory variables. The data are analyzed using the fractional multinomial logit model (FMLogit).[9] The dependent variables are the fractions of each of the four payment instruments used in transactions at stores in one zip code on one day between April 1, 2010, and March 31, 2015.[10] The explanatory variables are those introduced above.[11]

[9] The FMLogit model addresses the multiple fractional nature of the dependent variables, namely that the fraction of payments for each instrument should remain between zero and one, and the fractions add up to one. More details of the FMLogit model are provided in the Appendix.

[10] In our sample, most zip codes have only one store. Because we measure the fraction of payment instruments at the zip-code level, we do not distinguish locations with one store from those with multiple stores. In the latter case, we simply sum up the transactions of all the stores in the zip code.

[11] Note that the local characteristics data are from a single year, 2011, while the dependent variables and the median-transaction-size variable come from multiple years,

Table 2 reports the estimation results, expressed in terms of marginal effects.[12] We summarize the findings as follows.

(I) Median Transaction Size Aggregating transactions within a zip-code day, we expect to find that a rightward shift in the size distribution of transactions corresponds to a lower share of cash transactions, as consumers are less likely to use cash for larger transactions. Using median transaction size as a convenient summary of the size distribution, we find the expected result: evaluating at the mean of median transaction size, $6.65, the marginal effects indicate that a $1 increase in median transaction size reduces the predicted cash share by 1.8 percentage points but raises debit by 1.3 percentage points, credit by 0.4 percentage points, and check by 0.1 percentage points.

(II) Banking Variables We find that higher banking concentration corresponds to a higher cash share (lower card shares) in rural areas. However, higher concentration corresponds to a lower cash share (higher card shares) in MSAs. We conjecture that in rural areas HHI does a good job proxying for banks' market power, whereas in metro areas it may not: in metro areas, banking is inherently competitive, and a high level of concentration (as measured by HHI) may simply indicate the presence of one or more especially efficient banks.[13] In contrast, more bank branches per capita are associated with a higher cash share, mainly at the expense of debit and credit. These findings are consistent with our discussion in Section 1.

(III) Socioeconomic Variables As expected, a higher robbery rate is found to be associated with less cash use and more debit use. Our estimates show that a one-standard-deviation increase in the robbery rate (i.e., four more robbery incidences per 10,000 residents) reduces

---

2010-15. For robustness checks, we also ran regressions only on 2011 data as well as on data from other sample years. The results are largely consistent.

[12] For continuous variables, the marginal effects are calculated at the means of the independent variables. For dummy variables, the marginal effects are calculated by changing the dummy from zero to one, holding the other variables fixed at their means.

[13] When interpreting the relationship between market performance and HHI, two hypotheses are often tested. One is the Structure-Conduct-Performance (SCP) hypothesis, which assumes that the ability of banks in a local market to set relatively low deposit rates or high fees depends positively on market concentration. The other is the Efficient-Structure (ES) hypothesis, which takes an opposite view and argues that a concentrated market may reflect the efficiency advantages of leading banks in the market, so it may instead be associated with lower prices for banking services. The empirical evidence on these two hypotheses is mixed (Gilbert and Zaretsky [2003] provides a comprehensive literature review). Our findings suggest that both hypotheses are relevant for our sample, with the SCP hypothesis supported by the rural market evidence and the ES hypothesis supported by the MSA evidence.

Table 2 Marginal Effects for Zip-Code Variables

Variable		Cash	Debit	Credit	Check
Median transaction size		-0.018*	0.013*	0.004*	0.001*
Banking condition					
HHI		0.035*	-0.027*	-0.010*	0.002*
HHI*metro		-0.051*	0.042*	0.011*	-0.003*
Branches per capita		0.069*	-0.038*	-0.029*	-0.002*
Socioeconomic condition					
Robbery rate		-0.126*	0.121*	0.020*	-0.014*
Median household income		0.003*	-0.013*	0.019*	-0.009*
Population density		-0.450*	0.470*	0.077*	-0.097*
Family households		-0.089*	0.104*	-0.006*	-0.009*
Housing: Owner occupied		-0.006*	-0.030*	0.025*	0.011*
Vacant		-0.021*	-0.029*	0.043*	0.006*
Demographics					
Female		-0.043*	0.131*	-0.079*	-0.010*
Age	15-34	-0.272*	0.285*	0.000	-0.013*
	35-54	-0.366*	0.416*	-0.033*	-0.016*
	55-69	0.070*	-0.037*	-0.011*	-0.022*
	>70	-0.172*	0.161*	0.008*	0.004*
Race	Black	0.055*	-0.040*	-0.007*	-0.007*
	Hispanic	0.049*	-0.168*	0.114*	0.005*
	Native	0.105*	-0.060*	-0.040*	-0.004*
	Asian	0.037*	-0.018*	-0.018*	-0.001
	Pacific Islander	0.986*	0.811*	-1.595*	-0.202*
	Other	0.129*	0.111*	-0.220*	-0.019*
	Multiple	-0.019	-0.136*	0.251*	-0.096*
Education level					
High school		-0.280*	0.169*	0.108*	0.003*
Some college		-0.275*	0.184*	0.089*	0.002*
College		-0.271*	0.162*	0.106*	0.003*
Pseudo R^2		0.604	0.534	0.607	0.559
Zip-code-day Observations		1,021,764	1,021,764	1,021,764	1,021,764

Note: *1 percent significance level based on robust standard errors. The dependent variables are the fractions of each of the four general payment instruments used in transactions at stores in a zip code on a day between April 1, 2010, and March 31, 2015. The explanatory variables take their values in 2011. Banking HHI index is calculated by squaring each bank's share of deposits in a market (an MSA or a rural county) and then summing these squared shares. Metro is a dummy variable taking the value 1 when the banking market is an MSA, otherwise equal to zero. Branches per capita is measured as the number of bank branches per 100 residents in a zip code. Robbery rate is defined as the number of robberies per 100 residents in a county. Median household income is measured in units of $100,000 per household in a zip code. Population density is measured in units of 100,000 residents per square mile in a zip code. All the other variables are expressed as fractions.

the predicted cash share by 0.5 percentage points but raises debit by 0.49 percentage points.

High median household income in a zip code is associated with high credit use, mainly at the expense of debit. We find that for a one-standard-deviation increase ($12,666) in the median household income from its mean, the predicted credit share increases by 0.24 percentage points, but the debit and check shares drop by 0.16 and 0.11 percentage points, respectively. The effect on the cash share is small – it rises by 0.04 percentage points. The results suggest that median household income in our sample may largely proxy for access to credit.

We find that higher population density is associated with lower shares of paper payments and higher shares of card payments. This is consistent with McAndrews and Wang's (2012) theory of the scale economies of adopting relatively new payment instruments. A one-standard-deviation increase in population density (1,024 residents per square mile) reduces the predicted cash share by 0.46 percentage points and check by 0.10 percentage points, but it raises debit by 0.48 percentage points and credit by 0.08 percentage points. Although the stores in our sample accept both credit and debit cards, consumers' adoption decisions should be related to the policies of other stores, and those may vary systematically with population density.

(IV) Demographic Variables Consistent with some existing payments studies (e.g., Klee [2008]), we find that demographic characteristics such as gender, age, and race are systematically related to consumer payment choice.

We find that a higher female ratio is associated with less cash use and more debit use. A higher presence of older age groups is associated with greater use of debit but less use of cash and check relative to the baseline age group, under 15. This might be because minors do not have access to noncash payments or because families with children tend to use more cash and check. However, the age profile with respect to cash is nonmonotonic. A higher presence of the age group 55-69 is associated with a significantly higher cash fraction. These findings suggest that the age variables may capture a combination of age and cohort effects. We also find that compared to white, minority groups tend to be associated with higher cash shares but lower debit shares.

(V) Education Variables We find a more educated population (i.e., high school and above) is associated with a lower cash fraction relative to the baseline education group (i.e., below high school). For education levels at high school and above, however, the difference is quite small between the sub-groups.

Table 3 State Fixed Effects

State Dummies	Cash	Debit	Credit	Check
North Carolina	-0.069*	0.095*	-0.025*	-0.001
South Carolina	-0.058*	0.086*	-0.027*	-0.000
Virginia	-0.063*	0.067*	-0.006*	0.002*
West Virginia	-0.033*	0.042*	-0.010*	0.001*

Note: *1 percent significance level based on robust standard errors.

(VI) State Dummies Our results reveal some interesting state fixed effects, as shown in Table 3. Compared with the benchmark state, Maryland, other states show lower shares of cash use and higher shares of debit use. This is particularly significant for North Carolina, South Carolina, and Virginia. They each have a cash share that is 5.8-6.9 percentage points lower than Maryland but a debit share that is 6.7-9.5 percentage points higher. West Virginia is the intermediate case, of which cash share is 3.3 percentage points below Maryland, and debit share is 4.2 percentage points higher. These states also show a lower share of credit use than Maryland but the magnitude is fairly small compared with debit, and the state fixed effects on check are quantitatively negligible.

(VII) Time Dummies Figure 4 plots the marginal effects associated with our estimated day-of-week dummies. The cash and debit effects are nearly mirror images of each other: cash falls and debit rises from Monday through Thursday, then cash rises and debit falls on Friday and Saturday, and the pattern reverses again on Sunday. Although credit displays less variation than cash or debit, there are noticeable movements in credit from Friday through Sunday.

Figure 5 plots the marginal effects associated with our day-of-month dummies. Whereas most of the "substitution" within the week occurred between cash and debit, within the month the substitution with cash comes from both credit and debit, especially credit. Early in the month, cash is at its highest and credit and debit are at their lowest. Over the month, cash generally falls and credit rises. Debit has a similar pattern to credit, although the variation is smaller.

Figure 4 Day-of-Week Marginal Effects

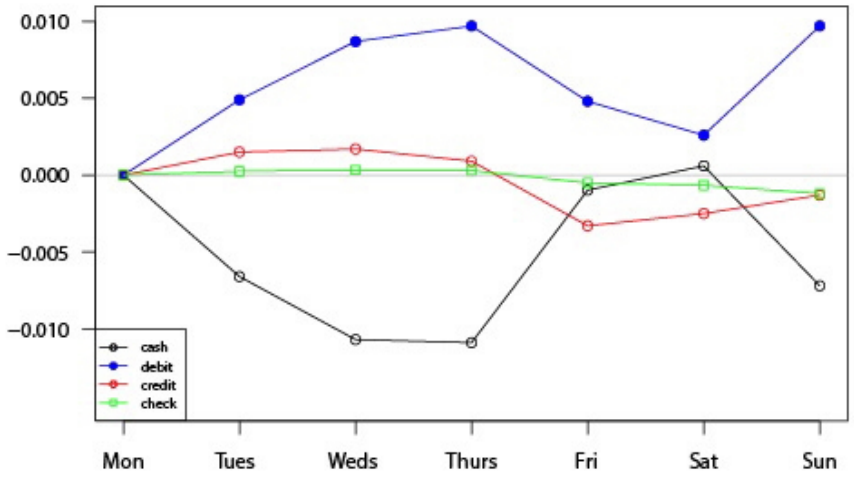
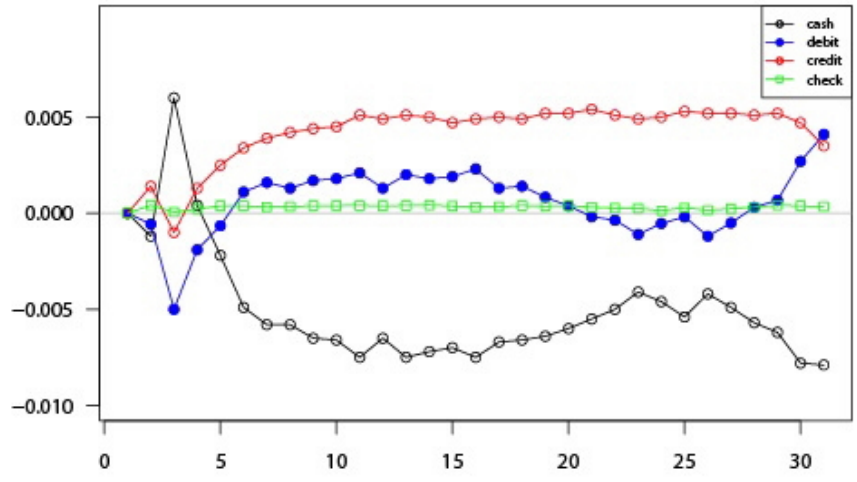
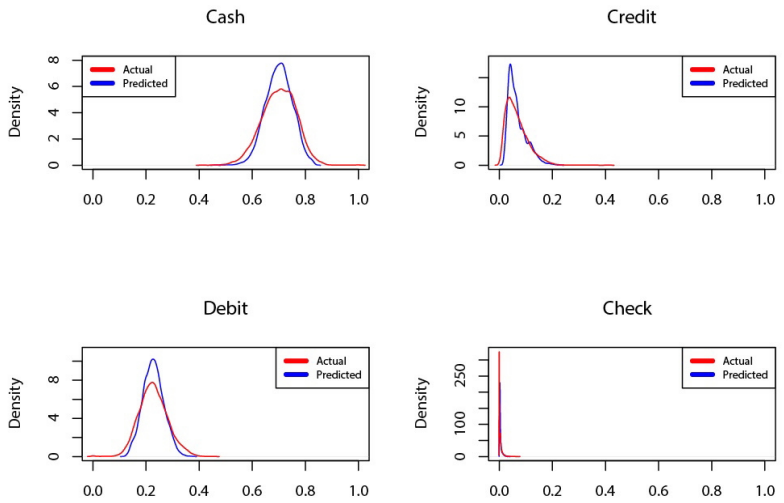


