FEDERAL RESERVE BANK of ST. LOUIS

RESEARCH DIVISION

Working Paper Series

Money, Liquidity and Welfare

Yi Wen

Working Paper 2014-003B
https://doi.org/10.20955/wp.2014.003

January 2015

FEDERAL RESERVE BANK OF ST. LOUIS
Research Division
P.O. Box 442
St. Louis, MO 63166

Money, Liquidity and Welfare*

Yi Wen

(This version: January 30, 2015)

Abstract

This paper develops an analytically tractable Bewley model of money demand to shed light on some important questions in monetary theory, such as the welfare cost of inflation. It is shown that when money is a vital form of liquidity to meet uncertain consumption needs, the welfare costs of inflation can be extremely large. With log utility and parameter values that best match both the aggregate money demand curve suggested by Lucas (2000) and the variance of household consumption, agents in our model are willing to reduce consumption by $3\% \sim 4\%$ to avoid 10% annual inflation. The astonishingly large welfare costs of inflation arise because inflation increases consumption risk by eroding the buffer-stock-insurance value of money, thus hindering consumption smoothing at the household level. Such an inflation-induced increase in consumption risk at the micro level cannot be captured by representative-agent models or the Bailey triangle. Although the development of financial intermediation can mitigate the problem, with realistic credit limits the welfare loss of moderate inflation still remains several times larger than estimations based on the Bailey triangle. Our findings provide a strong justification for adopting a low inflation target by central banks, especially in developing countries where money is the major form of household financial wealth.

Keywords: Liquidity Preference, Heterogenous Money Demand, Financial Intermediation, Velocity, Welfare Costs of Inflation.

JEL codes: D10, D31, D60, E31, E41, E43, E49, E51.

*This paper is a revised version of Wen (2009). I thank Pengfei Wang for discussions on issues related to this project. My gratitude also goes to two anonymous referees, Costas Azariadis, Jinghui Bai, Aleksander Berentsen, Gabriele Camera, Yongsung Chang, Mark Huggett, Ayse Imrohoroglu, David Levine, Narayana Kocherlakota, Qing Liu, Rodolfo Manuelli, Steve Williamson, Tao Zhu, and seminar participants at Georgetown University, Hong Kong University of Science & Technology, Taiwan University, the Federal Reserve Banks of San Francisco and St. Louis, and the 2013 SED conference in Seoul for comments; Judy Ahlers for editorial assistance; and Yuman Tam and Luke Shimek for research assistance. The views expressed do not reflect official positions of the Federal Reserve Bank of St. Louis and the usual disclaimer applies. Correspondence: Yi Wen, Research Department, Federal Reserve Bank of St. Louis, P.O. Box 422, St. Louis, MO, 63166, and School of Economics and Management, Tsinghua University, Beijing, China. Phone: 314-444-8559. Fax: 314-444-8731. Email: yi.wen@stls.frb.org.

1 Introduction

In developing countries, liquid money (cash and checking accounts) is the major form of household financial wealth and a vital tool of self-insurance (precautionary saving) to buffer idiosyncratic shocks because of the lack of a well-developed financial system. Based on recent data in China and India, more than 90% of the household financial wealth is held in the form of cash and checking accounts.[1] Even in developed countries, because of borrowing constraints and costs of participating in the financial markets, money remains one of the most important assets to provide liquidity to smooth consumption for low-income households. Mulligan and Sala-i-Martin (2000) document that the majority of households in the United States do not hold financial assets other than checking accounts. In particular, based on the Survey of Consumer Finances (SCF), on average (over the period 1989-2007) nearly 60% of U.S. households do not hold any nonmonetary financial assets (interest-bearing assets), and about 50% of those that hold checking accounts do not hold any interest-bearing assets. In addition, money demand is highly heterogeneous: The Gini coefficient of the distribution of money across households is greater than 0.85 in the United States. This degree of heterogeneity in money demand closely resembles the distribution of financial wealth instead of consumption (with a Gini coefficient less than 0.3). This suggests that the liquidity motive of money demand is at least as important (if not more so) as the transaction motive of money demand, even in developed countries such as the United States.[2]

When money is essential (as a store of value) for consumption smoothing and is unequally distributed across households, largely because of idiosyncratic needs for liquidity and the lack of sophisticated risk sharing, inflation can be far more costly than recognized by the existing literature, which suggests that the welfare cost of inflation is less than 1% of aggregate output (see, e.g., Lucas, 2000).[3]

This paper argues that to properly assess the welfare cost of inflation in developing countries (as well as for low income people in rich countries), it is desirable to use a theoretical model that takes

[1] Townsend (1995) points out that currency and crop inventory are the major forms of liquid assets to provide self-insurance against idiosyncratic shocks for farmers in India and Thailand, and surprisingly, purchases and sales of real capital assets, including livestock and consumer durables, do not play a role in smoothing income fluctuations.