Figure 5 Day-of-Month Marginal Effects



A natural explanation for the day-of-week and day-of-month effects is consumers' changing financial or cash-holding positions during the period. Presumably, the weekly pattern could be driven by consumers

Figure 6 Fitted and Actual Payment Shares for March 2015



who receive weekly paychecks, while the monthly patterns are likely driven by those who receive monthly pay, including those who receive certain government benefits. One notable feature of the monthly pattern is a transitory reversal of the broad trends on the third day of the month. In fact, many recipients of Social Security and Supplemental Security Income are usually paid on the third of the month. Early in the month, these customers may be financially unconstrained, and thus spend cash, whereas late in the month they rely more on credit while anticipating the next paycheck. In Wang and Wolman (2016), we provide more extensive discussions of the weekly and monthly patterns.

The month-of-sample dummies in our regression identify the seasonal cycles and longer-run trends in the payment mix, but we will defer that discussion to Section 4.

3. PAYMENT VARIATION ACROSS LOCATIONS

Our regression analysis helps shed light on payment variation across locations. In this section, we will first evaluate the relative importance of the explanatory variables in accounting for such variation, and then project payment variation across the entire Fifth District for retail outlets similar to those in our sample.

Relative Importance of Explanatory Variables

In Figure 6, we plot the actual and model-predicted distributions of payment fractions for March 2015, a counterpart to Figure 2.[14] The figure shows that our regression model does a good job of capturing observed payment variation. With many explanatory variables included in the regression analysis, an immediate question is what factors account for most of the variation. To answer this question, we conduct the following decomposition exercise. We first calculate the pseudo-R^2 statistics, defined as the square of the correlation between the model-predicted value and the actual data, for the March 2015 sample. We then fix each subgroup of explanatory variables one by one at the sample mean values and recalculate the pseudo-R^2 statistics. The reduction of the model fit is then used as a measure of explanatory power of the controlled explanatory variables. Finally, we compare the relative importance across all the subgroups of explanatory variables.

Table 4 reports the comparison results for cash and debit, the two most used means of payment in our data. The table shows that the day-of-week and day-of-month dummies account for little of the data variation (1 to 2 percent), so the payment variation in the one-month data is mostly cross-location variation. For cash, it is median transaction size, education levels, demographics, and state fixed effects that rank as the top four factors in explaining the variation in cash fractions, each accounting for 44 percent, 19 percent, 17 percent, and 14 percent, respectively. These are also the top four factors that explain the variation in debit fractions, though the ranking is a little different, with state fixed effects ranking first (44 percent), followed by median transaction size (23 percent), demographics (14 percent), and education levels (9 percent).

The decomposition exercise above takes the median transaction size as given and shows that it explains a large share of payment variation across locations. However, it is possible that median transaction size is not independent of other location-specific variables. This will in turn affect the interpretation of the decomposition. To account for that, we conduct an alternative exercise. First, we regress median transaction size for each zip-code day on all the other explanatory variables using a linear model and calculate the model-predicted median transaction sizes and the residuals. Second, we re-run the FMLogit model as before but replace the median transaction sizes with the residual median

[14] Note that the data plots in Figure 2 and Figure 6 are slightly different because a small fraction of stores is omitted from the regression analysis due to missing zip-code-level information.

Table 4 Relative Importance of Explanatory Variables (March 2015)

	Cash			Debit		
Scenarios	R^2	ΔR^2	$\frac{\Delta R^2}{sum(\Delta R^2)}$	R^2	ΔR^2	$\frac{\Delta R^2}{sum(\Delta R^2)}$
All variables included	0.531			0.374		
Constant median transaction size	0.342	0.189	44%	0.264	0.110	23%
Constant banking condition	0.521	0.010	2%	0.366	0.008	2%
Constant socioeconomic factors	0.522	0.009	2%	0.347	0.027	6%
Constant demographics	0.456	0.075	17%	0.306	0.068	14%
Constant education levels	0.450	0.081	19%	0.330	0.044	9%
Constant state fixed effects	0.472	0.059	14%	0.163	0.211	44%
Constant day of week & month effects	0.525	0.006	1%	0.367	0.007	2%

transaction sizes. Finally, we redo the decomposition exercise above based on this new FMLogit regression.[15]

Table 6 in the Appendix reports the results of the new FMLogit regression. Note that the new FMLogit model contains the same information as the original one, so it yields the same marginal effects for median transaction size, as well as the same model fit in terms of the pseudo-R^2 values as in Table 2. The only difference is that the new model attributes some additional payment variation to the location-specific variables through their impact on median transaction size, which results in different estimated marginal effects for those variables. Comparing Tables 2 and 6 confirms this, but the qualitative results found in Table 2 remain largely unchanged.

Based on the alternative regression model, we redo the decomposition exercise and report the results in Table 5. For cash, median transaction size, education, demographics, and state fixed effects remain the top four factors driving cash fractions, though the ranking and relative shares differ slightly from Table 4: demographics now comes in first (36 percent) followed by median transaction size (23 percent), education levels (16 percent), and state fixed effects (13 percent). A similar case is found for debit.

[15] For the purpose of estimating the effects of the other explanatory variables, the alternative model where we use residual median transaction size instead of median transaction size is equivalent to running a regression without the median transaction size. Also, in principle, we could run the alternative model for each subgroup of variables other than median transaction size, but we chose not to do so. One consideration is that median transaction size is likely to be affected by other, more fundamental variables (such as income and race) but not the other way around.

Table 5 Relative Importance of Explanatory Variables (An Alternative Model)

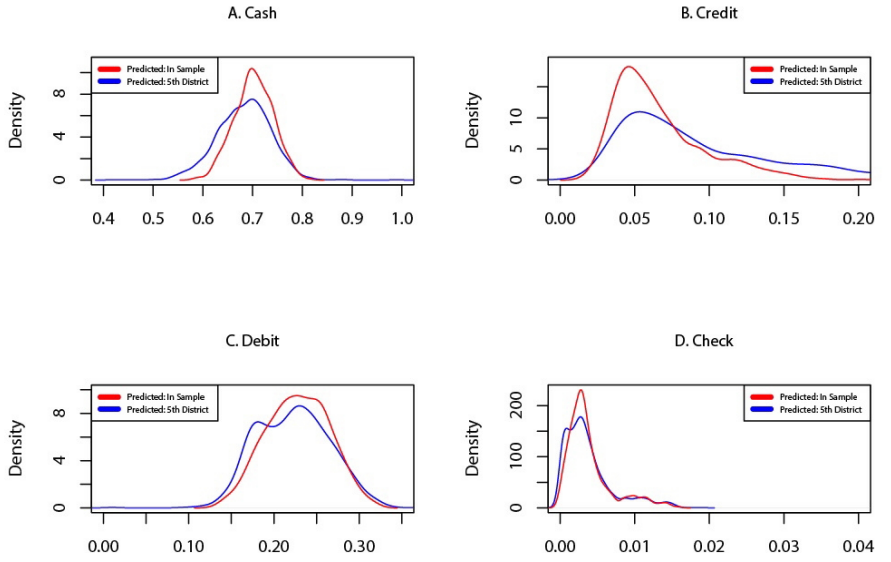
	Cash			Debit		
Scenarios	R^2	ΔR^2	$\frac{\Delta R^2}{sum(\Delta R^2)}$	R^2	ΔR^2	$\frac{\Delta R^2}{sum(\Delta R^2)}$
All variables included	0.531			0.374		
Constant median transaction size	0.425	0.106	23%	0.316	0.058	13%
Constant banking condition	0.527	0.004	1%	0.367	0.007	1%
Constant socioeconomic factors	0.490	0.041	9%	0.330	0.044	10%
Constant demographics	0.364	0.167	36%	0.259	0.115	25%
Constant education levels	0.457	0.074	16%	0.333	0.041	9%
Constant state fixed effects	0.472	0.059	13%	0.201	0.173	38%
Constant day of week & month effects	0.522	0.009	2%	0.359	0.015	3%

Payment Variation across the Entire Fifth District

The estimation results above allow us to project payment variation across the entire Fifth District for similar retail outlets. Comparing our data with the entire Fifth District, we notice that the store locations in our sample are not fully representative (Table 7 in the Appendix provides summary statistics for zip-code-level explanatory variables for the entire Fifth District). On average, store locations in our sample have fewer bank branches per capita, lower median household income, lower population density, and a smaller percentage of college graduates. The racial composition also differs from the rest of the Fifth District: there is a higher percentage of blacks, Hispanics, and Native Americans and a lower percentage of whites and Asians.

Based on the estimates from our regression model, we now address a counterfactual question: if the retail chain were to locate stores equally across the entire Fifth District, what would be the payment pattern? To answer the question, we first use the benchmark model to predict payment shares across the Fifth District with the assumption that all the zip-code locations in the Fifth District have the same median transaction size as the mean of the regression sample. The results are shown in Figure 7. We find that comparing with our regression sample, the entire Fifth District would show a similar pattern of payment variation: cash is being used most at this type of retail outlets, followed by debit, credit, and finally check. However, the relative share of these payment means would differ. We find that cash as well as debit and check would be used less in the rest of the Fifth District, while credit would be used more. This is consistent with the location bias of the stores in our sample, as discussed above.

Figure 7 Predicted Payment Variation: Sample Locations vs. Entire Fifth District

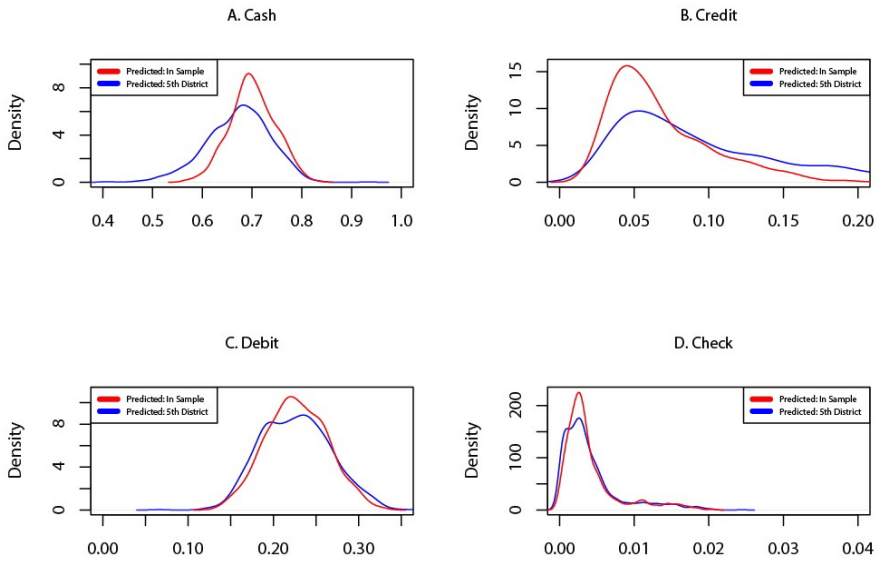


As a robustness check, we also redo the counterfactual exercise using the alternative regression model, in which we replace the median transaction size with the residual median transaction size. This takes into account that location-specific variables may also affect payment variation through their effects on transaction sizes. The results are plotted in Figure 8. As it turns out, Figures 7 and 8 are not very different.

4. PAYMENT VARIATION OVER THE LONGER TERM

The month-of-sample dummies in our regression identify changing payment mix over the longer term. Figure 9 plots the marginal effects for month-of-sample dummies. These effects combine seasonality with a time trend and idiosyncratic monthly variation. The vertical lines indicate each January in our sample years. The estimated annual time trends are -2.46 percentage points for cash, 1.69 percentage points for debit, 0.83 percentage points for credit, and -0.06 percentage points for check. This suggests a longer-term trend of declining cash shares at this retailer, largely replaced by debit.

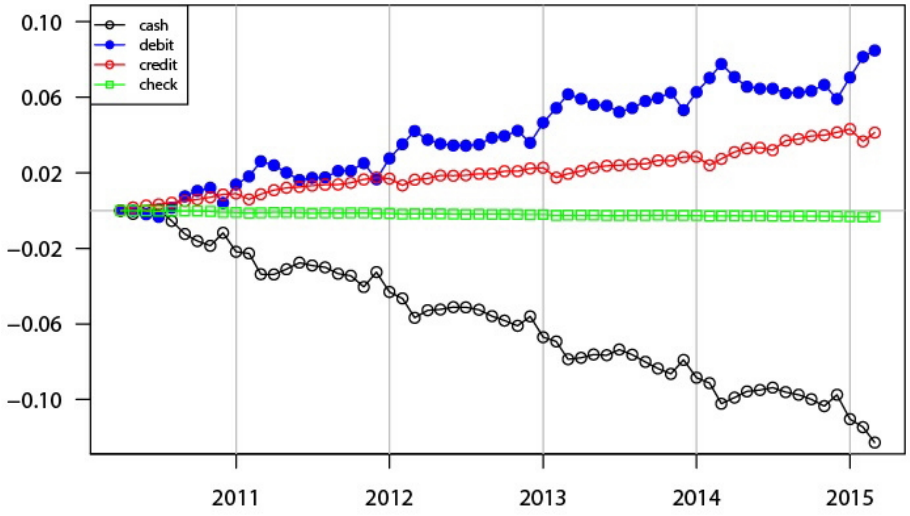
Figure 8 Predicted Payment Variation: A Robustness Check



The trend decline in the share of cash transactions is striking. Moreover, we plot the raw transactions data in Figure 10, which shows that this trend is not driven by any particular subset of stores or regions but is universal for the Fifth District. Exploring the driving forces behind the trend would be useful for understanding the changing demand for currency in retail transactions more broadly. We discuss several candidate factors below.

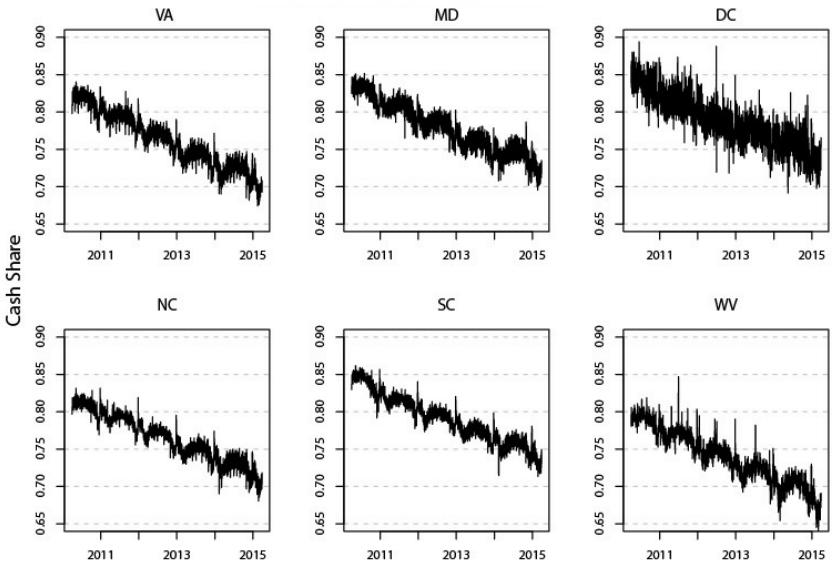
First, one may wonder whether the decline in cash over the five years of our sample could be driven by transitory factors, such as the Great Recession. According to the Boston Fed's latest report on consumer payments (Schuh and Stavins 2014), cash payments increased significantly after the financial crisis, replacing credit payments. Therefore, as the economy recovered from the recession, we may expect credit to have risen at the expense of cash. However, in our sample most of the cash decline was offset by an increase in debit. As Figure 1 shows, credit accounts for only about 4 percent of transactions at the beginning of the sample period and 7 percent at the end. And note that even 7 percent is an overestimate because our measure of credit includes signature debit and prepaid cards.

Figure 9 Month-of-Sample Marginal Effects



Another possible transitory factor is a change in the store's payment acceptance policy. However, as far as we know, there was a uniform payment policy in place across all the chain's stores during the sample period, with cash, debit, credit, and checks accepted on equal terms. Still, because our sample covers the implementation of the Durbin regulation on debit card interchange fees (effective on October 1, 2011), one may wonder if the chain had an incentive to steer customers toward more debit use. Again, this was unlikely. The Durbin regulation established a 21-cent cap on the debit interchange fees that financial institutions with more than $10 billion in assets can charge to merchants through merchant acquirers. However, we learned from the company that more than 50 percent of its debit transactions were exempt from the regulation because the debit cards used were issued by financial institutions with under $10 billion in assets. Moreover, the Durbin regulation is known for its unintended consequence of raising interchange fees to 21 cents for small-dollar transactions, which account for the vast majority of transactions at this retailer (Wang 2016). Therefore, if the new regulation were to have any impact on the stores in our sample, it should have caused them to try to reduce debit use rather than promote it.

Figure 10 The Decline in Cash Transactions in the Fifth District States and D.C.



Another question is whether the store altered the range of retail goods it sold during the sample period so that it attracted a clientele with different payment preferences. We cannot fully rule out this possibility given that we do not observe individual customers, but the company's annual financial reports indicate that the composition of goods sold did not undergo major changes during the period.

Given that the transitory factors discussed above are unlikely to explain the decline in the cash share at this retail chain, we then turn to longer-term factors. First, there could be an increasing trend of transaction sizes. It is indeed true that the average median transaction size at this retailer increased from \$6.27 to \$7.07 from 2010 to 2015. However, according to our estimation results, this could only account for a decline of cash shares of 1.47 percentage points out of the overall decline of 12.28 percentage points over the five years. Second, part of the time trend is presumably attributable to the change in zip-code-level variables. Recall that we treated all zip-code-level variables as fixed at their 2011 values across time in the regressions. Therefore any time trend is picked up by the month of sample dummies, even if some of the trend is actually associated with time variation in the

zip-code-level variables. However, as shown in Wang and Wolman (2016), the forecasted changes for the zip-code-level variables can only explain a relatively small portion of the decline of cash shares.

This leaves a large fraction of the time trend still to be explained. Prime candidates are technological progress and changing consumer perceptions of the attributes of debit payments relative to others. These attributes include adoption costs, marginal cost of transactions, speed of transactions, security, record keeping, general merchant acceptance, and ease of use, which are not directly included in our regressions.

While our data is from one retail chain, our exercise highlights the rise of debit in place of cash. In fact, debit has seen tremendous overall growth in the past decade. According to the latest Federal Reserve Payments Study (2014), it has risen to be the top noncash payment instrument in the U.S. economy: debit accounted for 19 percent of all noncash transactions in 2003, and its share doubled by 2012. Our study provides firsthand micro evidence that the increase in debit came at the expense of cash at a large cash-intensive retailer. Assuming that the shift from cash to debit is also occurring in retail more generally and that it continues, it could eventually be manifested in a decline in currency in circulation.

5. CONCLUSION

Using five years of transactions data from a large discount retail chain with hundreds of stores across the Fifth District, we study payment variation across locations and time. We find that the fraction of cash (noncash) transactions decreases (increases) with median transaction size and is affected by location-specific variables reflecting consumers' preferences and the opportunity costs of using cash relative to noncash means of payment. With the estimation results, we evaluate the relative importance of various factors in explaining the cross-location payment variation in our sample. We find that the median transaction size, demographics, education levels, and state fixed effects are the top factors. Taking those into consideration, we also project payment variation across the entire Fifth District for retail outlets similar to those in our sample.

We also identify interesting time patterns of payment variation. In particular, over the longer term, the shares of cash and check transactions decline steadily, while debit and credit shares rise. The overall cash fraction of transactions is estimated to have declined by 2.46 percentage points per year in our five-year sample period, largely replaced by debit. We show that the decline in cash at this particular retailer was likely not driven by transitory factors, and only a relatively small

fraction could be explained by changes in the median transaction size and the zip-code-level variables. This leaves a large fraction of the time trend to be explained, with prime candidates being technological progress in debit and changing consumer perceptions of debit relative to cash.

APPENDIX: THE FMLOGIT MODEL

The regression analysis in the paper uses the fractional multinomial logit model (FMLogit). The FMLogit model conforms to the multiple fractional nature of the dependent variables, namely that the fraction of payments for each instrument should remain between 0 and 1, and the fractions add up to 1. The FMLogit model is a multivariate generalization of the method proposed by Papke and Wooldridge (1996) for handling univariate fractional response data using quasi-maximum likelihood estimation. Mullahy (2010) provides more econometric details.

Formally, consider a random sample of $i = 1, ..., N$ zip-code-day observations, each with M outcomes of payment shares. In our context, $M = 4$, which correspond to cash, debit, credit, and check. Letting s_{ik} represent the k^{th} outcome for observation i, and x_i, $i = 1, ..., N$, be a vector of exogenous covariates. The nature of our data requires that

$$s_{ik} \in [0, 1] \qquad k = 1, ..., M;$$

$$\Pr(s_{ik} = 0 \mid x_i) \geq 0 \quad \text{and} \quad \Pr(s_{ik} = 1 \mid x_i) \geq 0;$$

$$\text{and} \quad \sum_{m=1}^{M} s_{im} = 1 \quad \text{for all } i.$$

Given the properties of the data, the FMLogit model provides consistent estimates by enforcing conditions (1) and (2),

$$E[s_k|x] = G_k(x; \beta) \in (0, 1), \quad k = 1, ..., M; \tag{1}$$

$$\sum_{m=1}^{M} E[s_m \mid x] = 1; \tag{2}$$

and also accommodating conditions (3) and (4),

$$\Pr(s_k = 0 \mid x) \geq 0 \qquad k = 1, ..., M; \tag{3}$$

$$\Pr(s_k = 1 \mid x) \geq 0 \qquad k = 1, ..., M; \tag{4}$$

where $\beta = [\beta_1, ..., \beta_M]$.[16] Specifically, the FMLogit model assumes that the M conditional means have a multinomial logit functional form in linear indexes as

$$E[s_k \mid x] = G_k(x; \beta) = \frac{\exp(x\beta_k)}{\sum_{m=1}^{M} \exp(x\beta_m)}, \qquad k = 1, ..., M. \tag{5}$$

As with the multinomial logit estimator, one needs to normalize $\beta_M = 0$ for identification purposes. Therefore, Equation (5) can be rewritten as

$$G_k(x; \beta) = \frac{\exp(x\beta_k)}{1 + \sum_{m=1}^{M-1} \exp(x\beta_m)}, \qquad k = 1, ..., M-1; \tag{6}$$

and

$$G_M(x; \beta) = \frac{1}{1 + \sum_{m=1}^{M-1} \exp(x\beta_m)}. \tag{7}$$

Finally, one can define a multinomial logit quasi-likelihood function $L(\beta)$ that takes the functional forms Equations (6) and (7) and uses the observed shares $s_{ik} \in [0, 1]$ in place of the binary indicator that would otherwise be used by a multinomial logit likelihood function, such that

$$L(\beta) = \prod_{i=1}^{N} \prod_{m=1}^{M} G_m(x_i; \beta)^{s_{im}}. \tag{8}$$

The consistency of the resulting parameter estimates $\hat{\beta}$ then follows from the proof in Gourieroux et al. (1984), which ensures a unique maximizer. In our regression analysis, we use Stata code developed by Buis (2008) for estimating the FMLogit model.

[16] To simplify the notation, the "i" subscript is suppressed in Equations (1)-(7).

Table 6 Marginal Effects for Zip-Code Variables

Variable		Cash	Debit	Credit	Check
Residual median transaction size		-0.018*	0.013*	0.004*	0.001*
Banking condition					
HHI		0.036*	-0.028*	-0.010*	0.002*
HHI*metro		-0.042*	0.036*	0.009*	-0.003*
Branches per capita		0.040*	-0.017*	-0.022*	-0.001*
Socioeconomic condition					
Robbery rate		-0.177*	0.158*	0.032*	-0.012*
Median household income		-0.018*	0.003*	0.024*	-0.008*
Population density		-0.623*	0.595*	0.118*	-0.091*
Family households		-0.158*	0.154*	0.010*	-0.006*
Housing: Owner occupied		0.005*	-0.038*	0.023*	0.010*
Vacant		-0.033*	-0.020*	0.046*	0.007*
Demographics					
Female		-0.074*	0.154*	-0.072	-0.009*
Age	15-34	-0.364*	0.351*	0.022*	-0.009*
	35-54	-0.485*	0.502*	-0.005	-0.012*
	55-69	-0.018*	0.027*	0.010*	-0.019*
	>70	-0.148*	0.144*	0.002	0.003*
Race	Black	0.093*	-0.068*	-0.016*	-0.009*
	Hispanic	-0.022*	-0.117*	0.131*	0.008*
	Native	0.125*	-0.075*	-0.045*	-0.005*
	Asian	0.115*	-0.074*	-0.036*	-0.004*
	Pacific Islander	-0.153	1.637*	-1.324*	-0.159*
	Other	0.257*	0.017*	-0.251*	-0.024*
	Multiple	0.220*	-0.309*	0.194*	-0.105*
Education level					
High school		-0.271*	0.162*	0.106*	0.003*
Some college		-0.278*	0.186*	0.090*	0.002*
College		-0.257*	0.153*	0.102*	0.002*
Pseudo R^2		0.604	0.534	0.607	0.559
Zip-code-day observations		1,021,764	1,021,764	1,021,764	1,021,764

Note: *1 percent significance level based on robust standard errors. The dependent variables are the fractions of each of the four general payment instruments used in transactions at stores in a zip code on a day between April 1, 2010, and March 31, 2015. The explanatory variables take their values in 2011. Banking HHI index is calculated by squaring each bank's share of deposits in a market (an MSA or a rural county) and then summing these squared shares. Metro is a dummy variable taking the value 1 when the banking market is an MSA, otherwise equal to zero. Branches per capita is measured as the number of bank branches per 100 residents in a zip code. Robbery rate is defined as the number of robberies per 100 residents in a county. Median household income is measured in units of $100,000 per household in a zip code. Population density is measured in units of 100,000 residents per square mile in a zip code. All the other variables are expressed as fractions.

Table 7 Summary Statistics of Zip-Code Variables (Entire Fifth District)

Variable (unit)		Mean	Std. dev.	1%	99%
Banking condition					
HHI metro		0.192	0.148	0.059	0.735
HHI rural		0.326	0.171	0.125	1.000
Branches per capita ($1/10^3$)		0.66	2.97	0.05	4.68
Socioeconomic condition					
Robbery rate ($1/10^5$)		30.13	41.30	0.00	235.07
Median household income ($)		50910	24859	22214	140093
Population density (per mile^2)		1157	2473	15	11514
Family households (%)		66.87	9.92	28.17	88.16
Housing (%): Renter occupied		27.17	13.73	6.52	77.14
Owner occupied		59.69	15.10	3.87	89.03
Vacant		13.14	10.70	2.41	64.89
Demographics (%)					
Female		50.73	3.22	35.64	55.18
Age	<15	18.22	3.92	5.76	27.27
	15-34	25.65	8.40	13.92	60.96
	35-54	27.46	3.82	12.12	35.05
	55-69	18.71	4.66	2.42	33.53
	>70	9.97	3.86	0.79	22.71
Race	White	72.18	21.73	10.56	98.95
	Black	19.80	19.70	0.07	82.09
	Hispanic	5.97	6.40	0.30	32.24
	Native	0.69	3.54	0.00	6.67
	Asian	2.44	4.27	0.00	23.34
	Pacific Islander	0.06	0.09	0.00	0.46
	Other	2.74	3.69	0.00	18.48
	Multiple	2.10	1.14	0.38	5.60
Education level (%)					
Below high school		15.20	11.38	0.00	54.00
High school		34.60	13.18	0.00	70.60
Some college		20.91	8.89	0.00	49.60
College		29.30	16.71	0.00	80.40

REFERENCES

Arango, Carlos, Kim P. Huynh, and Leonard Sabetti. 2011. "How Do You Pay? The Role of Incentives at the Point-of-Sale." Bank of Canada Working Paper 2011-23 (October).

Borzekowski, Ron, Elizabeth K. Kiser, and Shaista Ahmed. 2008. "Consumers' Use of Debit Cards: Patterns, Preferences, and Price Response." *Journal of Money, Credit, and Banking* 40 (February): 149–72.

Borzekowski, Ron, and Elizabeth K. Kiser. 2008. "The Choice at the Checkout: Quantifying Demand Across Payment Instruments." *International Journal of Industrial Organization* 26 (July): 889–902.

Buis, Maarten L. 2008. "FMLogit: Stata Module Fitting a Fractional Multinomial Logit Model by Quasi Maximum Likelihood." Statistical Software Components, Department of Economics, Boston College (June).

Ching, Andrew T., and Fumiko Hayashi. 2010. "Payment Card Rewards Programs and Consumer Payment Choice." *Journal of Banking and Finance* 34 (August): 1773–87.

Cohen, Michael, and Marc Rysman. 2012. "Payment Choice with Consumer Panel Data." Memo, Boston University Department of Economics.

Federal Reserve System. 2014. "The 2013 Federal Reserve Payments Study." https://www.frbservices.org/communications/payment_system_research.html (July).

Gilbert, R. Alton, and Adam M. Zaretsky. 2003. "Banking Antitrust: Are the Assumptions Still Valid?" Federal Reserve Bank of St. Louis *Review* 85 (November/December): 29–52.

Gourieroux, Christian, Alain Monfort, and Alain Trognon. 1984. "Pseudo Maximum Likelihood Methods: Theory." *Econometrica* 52 (May): 681–700.

Klee, Elizabeth. 2008. "How People Pay: Evidence from Grocery Store Data." *Journal of Monetary Economics* 55 (April): 526–41.

Koulayev, Sergei, Marc Rysman, Scott Schuh, and Joanna Stavins. 2016. "Explaining Adoption and Use of Payment Instruments by U.S. Consumers." *RAND Journal of Economics* 47 (Summer): 293–325.

McAndrews, James, and Zhu Wang. 2012. "The Economics of Two-Sided Payment Card Markets: Pricing, Adoption and Usage." Federal Reserve Bank of Richmond Working Paper 12-06 (October).

Mullahy, John. 2010. "Multivariate Fractional Regression Estimation of Econometric Share Models." National Bureau of Economic Research Working Paper 16354 (September).

Papke, Leslie E., and Jeffrey M. Wooldridge. 1996. "Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates." *Journal of Applied Econometrics* 11 (November/December): 619–32.

Schuh, Scott, and Joanna Stavins. 2014. "The 2011 and 2012 Surveys of Consumer Payment Choice." Federal Reserve Bank of Boston Research Data Report 14-1 (September).

Schuh, Scott, and Joanna Stavins. 2012. "How Consumers Pay: Adoption and Use of Payments." Federal Reserve Bank of Boston Working Paper 12-2.

Wang, Zhu. 2016. "Price Cap Regulation in a Two-sided Market: Intended and Unintended Consequences." *International Journal of Industrial Organization* 45 (March): 28–37.

Wang, Zhu, and Alexander L. Wolman. 2016. "Payment Choice and Currency Use: Insights from Two Billion Retail Transactions." *Journal of Monetary Economics* 84 (December): 94–115.

Zinman, Jonathan. 2009. "Debit or Credit?" *Journal of Banking and Finance* 33 (February): 358–66.

Economic Quarterly— Volume 102, Number 1— First Quarter 2016— Pages 79– 103

How Large Are Returns to Scale in the U.S.? A View Across the Boundary

Thomas A. Lubik

In this article, I investigate the size of the returns to scale in aggregate U.S. production. I do so by estimating the aggregate returns to scale within a theory-consistent general equilibrium framework using Bayesian methods. This approach distinguishes this article from much of the empirical literature in this area, which is largely based on production-function regressions and limited-information methods. The production structure within a general equilibrium setting, on the other hand, is subject to cross-equation restrictions that can aid and sharpen inference. My investigation proceeds against the background that increasing returns are at the core of business cycle theories that rely on equilibrium indeterminacy and sunspot shocks as the sources of economic fluctuations (e.g., Benhabib and Farmer 1994; Guo and Lansing 1998; Weder 2000).

Specifically, the theoretical literature has shown that multiple equilibria can arise when the degree of returns to scale is large enough. At the same time, the consensus of a large empirical literature is that aggregate production exhibits constant returns. However, equilibrium indeterminacy is a characteristic of a system of equations and can therefore not be assessed adequately with production function regressions. Instead, empirical researchers should apply full-information, likelihood-based methods to conduct inference along these lines. as not

I am grateful for useful comments by Huberto Ennis, Andreas Hornstein, Allen Sirolly, and John Weinberg that greatly improved the motivation and exposition of the article. The views expressed in this article are those of the author and should not necessarily be interpreted as those of the Federal Reserve Bank of Richmond or the Federal Reserve System. Correspondence address: Research Department, Federal Reserve Bank of Richmond. P.O. Box 27622, Richmond, VA 23261. Email: thomas.lubik@rich.frb.org.

DOI: http://doi.org/10.21144/eq1020103

allowing for indeterminacy leaves the empirical model misspecified. I therefore estimate the returns to scale in a theory-consistent manner using econometric methods that allow for indeterminate equilibria. I apply the methodology developed by Lubik and Schorfheide (2004) to bridge the boundary between determinacy and indeterminacy and estimate a theoretical model over the entire parameter space, including those parameter combinations that imply indeterminacy. This view across the boundary allows me to detect the possibility that data were generated under indeterminacy and provides the correct framework for estimating the returns to scale.

I proceed in three steps. First, I estimate a standard stochastic growth model with increasing returns to scale in production. In this benchmark specification, I estimate the model only on that region of the parameter space that implies a unique, determinate equilibrium to get an assessment of what a standard approach without taking into account indeterminacy would result in. The estimated model is based on the seminal paper of Benhabib and Farmer (1994). The mechanism that leads to increasing returns is externalities in the production process: individual firms have production functions with constant returns, but these are subject to movements in an endogenous productivity component that depends on the production decisions by all other firms in the economy. The key assumption is that individual firms take this productivity component as given and thereby do not take into account that increases in individual factor inputs also raise this productivity component. In the aggregate, the feedback effect from this mechanism can lead to increasing returns in the economy-wide production function. Benhabib and Farmer (1994) show analytically that if the strength of this feedback effect, tied to an externality parameter, is large enough, the resulting equilibria can be indeterminate in the sense that there are multiple adjustment paths to the steady state.

In this benchmark model with externalities, I find estimates that are tightly concentrated around the case of constant returns. Moreover, I also find that aggregate labor supply is fairly inelastic. This finding presents a problem for the existence of indeterminate equilibria due to increasing returns. It can be shown algebraically that the threshold required for an indeterminate equilibrium to arise depends on how elastic the labor supply is. Even with only mildly increasing returns, crossing the boundary into indeterminacy requires a perfectly elastic labor supply, both of which factors I can rule out from my estimation. Based on this baseline model with externalities, it would therefore seem unlikely that equilibrium indeterminacy would arise since the parameter estimates are far away from their threshold values.

In the second step, I therefore estimate a modified version of the benchmark model that allows for variable capacity utilization based on the influential paper by Wen (1998). He shows that the indeterminacy threshold is considerably closer to the constant-returns case when production is subject to variable capacity utilization, that is, when firms can vary the intensity with which the capital stock is used. Given typical parameter values from the literature, the required degree of increasing returns for an indeterminate equilibrium is within the range of plausible empirical estimates. When I estimate the model with variable capacity utilization, I find mildly increasing returns, but the statistical confidence region includes the constant-returns case. As in the benchmark model, I find an inelastic labor supply. In Wen's model, the threshold value of the returns-to-scale parameter is a function of the labor supply elasticity. The threshold attains a minimum for a perfectly elastic labor supply but rises sharply when labor becomes less elastic. Even with mildly increasing returns, these results indicate that indeterminacy will likely not arise in the framework with variable capacity utilization on account of the labor supply parameter.

A caveat to this conclusion is that the results are obtained by restricting the estimation to the determinate region of the parameter space. If the data are generated under parameters that imply indeterminacy, the thus-estimated model would be misspecified and the estimates biased. This potential misspecification would manifest itself as a piling up of parameter estimates near or at the boundary between determinacy and indeterminacy (Canova 2009; Morris 2016) or it might not be detected at all if there is a local mode of the likelihood function in the determinacy region.

In a third step, I therefore apply the methodology developed by Lubik and Schorfheide (2004) that takes the possibility of indeterminacy into account and allows a researcher to look across the boundary.[1] Reestimating the two models over the entire parameter space leave the original results virtually unchanged. Using measures of fit, I find that it is highly unlikely that U.S. data are generated from an indeterminate equilibrium and are driven by nonfundamental or sunspot shocks. The combination of at best mildly increasing returns and inelastic labor supply rule out indeterminacy even after correcting for potential biases in the estimation algorithm.[2]

[1] This notion is discussed in further detail in Lubik and Schorfheide (2004) and An and Schorfheide (2007).

[2] Conceptually, this article is closest to Farmer and Ohanian (1999). They estimate a model with variable capacity utilization and preferences that are nonseparable in consumption and leisure. This specification requires only a small degree of increasing returns to generate indeterminacy. Their empirical estimates indicate that returns

The article is structured as follows. In the next section, I specify the benchmark model, namely a standard stochastic growth model with externalities in production, and I discuss how this can imply increasing returns to scale and equilibrium indeterminacy. Section 2 describes my empirical approach and discusses the data used in the estimation. In the third section, I present and discuss results from the estimation of the benchmark model, while I extend the standard model in Section 4 to allow for variable capacity utilization. I address the issue of an indeterminate equilibrium as the source of business cycle fluctuations within this context in Section 5. The final section concludes and discusses limitations and extensions of the work contained in this article.

1. A FIRST PASS: THE STANDARD RBC MODEL WITH EXTERNALITIES

The benchmark model for studying returns to scale is the standard stochastic growth model with an externality in production. I use this model as a data-generating process from which I derive benchmark estimates for the returns to scale from aggregate data. Moreover, this model has been used by Benhabib and Farmer (1994) and Farmer and Guo (1994) to study the implications of indeterminacy and sunspot-driven business cycles. It will therefore also serve as a useful benchmark for capturing the degrees to scale when the data are allowed to cross the boundary between determinacy and indeterminacy.

In the model economy, a representative agent is assumed to maximize the intertemporal utility function:

$$E_0 \sum_{t=0}^{\infty} \beta^t \left[\log c_t - \chi_t \frac{n_t^{1+\gamma}}{1+\gamma} \right], \tag{1}$$

subject to sequences of the budget constraint:

$$c_t + k_{t+1} = A_t \bar{e}_t k_t^{\alpha} n_t^{1-\alpha} + (1-\delta) k_t, \tag{2}$$

by choosing sequences of consumption $\{c_t\}_{t=0}^{\infty}$, labor input $\{n_t\}_{t=0}^{\infty}$, and the capital stock $\{k_{t+1}\}_{t=0}^{\infty}$. The structural parameters satisfy the restrictions: $0 < \beta < 1$, $\gamma \geq 0$, $0 < \alpha < 1$, $0 < \delta < 1$, whereby β is the discount factor, γ the inverse of the Frisch labor supply elasticity, α the capital share, and δ the depreciation rate.

---

to scale are, in fact, increasing, but that U.S. data are nevertheless better described by the standard RBC model without sunspot shocks. This paper differs from theirs in that they estimate the model equation by equation without imposing cross-equation restrictions. Secondly, they do not formally test whether U.S. time series are better represented by a specification that allows for sunspot shocks. In this article, I conduct a formal test that can distinguish between the two variants.

The externality in the production process, $\overline{e}_t$, is taken parametrically by the agent. Conceptually, this means that when computing first-order conditions for the agent's problem, $\overline{e}_t$ is taken as fixed. It is only when equilibrium conditions are imposed ex post that the functional dependence of $\overline{e}_t$ on other endogenous variables is realized.[3] I assume that $\overline{e}_t$ depends on the average capital stock $\overline{k}_t$ and labor input $\overline{n}_t$:

$$\overline{e}_t = \left[\overline{k}_t^{\alpha}\overline{n}_t^{1-\alpha}\right]^{\eta-1}, \tag{3}$$

where the externality parameter $\eta \geq 0$ captures the returns to scale. When $\eta = 1$, production exhibits constant returns, while for $\eta > 1$ increasing returns are obtained. In equilibrium, $\overline{k}_t = k_t$ and $\overline{n}_t = n_t$. The social production function is thus given by:

$$y_t = A_t k_t^{\alpha\eta} n_t^{(1-\alpha)\eta}. \tag{4}$$

The model economy is driven by two exogenous shocks, technology A_t and preference χ_t, which captures variations in the disutility of working. I assume that A_t is a stationary first-order autoregressive process. Specifically, the level of technology is assumed to evolve according to:

$$A_t = (A_{t-1})^{\rho_A} e^{\varepsilon_t^A},\ \varepsilon_t^A \sim \mathcal{N}\left(0, \sigma_A^2\right), \tag{5}$$

where $0 \leq \rho_A < 1$ and mean technology is normalized to one. The shock ε_t^A is a zero-mean Gaussian innovation with variance σ_A^2. The preference process χ_t is also assumed to follow a stationary AR(1) process:

$$\chi_t = \left(\chi_{t-1}\right)^{\rho_\chi} e^{\varepsilon_t^\chi},\ \varepsilon_t^\chi \sim \mathcal{N}\left(0, \sigma_\chi^2\right), \tag{6}$$

where $0 \leq \rho_\chi < 1$. The preference shock alters the marginal rate of substitution between consumption and leisure.

The first-order conditions for this model form a system of equations that needs to be solved in order to provide a reduced form representation that serves as an input into the estimation procedure. This can be accomplished by approximating the equilibrium conditions in the neighborhood of the steady state using log-linearization. The resulting linear rational expectations model can then be solved using standard methods. I list the linearized equations that are used to estimate the model in the Appendix.

[3] A social planner would recognize this dependence and impose it ex ante, that is, before taking first-order conditions. It is this asymmetry that leads to lower social welfare in the benchmark case and creates a channel for welfare-improving tax policy, for instance. In addition, it creates the underpinning for equilibrium indeterminacy as Benhabib and Farmer (1994) show.

In a seminal paper, Benhabib and Farmer (1994) demonstrate that if the degrees of scale in production are large enough, then the model exhibits equilibrium indeterminacy. This has two implications for the behavior of the model. First, there are multiple adjustment paths to the unique steady state. Second, equilibrium dynamics can change markedly when compared to the determinate case in that nonfundamental shocks, "sunspots," can affect equilibrium outcomes and generate additional volatility. Benhabib and Farmer (1994) derive a simple analytical threshold condition for indeterminacy to arise in a continuous-time framework. The corresponding conditions for the discrete-time case, which are relevant for the model that I take to the data, are considerably more complicated, lengthy, and in parts not very intuitive. I list and discuss them in the Appendix. In order to develop intuition, I therefore derive insights based on the well-known Benhabib-Farmer condition first.

A necessary condition for indeterminacy to arise in Benhabib and Farmer (1994) is that the returns-to-scale parameter η is above a certain threshold given by the following:

$$\eta > \frac{1+\gamma}{1-\alpha}. \tag{7}$$

It has to be larger than the ratio between the exponent on the disutility of labor $1+\gamma$ and the labor share in production. Since the latter is a value between zero and one and typically found to be around two-thirds, indeterminacy in this model requires quite high increasing returns. This high level of a threshold is further exacerbated if the labor supply is less than perfectly elastic, that is, if $\gamma > 0$.

The intuition behind the condition is that if the returns to scale are large enough, the aggregate labor demand schedule is upward-sloping. In the standard case, workers are employed until their marginal product equals their wage. Hiring an additional worker reduces firm profits since the competitive wage would be higher than what the worker could produce at the margin. With production externalities as in (3), however, an additional feedback effect arises. At the margin, additional labor input raises economy-wide total factor productivity through its effect on $\overline{e}_t$, which feeds back on the competitive wage and counters the declining marginal product of labor. When this effect is large enough, labor demand starts sloping upward since the externality factor becomes dominant. In this scenario, the economy becomes susceptible to the influence of sunspot shocks that are unrelated to fundamentals such as productivity disturbances. When firms believe employment is higher than it should be given the fundamentals, this belief is self-validating in an indeterminate equilibrium: higher labor input leads to a stronger

externality, which raises production and, in turn, requires more labor input.

Since I am interested in taking this model to the data, I employ a discrete-time model. I list the corresponding analytical determinacy conditions in the Appendix. Generally speaking, the intuition from the continuous-time condition (7) carries over to discrete time, specifically the fact that the labor-demand schedule needs to be upward-sloping. I now turn to the first empirical exercise, where I estimate the standard RBC model with externalities to determine the returns to scale in the aggregate production function for the U.S. economy. I will do so against the background of the possibility of an indeterminate equilibrium in the data in case the indeterminacy conditions apply. Whether they do so is naturally an empirical question.

2. EMPIRICAL APPROACH

Bayesian Estimation

My empirical approach to the questions raised in this article is Bayesian DSGE estimation. This methodology is discussed in detail in An and Schorfheide (2007). The main object of investigation is the parameter vector θ, on which inference is conducted by extracting information from the observed data $Y^T = \{y_t\}_{t=1}^T$, with a sample size of T. The data are interpreted through the lens of a structural model, which provides restrictions necessary for parameter identification. A log-linear DSGE model can be written in terms of a state-space representation for y_t:

$$y_t = \Xi(\theta)\, s_t, \quad \Gamma_0(\theta)\, s_t = \Gamma_1(\theta)\, s_{t-1} + \Psi(\theta)\, \epsilon_t + \Pi(\theta)\, \eta_t, \qquad (8)$$

where the vector s_t collects the state variables of the theoretical model and where the coefficient matrices are shown as generally dependent on the structural parameters θ. ϵ_t is a vector of fundamental shocks, and the vector η_t collects the endogenous forecast errors of the rational expectations formation process in the parlance of Sims (2002). The model can be solved under determinacy and indeterminacy by the method described in Lubik and Schorfheide (2003).

Empirical evaluation in this Bayesian framework starts by specifying a probability distribution of the structural shocks $\{\epsilon_t\}_{t=1}^T$, from which a likelihood function $L\left(\theta|Y^T\right)$ can be obtained by means of the Kalman filter. The next step is to specify a prior distribution $p(\theta)$ over the structural parameters. The data Y^T are then used to update the prior through the likelihood function. The main concept in Bayesian inference is the posterior distribution $p(\theta|Y^T)$, which is the distribution of the parameters conditional on having seen the data. Moments of the posterior can then be used to characterize the parameter estimates.

The posterior distribution is computed according to Bayes' Theorem:

$$p(\theta|Y^T) = \frac{L(\theta|Y^T)p(\theta)}{\int L(\theta|Y^T)p(\theta)d\theta}, \tag{9}$$

whereby the denominator is the marginal data density, which can serve as a measure of overall model fit. Finally, the prior and posterior can be used to directly compare two different models or specifications, H_0 and H_1, as to which explains a given data set better. This is done by conducting a posterior odds test, which is similar to computing likelihood ratios. I apply this test later on to assess whether U.S. data are more likely to have been generated under determinacy or indeterminacy.

Data and Priors

I estimate all models in this article on quarterly U.S. data from 1954:3 to 2007:4.[4] I estimate the benchmark specifications on two data series, namely output and employment. Aggregate output y_t is measured as (the natural logarithm of) real per capita GDP. Since I assume that the model is driven by stationary shock processes, I need to remove any trends. I do so by passing the output series through an HP filter with smoothing parameter $\lambda = 1600$, which is standard for quarterly data. Employment n_t is measured as average weekly hours times employment from the Household Survey divided by population. I assume that the employment series is stationary, so that no further transformation is necessary. In the extended model discussed in Section 4, I also include a measure of capacity utilization in the data set. This is measured by the series available from the Board of Governors and reported as a percentage of industrial production. No further transformation is applied to these data series.

In a Bayesian DSGE estimation approach, a prior distribution needs to be specified for the model parameters. I largely choose prior means to be consistent with values established previously in the literature. The prior distributions are reported in Table 1. The specific form of the density is predicated by the type of parameter. A parameter restricted to lie on the unit interval is assumed to have a beta-distribution, while parameters on the real line are typically chosen to have gamma distributions, whereas variances are described by inverse gamma densities. I choose tight priors on the capital share and depreciation, but looser priors on the labor-supply elasticity and the returns-to-scale parameter.

[4] I choose to end my sample period at the onset of the Great Recession. The sharp decline in GDP would be difficult to capture even with HP-filtered data. Moreover, the

Table 1 Prior Distribution

Name	Range	Density	Mean	Std. Deviation
α	$[0,1)$	Beta	0.34	0.020
β	$[0,1)$	Beta	0.99	0.002
γ	$\mathbb{R}^+$	Gamma	2.00	0.500
δ	$[0,1)$	Beta	0.025	0.005
η	$\mathbb{R}^+$	Gamma	1.00	0.500
ρ_A	$[0,1)$	Beta	0.20	0.100
ρ_χ	$[0,1)$	Beta	0.95	0.050
σ_A	$\mathbb{R}^+$	InvGamma	N.A.	N.A.
σ_χ	$\mathbb{R}^+$	InvGamma	N.A.	N.A.

Note: The inverse gamma priors are of the form $p(\sigma|\nu,s) \propto \sigma^{-\nu-1}e^{-\nu s^2/2\sigma^2}$, where $\nu = 1$ and s equals 0.015. The prior is truncated at the boundary of the determinacy region.

Specifically, I set a very tight prior for the discount factor β with a mean of 0.99. The prior on the capital share α has a mean of 0.34 with a standard deviation of 0.02, while the depreciation rate δ has a mean of 0.025 with a standard deviation of 0.005. These are standard values in the calibration literature, but I allow for some flexibility in these parameters to somewhat adjust to the model environment at hand. I impose a more agnostic prior on the labor-supply elasticity parameter γ, where I impose some curvature on the disutility of labor with a mean of 2.0 and a standard deviation of 0.5. This value is somewhat distant from the case of a perfectly elastic labor supply with $\gamma = 0$. I choose the higher value since there is considerable evidence, both microeconomic and macroeconomic, that labor supply is not perfectly elastic and can be quite inelastic. I allow for some variation in this parameter because of the uncertainty surrounding this value.

The key parameter in this article is the degree of returns to scale, η. I center this value at the constant-returns case of 1 but assume a large standard deviation of 0.5. My underlying motivation is that I want the data to clearly dominate the posterior estimate. Finally, the parameters governing the two exogenous shock processes, technology A_t and preferences χ_t, are based on prior experience. The autocorrelation parameter for the technology process ρ_A has a mean of 0.95, while the corresponding value for ρ_χ is a slightly less persistent 0.9.

---

apparent shift in the level path of GDP that is visible in the data from 2008 on might affect parameter estimates.

3. SOME BASELINE ESTIMATION RESULTS

As my benchmark, I estimate the RBC model with externalities by letting all parameters vary freely over the admissible range as discussed in the model section above. Some of the parameter combinations would imply indeterminacy given the condition (7). As discussed before, the RBC model with externalities requires both very high labor supply elasticity (a small γ) and increasing returns for indeterminacy (a high enough $\eta > 1$). In particular, it would require values that are beyond those usually found in the literature. Studies using production function data such as Basu and Fernald (1997) typically find at best only mildly increasing returns at the aggregate level.

In the benchmark specification, I adopt a naive approach to the potential presence of indeterminate equilibria. I let myself be guided by prior studies that use limited information or single-equation methods that have nothing to say about indeterminacy since it is a property of a dynamic general equilibrium system (see the discussion in Lubik and Schorfheide [2004]). Prior inspection shows that it is highly unlikely that the returns to scale are large enough to meet the indeterminacy threshold. For instance, even with perfectly elastic labor supply, the indeterminacy condition in the continuous case would require a returns-to-scale parameter of $\eta > 1.5$ for $\alpha = 1/3$. It therefore seems a priori unlikely that the benchmark model would produce indeterminate outcomes based on typical parameter values found in the literature.

I thus proceed by estimating the model only over the determinate region over the parameter space. This procedure establishes a baseline as to what the parameter estimates that define the threshold between determinacy and indeterminacy would be if the model were restricted to a subset of the full admissible parameter space. I implement this numerically by penalizing the region of the parameter space that would imply indeterminacy for all possible draws from the joint prior distribution. I do so by throwing out all parameter combinations for which the solution algorithm of Sims (2002) returns an indeterminate equilibrium. More precisely, the solution algorithm rejects all draws that fall outside the determinacy bounds established by the analytical conditions given in the Appendix. This procedure implies that the prior distribution is restricted to the determinacy region only so that the search algorithm for the maximum of the likelihood function cannot venture into the indeterminacy region.

Table 2 reports the estimation results from the RBC model. The column labeled "Baseline Model" contains those from the baseline specification described above. The estimates of the capital share α, the discount factor β, and the depreciation rate δ are consistent with those commonly used in the calibration literature, respectively, 0.33, 0.99,

Table 2 Parameter Estimation Results, RBC Model

	Baseline Model		Restricted Model: $\eta = 1$		Restricted Model: $\gamma = 0$	
	Mean	**90% Interval**	**Mean**	**90% Interval**	**Mean**	**90% Interval**
α	0.331	[0.301, 0.349]	0.335	[0.310, 0.368]	0.329	[0.285, 0.371]
β	0.986	[0.979, 0.995]	0.990	[0.982, 0.994]	0.988	[0.979, 0.995]
γ	2.061	[1.573, 2.671]	2.332	[1.871, 2.904]	0.000	–
δ	0.022	[0.014, 0.030]	0.025	[0.017, 0.031]	0.026	[0.021, 0.030]
η	0.982	[0.894, 1.060]	1.000	–	0.912	[0.796, 0.923]
ρ_A	0.980	[0.968, 0.998]	0.981	[0.971, 0.998]	0.987	[0.971, 0.999]
ρ_χ	0.945	[0.901, 0.987]	0.974	[0.921, 0.995]	0.979	[0.960, 0.985]
σ_A	0.005	[0.003, 0.008]	0.018	[0.009, 0.029]	0.018	[0.014, 0.022]
σ_χ	0.019	[0.010, 0.025]	0.042	[0.030, 0.051]	0.030	[0.021, 0.040]

Note: The table reports posterior means and 90 percent coverage regions (in brackets). The posterior summary statistics are calculated from the output of the posterior simulator.

and 0.02, although the latter is contained in a fairly wide 90 percent probability interval. Posterior estimates of the autoregressive parameters tend to be high, which is a common observation in small-scale Bayesian DSGE models. The posterior mean of the scale parameter η is 0.98 with a 90 percent coverage range of $[0.89, 1.06]$. The posterior for this parameter is thus firmly centered on a small region around the constant-returns-to-scale case, which would rule out any possibility of indeterminacy. In addition, the labor supply parameter γ has a posterior mean of 2.06. For this value, the minimum required degree of returns to scale to result in indeterminacy would have to be 4.57.

To be fair, the joint prior distribution over the parameter space put virtually no probability mass on the indeterminacy region even before restricting the solution to determinate equilibria, and it was centered on constant returns. To gauge the sensitivity of the estimation, I experimented with various alternative starting values and priors. The results proved to be robust to different starting values, as the iterations of the algorithm quickly approached the benchmark posterior mode, even for high values of η. There was also no evidence that the algorithm would pile up at the boundary of the parameter space, that is, the threshold between determinacy and indeterminacy, which Morris (2016) suggests is evidence of misspecification. I obtained similar results when varying the prior distribution, specifically in the direction of a higher mean of η and a tighter distribution. Posterior mode estimates quickly converged to the benchmark case. This suggests the conclusion that restricting the model to the determinacy regions does not bias the findings since

the indeterminacy regions are far away from plausible parameterizations consistent with the data.

As a second exercise, I estimate the model under the restriction $\eta = 1$. The results are reported in Table 2 in the column labeled "Restricted Model: $\eta = 1$." This is the case of the standard RBC model as in King, Plosser, and Rebelo (1988). It is well-known that the standard RBC model does not admit indeterminate equilibria, so that I do not have to restrict the parameter space over which the model is estimated. The parameter estimates were virtually unchanged with the exception of the labor parameter γ, which increased to 2.33. As the benchmark results indicated before, the estimation algorithm settles quickly and closely on the constant-returns-to-scale case. Therefore, conditioning on this value, $\eta = 1$, should not affect the other parameter estimates much.

Alternatively, I fix $\gamma = 0$ (see Table 2, last column). This specification corresponds to the benchmark case of Benhabib and Farmer (1994) with perfectly elastic labor supply. Under this specification, the required returns to scale for equilibrium indeterminacy are considerably lower, namely at 1.5 given the standard parameterization of $\alpha = 1/3$. This restriction results in a posterior mean of $\eta = 0.91$. The algorithm thus pushed the returns-to-scale parameter in an opposite direction of what would be needed to cross the indeterminacy threshold. I explored this specification a bit further by imposing a tight prior on η with a mean of 1.60 and a standard deviation of 0.05. Even in this case, the resulting posterior mean is 0.99, as in the unrestricted benchmark case. It seems clear that the information in the data strongly prefers mildly decreasing returns to scale in production.[5]

I can formally compare the different specifications by computing their marginal data densities (MDD). These can be thought of as comparable to maximum likelihood values in that they capture the value of the posterior with all parameters integrated out. They also form the basis of posterior odds tests, which allow econometricians to discriminate between two alternative models in terms of overall fit. Given even prior probabilities on the two competing models, the model with the higher MDD can be considered as the better descriptor of the data. I report the MDDs in Table 3. Clearly, the information in the data draws the posterior strongly toward decreasing returns. The restricted model with $\eta = 1$ dominates all others, as can be seen in the first row.

[5] Arguably, the standard RBC model is misspecified in that it assumes constant returns to scale. However, the degree of uncertainty around this value is such that it encompasses constant returns.

Table 3 Marginal Data Densities and Posterior Odds Tests

Marginal Data Densities				
	Baseline	$\eta = 1$	$\gamma = 0$	Sunspot
RBC Model	128.75	129.81	88.43	–
Cap. Util.	138.92	130.01	–	120.34

Note: Marginal data densities are approximated by Geweke's (1999) harmonic mean estimator.

Moreover, comparison of the MDDs allows us to reject the specification with a perfectly elastic labor supply by a wide margin.[6]

In order to get a sense of the driving forces behind the data as interpreted through this specific model, I also compute variance decompositions. The results are broadly similar across different model specifications. Therefore, I only report those for the baseline model in Table 4. Technology shocks determine about 80 percent of fluctuations in output, the remainder are made up by shocks to preferences, namely the disutility of working. In contrast, these labor supply shocks are the main determinants of labor input in the amount of roughly two-thirds of the overall variability.

I can draw some preliminary conclusions at this point. Overall, I do not find any evidence of increasing returns in aggregate U.S. data under the assumption that the standard RBC model with production externalities is the data-generating process. The results show that the estimates are tightly clustered around the constant-returns case with more probability mass on decreasing returns. Even if we are willing to allow for increasing returns in contrast to what the data say, estimates for η are not at the level required for indeterminacy in an environment with perfectly elastic labor supply. In addition, the estimated aggregate labor supply elasticity is far too low to generate indeterminacy at any remotely plausible level of increasing returns.[7,8]

[6] The difference between the two values of the MDDs is almost 50 on a log scale. With even prior odds, that is equal prior probability on each model being the data-generating process, this amounts to a probability one acceptance of the constant-returns-to-scale model with inelastic labor supply.

[7] To the best of my knowledge, no empirical study has found increasing returns of that magnitude. Baxter and King (1991) come closest with $\eta = 1.6$.

[8] The main caveat for this conclusion is that the model is estimated under the restriction that the equilibrium is determinate. By doing so, I rule out any possibility of finding considerable returns to scale a priori. In effect, the model is misspecified along this dimension. The robustness checks that I performed show, however, that this is not the case. In that sense, the additional restriction to the space of determinate equilibria is not much of a restriction at all. This may be different for other models.

Table 4 Variance Decompositions

	Technology		Preference		Sunspot/ Measurement	
	Mean	90% Interval	Mean	90% Interval	Mean	90% Interval
	Standard RBC					
Output	0.81	[0.74, 0.90]	0.19	[0.08, 0.29]		
Labor	0.36	[0.30, 0.51]	0.64	[0.58, 0.71]		
	Variable Capacity Utilization					
Output	0.94	[0.89, 0.98]	0.06	[0.02, 0.11]		
Labor	0.13	[0.09, 0.19]	0.87	[0.81, 0.89]		
	Variable Capacity Utilization with Sunspots					
Output	0.81	[0.74, 0.86]	0.04	[0.01, 0.06]	0.15	[0.12, 0.23]
Labor	0.08	[0.04, 0.12]	0.21	[0.15, 0.29]	0.71	[0.60, 0.82]
	Variable Capacity Utilization with Utilization Data					
Output	0.92	[0.88, 0.95]	0.02	[0.00, 0.04]	0.06	[0.01, 0.12]
Labor	0.01	[0.00, 0.02]	0.67	[0.58, 0.76]	0.32	[0.22, 0.42]
Utilization	0.32	[0.24, 0.39]	0.06	[0.02, 0.09]	0.62	[0.52, 0.74]

4. VARIABLE CAPACITY UTILIZATION AND INCREASING RETURNS

The main conclusion from my empirical analysis of the Benhabib and Farmer (1994) model is that the degree of returns to scale necessary for indeterminacy to arise is implausibly large. This has been noted in the literature, which evolved toward developing frameworks that lead to a lower threshold value. A key paper following up on this issue is Wen (1998), who introduces variable capacity utilization into an otherwise standard Benhabib-Farmer model.[9] He is able to show that the degree of increasing returns required for indeterminacy is considerably less than in the standard model. I now use his framework to reassess the conclusion drawn in the previous section. I proceed as before in that I first estimate the model by restricting the parameter space to the determinacy region. This establishes a baseline to assess whether disregarding the possibility of indeterminacy has an effect on parameter estimates. This issue is addressed in the subsequent section.

[9] It has long been recognized that variable capacity utilization is an important component of business cycle analysis. In a key paper, Burnside and Eichenbaum (1996) demonstrate that variable capital utilization can significantly reduce the volatility of technology shocks required to replicate observed business cycles in otherwise standard models. Moreover, Basu and Fernald (1997) point out that production function regressions need to allow for variable capacity utilization in order to be able to remove endogenous components from total factor productivity and to get unbiased estimates of the returns to scale.

I assume that a representative agent maximizes the intertemporal utility function (1) as before. The budget constraint is modified by introducing variable capacity utilization u_t:

$$c_t + k_{t+1} = A_t \overline{e}_t \left(u_t k_t\right)^{\alpha} \left(n_t\right)^{1-\alpha} + (1-\delta_t)k_t. \tag{10}$$

$u_t \in (0,1)$ is the rate of capacity utilization. Given the capital stock k_t, which is predetermined in the current period, changes in utilization affect production and present an additional margin of adjustment. This captures the idea that the capital stock is sometimes left idle and that in general the utilization rate of machinery varies over time, depending on demand conditions, shift work, the work week, and other factors. Varying productive capacity gives firms a margin along which profits can be optimized by preemptively hoarding capital in anticipation of future demand conditions. However, changes in utilization come at a cost since capacity variation affects the depreciation rate. The more intensely the capital stock is utilized, the faster it depreciates. As in Wen (1998), I assume for simplicity a monotonic relationship between u_t and the depreciation rate δ_t:

$$\delta_t = \frac{1}{\theta} u_t^{\theta}, \tag{11}$$

where θ is a parameter. I can find the first-order conditions by maximizing the utility function (1) subject to the budget constraint and the definition of the depreciation rate by choosing sequences of consumption $\{c_t\}_{t=0}^{\infty}$, labor input $\{n_t\}_{t=0}^{\infty}$, capacity utilization $\{u_t\}_{t=0}^{\infty}$, and capital stock $\{k_{t+1}\}_{t=0}^{\infty}$.

As in the standard RBC model, I assume that $\overline{e}_t$ captures the externality in the production process and is taken parametrically by the agent. Under this specification, $\overline{e}_t$ depends on the average capital stock $\overline{k}_t$, labor input $\overline{n}_t$, and capacity utilization $\overline{u}_t$:

$$\overline{e}_t = \left[\left(\overline{u}_t \overline{k}_t\right)^{\alpha} \left(\overline{n}_t\right)^{1-\alpha}\right]^{\eta-1}, \tag{12}$$

where the externality parameter $\eta \geq 0$ captures the returns to scale. As before, production exhibits constant returns when $\eta = 1$ and returns to scale are increasing for $\eta > 1$. The determinacy conditions for this model are listed in the Appendix.

I estimate the model using Bayesian methods as discussed above. For comparison purposes, I estimate the model on the same two data series, output and labor input, and for the same two shocks, technology and labor disutility. In a robustness check, I further utilize data on capacity utilization and allow for the presence of sunspot shocks and measurement error. A convenient feature of the choice of the depreciation cost function is that it implies the same number of independent parameters to be estimated. The existence of a steady state

imposes a parametric restriction between θ and the depreciation rate: $\theta = \frac{1-\beta(1-\delta)}{\beta\delta}$. That is, the depreciation cost elasticity is not an independent parameter, but is determined by the steady-state depreciation rate and vice versa. I can therefore choose to treat steady-state depreciation parametrically. Consequently, I impose the same prior on δ and on the other parameters in the model. This implies a prior mean of $\theta = 1.40$. The empirical difference between the benchmark and the extended model only lies in the different dynamics via the introduction of capacity utilization and endogenous depreciation but not in different priors.

The estimation results for the extended model are reported in Table 5. The first set of results is contained in the left column, labeled "Baseline Model," where I allow all parameters to vary freely over the determinacy regions. That is, I throw out all parameter draws that would imply an indeterminate equilibrium just as I did in the benchmark case for the standard model. The parameter estimate that stands out is a high $\gamma = 8.46$, which implies a very inelastic labor supply and thereby likely rules out the possibility of indeterminate equilibria on account of increasing returns. The baseline estimates also show a lower capital elasticity of $\alpha = 0.27$ and a higher depreciation rate of $\delta = 0.05$ than in the standard RBC model. These estimates are consistent with those found in the literature on variable capacity utilization and reflect the impact of the latter on adjusting input margins in production as suggested by Burnside, Eichenbaum, and Rebelo (1995). Moreover, the implied estimate at the posterior means of the depreciation cost parameter is $\theta = 1.20$.

As to the question of increasing returns, I estimate the externality parameter $\eta = 1.09$ with a 90 percent coverage region of $[0.98, 1.17]$. This is higher than in the standard RBC model, although the constant-returns case is included in this coverage region. Incidentally, this value is right at the preferred estimate of Laitner and Stolyarov (2004), who estimate a full set of structural equations derived from a business cycle model using a methods of moments approach that is independent of whether the data are generated from a determinate or indeterminate equilibrium. What is intriguing about this result is that the returns to scale are at the threshold for indeterminacy in the baseline calibration in Wen (1998). Yet, as I argued in the previous section, the other critical parameter is the labor supply elasticity. In his benchmark calibration, Wen (1998) assumes perfectly elastic supply with $\gamma = 0$, whereas the posterior mean in my estimation is considerably higher. While I cannot rule out mild increasing returns empirically, the other parameter estimates imply that the equilibrium is not indeterminate.

Table 5 Parameter Estimation Results, Variable Capacity Utilization

	Baseline Model		Restricted Model: $\eta = 1$		Restricted Model: $\gamma = 0$	
	Mean	90% Interval	Mean	90% Interval	Mean	90% Interval
α	0.274	[0.261, 0.286]	0.254	[0.241, 0.275]	0.201	[0.182, 0.259]
β	0.991	[0.989, 0.994]	0.993	[0.987, 0.996]	0.994	[0.989, 0.999]
γ	8.459	[6.987, 9.801]	12.90	[10.45, 14.86]	0.903	[0.420, 1.681]
δ	0.049	[0.043, 0.055]	0.058	[0.053, 0.064]	0.089	[0.082, 0.099]
η	1.087	[0.975, 1.174]	1.000	–	1.384	[1.121, 1.605]
ρ_A	0.982	[0.969, 0.996]	0.067	[0.059, 0.072]	0.966	[0.958, 0.980]
ρ_χ	0.958	[0.949, 0.968]	0.965	[0.944, 0.991]	0.850	[0.791, 0.921]
σ_A	0.036	[0.030, 0.041]	0.041	[0.036, 0.044]	0.048	[0.042, 0.056]
σ_χ	0.094	[0.085, 0.099]	0.086	[0.070, 0.110]	0.079	[0.054, 0.097]
σ_ζ					0.163	[0.081, 0.303]

Note: The table reports posterior means and 90 percent probability intervals (in brackets). The posterior summary statistics are calculated from the output of the posterior simulator.

As a first robustness check, I estimate a restricted version of the model where I fix $\eta = 1$, which shuts down the externality feedback. The effects on the parameter estimates are somewhat larger than in the corresponding exercise for the RBC model. The posterior mean of α declines to 0.25, while depreciation rate δ increases to 0.06. The labor supply parameter is now estimated at 12.90. However, as the MDDs in the second row of Table 3 show, the unrestricted version is much preferred in terms of overall fit. Interestingly enough, the model with variable capacity utilization also dominates the standard RBC model in explaining labor input and GDP. Finally, I also compute the variance decompositions, which are reported in Table 4. The relative importance of the two shocks, technology and preference, in explaining the two data series is unchanged compared to the first model specification.

As a second robustness check, I also estimate the model using the Federal Reserve's data on capital utilization.[10] Adding a third observable variable to the model requires an additional source of uncertainty in order to avoid a singular likelihood function. I choose to add a measurement error to the observation equation that links the data series to its counterpart in the model instead of introducing an additional

[10] Available at: https://www.federalreserve.gov/releases/g17/

shock. I find that the parameter estimates do not change in any significant manner. The likely reason is that the utilization series mirrors the output series very closely and thus does not contain enough information to improve the empirical model.[11]

Bayesian estimates of a standard RBC model with variable capacity utilization that allows for increasing returns to scale via production externalities show that the U.S. economy is characterized by mildly increasing returns. This stands in contrast with the results derived from the model without capacity utilization, which found constant returns. This leaves open the possibility that the equilibrium in the U.S. economy may be indeterminate given the mechanism described in this article. What goes against this argument is that indeterminacy also requires a low labor supply elasticity. Estimates from both models show that labor is, in fact, fairly inelastically supplied. However, since I restricted the estimation to the determinacy region of the parameter space, I cannot be confident of the soundness of this conclusion. In the next section, I therefore look across the boundary of the determinacy region and estimate the model under indeterminacy.

5. ARE U.S. BUSINESS CYCLES DRIVEN BY SUNSPOT FLUCTUATIONS?

I now follow the implications of the theoretical model to their logical end and assess whether the observed U.S. data are generated under indeterminacy. As the discussion above shows, equilibrium indeterminacy requires a high degree of increasing returns (a large enough estimate of η) and a high labor supply elasticity (a low enough estimate of γ). In all estimated specifications, the labor supply elasticity turned out to be too low for the equilibrium to be indeterminate even if the externalities parameter was within a range that would otherwise have put the economy across the boundary, namely in the model with variable capacity utilization. However, these estimates should be understood against the background that I ruled out indeterminate equilibria a priori by restricting the prior to that region of the parameter space where there is a unique equilibrium.[12]

[11] At the same time, the measurement error explains about one-third of the fluctuations in the utilization series (see Table 4), which does suggest the model is not well-specified to capture movements in utilization that are independent of the output series.

[12] I do not find any indication across the various specifications that the posterior estimates are clustering near the indeterminacy threshold. As discussed in Canova (2009) and Morris (2016), this pile-up of probability mass near the boundary could be seen as evidence that the model is misspecified since indeterminacy is not explicitly accounted for. Nevertheless, it cannot be ruled out that a posterior mode is well within the indeter-

I therefore reestimate the two model specifications over the full parameter space using the methodology developed by Lubik and Schorfheide (2003, 2004), who show how to write the full set of indeterminate equilibria in a reduced form. The estimation algorithm can be used to reveal which of the many indeterminate equilibria the data reflect. At the same time, the indeterminate solution allows for the influence of an additional exogenous disturbance, namely nonfundamental sunspot shocks, in addition to the two fundamental shocks from before. I use the same data series for the estimation as in the benchmark case to ensure comparability across the result. I should note, however, that allowing for indeterminacy and sunspot shocks gives the estimation algorithm additional degrees of freedom to fit the data.

In estimating the models under indeterminacy, the first issue I face is that my chosen benchmark prior puts only small probability mass on the indeterminacy region. This is particularly problematic in the standard RBC model where even in the case of $\gamma = 0$, the required threshold value for η equals 1.5. Allowing for a wider dispersion in these two key parameters does not seem to make much of a difference. I therefore experimented with shifting the prior means. I found that a prior mean of $\eta = 2.6$ with a standard deviation of 0.1 and almost perfectly elastic labor supply would be needed to support a posterior estimate in the indeterminacy region. Since these values are far outside what can be considered a plausible range, it seems safe to rule out estimates based on this prior. Consequently, I argue that the Benhabib and Farmer (1994) model cannot be used to support the notion of sunspot-driven business cycles since it is simply inconsistent with the data.

I face a similar issue in the case of the Wen (1998) model. Under the benchmark prior, there is not much mass in the indeterminacy region. The limiting factor is again the labor supply elasticity parameter γ, which needs to be close to zero to be able to support an indeterminate equilibrium. Experimenting with the prior, I find, however, that a prior mean for η of 1.7 with a standard deviation of 0.2 puts enough mass beyond the boundary. Using this prior, I reestimate the specification with variable capacity utilization. The results are reported in the last column of Table 5. The posterior mean of the elasticity parameter $\eta = 1.38$, which is higher than in the benchmark case. At the same time, the estimate of $\gamma = 0.90$, which guarantees that the equilibrium is indeterminate. As Table 3 shows, however, the MDDs indicate that the indeterminacy specification is rejected relative to the

---

minacy region and can therefore not be detected when the parameter space is restricted to determinacy.

benchmark specification even when taking into account the higher degrees of freedom afforded by the model solution under indeterminacy. Table 4 reports the variance decompositions for the indeterminacy specification. Although I can conclude that the data are unlikely to have been generated under indeterminacy, it is interesting to determine how much of an effect sunspot shocks may have on economic fluctuations. The contribution to output fluctuations is small, around 15 percent, whereas sunspots drive a substantial fraction of labor input.

Are U.S. business cycles driven by sunspot fluctuations? Not if one believes that the source of these sunspot fluctuations lies in increasing returns to scale. Based on the results in this section, I can rule out the standard RBC model with production externalities as in Benhabib and Farmer (1994) as the data-generating process for a possible sunspot equilibrium. The extension of Wen (1998) to include variable capacity utilization is more promising, but the statistical support for indeterminacy is quite weak. As the results from the preceding sections show, aggregate U.S. production likely exhibits constant returns to scale, which rules out equilibrium indeterminacy a priori.

6. CONCLUSION

This article studies the returns to scale in aggregate U.S. data by estimating various specifications of the standard RBC model. In order to allow for the possibility of increasing returns in production, so as not to impose constant returns a priori, I introduce aggregate production externalities as in the framework of Benhabib and Farmer (1994). The degree of returns to scale can then be tied to a single parameter that measures the strength of the externality effect. In a second model specification, I also introduce variable capacity utilization as in Wen (1998), who generally reduces the required degree of increasing returns needed to support indeterminacy. All model specifications present in this paper admit the possibility of equilibrium indeterminacy to the effect that business cycles could be driven by extraneous, nonfundamental shocks.

I estimate the various specifications using Bayesian DSGE methods. I find strong evidence for constant returns to scale in aggregate U.S. data. Specifications that impose increasing returns are rejected based on standard model selection criteria. I show in a simple robustness exercise that a substantial degree of increasing returns can only be supported by imposing implausible priors. Equilibrium indeterminacy in the modeling frameworks used in this article requires a high enough degree of increasing returns and a low enough labor supply elasticity. My estimates show that even if increasing returns were present, we can rule out indeterminacy on account of an inelastic labor supply.

Therefore, a theory of sunspot-driven business cycles should not rely on increasing returns to scale in production.

The empirical results are to some extent model-dependent. My conclusion as to the possibility of indeterminate equilibria appears robust as the framework based on production externalities requires implausible labor supply elasticities. Nevertheless, alternative model setups may imply different, less stringent requirements for indeterminacy. Prime candidates are models with alternative utility functions, such as nonseparability in consumption and leisure (see Bennett and Farmer 2000) or models with multiple sectors (see Benhabib, Meng, and Nishimura 2000). Finally, if researchers are interested in sunspot shocks as potential driving forces for business cycles, exploring other avenues than production externalities seems a more promising option. For instance, Lubik and Schorfheide (2004) show that the Great Inflation of the 1970s was caused by sunspot shocks since the Federal Reserve pursued monetary policy that was not aggressive enough in fighting inflation. More recently, Golosov and Menzio (2015) have proposed a novel theoretical framework that generates sunspot-driven business cycles through idiosyncratic and firm-specific uncertainty over the quality of their workers.

REFERENCES

An, Sungbae, and Frank Schorfheide. 2007. "Bayesian Analysis of DSGE Models." *Econometric Reviews* 26 (May): 113–72.

Basu, Susanto, and John G. Fernald. 1997. "Returns to Scale in U.S. Production: Estimates and Implications." *Journal of Political Economy* 105 (April): 249–83.

Baxter, Marianne, and Robert G. King. 1991. "Productive Externalities and Business Cycles." Institute for Empirical Macroeconomics at the Federal Reserve Bank of Minneapolis Discussion Paper 53 (November).

Benhabib, Jess, and Roger E. A. Farmer. 1994. "Indeterminacy and Increasing Returns." *Journal of Economic Theory* 63 (June): 19–41.

Benhabib, Jess, Qinglai Meng, and Kazuo Nishimura. 2000. "Indeterminacy under Constant Returns to Scale in Multisector Economies." *Econometrica* 68 (November): 1541–48.

Bennett, Rosalind L., and Roger E. A. Farmer. 2000. "Indeterminacy with Non-separable Utility." *Journal of Economic Theory* 93 (July): 118–43.

Burnside, Craig, and Martin Eichenbaum. 1996. "Factor-Hoarding and the Propagation of Business-Cycle Shocks." *American Economic Review* 86 (December): 1154–74.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo. 1995. "Capital Utilization and Returns to Scale." In *NBER Macroeconomics Annual 1995*, vol. 10, edited by Ben S. Bernanke and Julio Rotenberg. Cambridge, Mass.: MIT Press, 67–124.

Canova, Fabio. 2009. "What Explains The Great Moderation in the U.S.? A Structural Analysis." *Journal of the European Economic Association* 7 (June): 697–721.

Farmer, Roger E. A., and Jang-Ting Guo. 1994. "Real Business Cycles and the Animal Spirits Hypothesis." *Journal of Economic Theory* 63 (June): 42–72.

Farmer, Roger E. A., and Lee Ohanian. 1999. "The Preferences of the Representative American." Manuscript.

Geweke, John. 1999. "Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communications." *Econometric Reviews* 18 (February): 1–73.

Golosov, Mikhail, and Guido Menzio. 2015. "Agency Business Cycles." National Bureau of Economic Research Working Paper 21743 (November).

Guo, Jang-Ting, and Kevin J. Lansing. 1998. "Indeterminacy and Stabilization Policy." *Journal of Economic Theory* 82 (October): 481–90.

King, Robert G., Charles I. Plosser, and Sergio T. Rebelo. 1988. "Production, Growth and Business Cycles: I. The Basic Neoclassical Model." *Journal of Monetary Economics* 21 (March–May): 195–232.

Laitner, John, and Dmitriy Stolyarov. 2004. "Aggregate Returns To Scale and Embodied Technical Change: Theory and Measurement Using Stock Market Data." *Journal of Monetary Economics* 51 (January): 191–233.

Lubik, Thomas A., and Frank Schorfheide. 2003. "Computing Sunspot Equilibria in Linear Rational Expectations Models." *Journal of Economic Dynamics and Control* 28 (November): 273–85.

Lubik, Thomas A., and Frank Schorfheide. 2004. "Testing for Indeterminacy: An Application to U.S. Monetary Policy." *American Economic Review* 94 (March): 190–217.

Meng, Qinglai, and Jianpo Xue. 2009. "Indeterminacy and E-stability in Real Business Cycle Models with Factor-Generated Externalities." Manuscript.

Morris, Stephen D. 2016. "DSGE Pileups." *Journal of Economic Dynamics and Control*, forthcoming.

Sims, Christopher A. 2002. "Solving Linear Rational Expectations Models." *Computational Economics* 20 (October): 1–20.

Weder, Mark. 2000. "Animal Spirits, Technology Shocks and the Business Cycle." *Journal of Economic Dynamics and Control* 24 (February): 273–95.

Wen, Yi. 1998. "Capacity Utilization under Increasing Returns to Scale." *Journal of Economic Theory* 81 (July): 7–36.

APPENDIX: INDETERMINACY CONDITIONS

Benhabib and Farmer (1994) derive analytical conditions that are necessary for indeterminacy in a continuous-time version of the RBC model with externalities. As it turns out, the corresponding conditions for the discrete-time version are considerably more complex. Meng and Xue (2009) derive these conditions for general forms of utility and production with externalities. Under the restriction $\eta \geq 1$ and logarithmic utility, the necessary and sufficient conditions for indeterminacy are (see Meng and Xue [2009], Proposition 4, case (i)):[13]

$$\Gamma_2 \eta < \frac{1+\gamma}{1-\alpha} < \Gamma_1 \eta,$$

where $\Gamma_1 = \frac{(2-\delta)(1+\beta(1-\delta))-\frac{\beta}{\alpha}\left(\frac{1-\beta}{\beta}+\delta\right)^2}{2\left(2+\eta\frac{1-\beta}{\beta}+\delta(\eta-1)\right)}$ and $\Gamma_2 = \frac{(1-\beta)(1-\delta)}{\eta\frac{1-\beta}{\beta}+\delta(\eta-1)}$. The condition has the familiar form that links a minimum value of the externalities parameter η to the labor supply elasticity and the capital share but is harder to interpret than the corresponding continuous-time restriction. If we just look at the necessary condition, then we have:

$$\eta > \frac{1+\gamma}{1-\alpha}\frac{1}{\beta(1-\delta)}.$$

Wen (1998) derives necessary and sufficient conditions for equilibrium indeterminacy in his model with capacity utilization. The general analytical conditions are more cumbersome than those for the standard RBC model with externalities. Wen (1998) therefore restricts his analysis to the case such that $\alpha\eta < \theta$, whereby $\theta = \frac{1-\beta(1-\delta)}{\beta\delta}$, based on the steady-state restriction linking the endogenous depreciation rate δ and the parameter θ. Under this restriction, necessary and sufficient conditions for indeterminacy are:

$$\begin{aligned}
\eta &< \frac{1}{\alpha}, \\
\eta &> 1 + \frac{\theta\left(1+\gamma-\beta\left(1-\alpha\right)\right)-\left(1+\gamma\right)\alpha}{\beta\left(1-\alpha\right)\theta+\left(1+\gamma\right)\alpha-\frac{1-\beta}{1+\beta}\left(1+\gamma\right)\theta}, \\
\eta &> 1 + \frac{\theta\left(1+\gamma-\beta\left(1-\alpha\right)\right)-\left(1+\gamma\right)\alpha+\frac{1-\beta}{1+\beta}\left(1+\gamma\right)\beta\delta\left(\theta-\alpha\right)\frac{1-\alpha\theta}{2\alpha}}{\beta\left(1-\alpha\right)\theta+\left(1+\gamma\right)\alpha-\frac{1-\beta}{1+\beta}\left(1+\gamma\right)\left(\theta-\frac{1}{2}\left(\theta-\alpha\right)\left(1-\beta\right)\right)}.
\end{aligned}$$

[13] Meng and Xue (2009) consider two additional cases where indeterminacy arises when $\eta < 1$, that is, when there are decreasing returns to scale. Although I allowed for these cases in the benchmark specification based on a wide prior centered on $\eta = 1$, I did not encounter indeterminate equilibria in this region when estimating the model.

The third condition differs from the second by additional terms in the numerator and the denominator. As Wen (1998) demonstrates, they are virtually identical for β closest to one. It is fairly straightforward to show that the threshold value for η, beyond which an indeterminate equilibrium arises is increasing in γ. That is, the less elastic labor supply is, the less likely is an indeterminate equilibrium.