[2] Ragot (2009) reports that this stylized fact holds for other developed countries and argues that this is a problem for theories that directly link money demand to consumption, such as cash-in-advance (CIA), money-in-the-utility (MIU), or shopping-time models, but is consistent with incomplete-market models in which money is held as a form of financial asset that provides liquidity to smooth consumption.

[3] Small welfare costs of inflation are also obtained by many others, such as Cooley and Hansen (1989), Dotsey and Ireland (1996), Henriksen and Kydland (2010), and Khan, King, and Wolman (2003) in different models. Lagos and Wright (2005) obtain a significantly higher welfare cost of inflation in a search model of money—about 4% of aggregate consumption with a 10% inflation rate. Our welfare results are comparable to those obtained by Lagos and Wright (2005) in the order of magnitude, but with an entirely different mechanism and micro foundation.

the liquidity function of money and the precautionary motives of money demand into account, so as to capture the buffer-stock-insurance value of cash in addition to the opportunity cost of forgone interest as suggested by Bailey (1956). The loss of the insurance value of money under inflation may generate far larger welfare costs than implied by the Bailey triangle because inflation reduces real money demand and exposes more cash-poor households to idiosyncratic risks by destroying the liquidity value of cash.

This paper constructs such a model by generalizing Bewley's (1980, 1983) precautionary money demand model into a tractable, dynamic stochastic general equilibrium (DSGE) framework where money can coexist with other assets (such as capital).[4] The key feature distinguishing Bewley's model from the related monetary literature, such as the heterogeneous-agent cash-in-advance (CIA) model of Lucas (1980) and the (S,s) inventory-theoretic model of Baumol (1952) and Tobin (1956), is that money is held solely as a store of value, completely symmetric to any other asset, and is not imposed from outside as the means of payments. Agents can choose whether to hold money depending on the costs and benefits. By freeing money from its role of medium of exchange, Bewley's approach allows us to focus on the inventory function of money as a pure form of liquidity, so that the welfare implications of the liquidity-preference theory of money demand can be investigated in isolation. Beyond Bewley (1980, 1983), my generalized model is analytically tractable; hence, it greatly simplifies the computation of equilibrium in DSGE environments with both idiosyncratic and aggregate shocks, capital accumulation, financial intermediation, and nontrivial distributions of cash balances, thus facilitating welfare and business-cycle analysis. Analytical tractability also makes the mechanisms of the model highly transparent.

The major finding of the paper is that persistent money growth is very costly. When the model is calibrated to match not only the interest elasticity of aggregate money demand but also the extent of idiosyncratic risk faced by households in the data, the implied welfare cost of increasing the inflation rate from 0% to 10% per year is around $3\% \sim 4\%$ of consumption (or even higher).

Since holding money is both beneficial (providing liquidity) and costly (forgoing interest payment and bearing the inflation tax), agents opt to hold different amounts of cash depending on income levels and consumption needs. As a result, a key property of the model is an endogenously determined distribution of money holdings across households, with a strictly positive fraction of households being cash-constrained (i.e., with zero cash balances) in equilibrium. Hence, lump-sum money injections have an immediate positive impact on consumption for the cash-constrained agents, but not for agents with idle cash balances. Consequently, the aggregate price does not

[4] General-equilibrium analysis with capital accumulation is important. Cooley and Hansen (1989) emphasize the general-equilibrium effect of inflation on output through substituting leisure for consumption in the face of positive inflation, which causes labor supply and output to decline. However, because these authors assume that money is held only for transaction purposes, the welfare cost of inflation is still small despite the general-equilibrium effects of inflation on output, about $0.4\% \sim 0.5\%$ of GDP with 10% inflation.

increase with the aggregate money supply one for one, so transitory monetary shocks are expansionary to aggregate output (even without open market operations), the velocity of money is countercyclical, and the aggregate price appears "sticky."

However, with anticipated inflation, permanent money growth reduces welfare significantly for several reasons: (i) Precautionary money demand induces agents to hold excessive amounts of cash to avoid liquidity constraints, raising the inflation tax on the population. (ii) Cash-poor agents suffer disproportionately more from the inflation tax because they are more likely to be subject to idiosyncratic risks without self-insurance; thus, for the same amount of reduction in real wealth, inflation reduces their expected utility more than it does for liquidity-abundant agents.[5] (iii) The size of the liquidity-constrained population (with zero cash balances) rises rapidly with inflation, leading to an increased portion of the population unable to smooth consumption against idiosyncratic shocks.[6] This last factor can dramatically raise social welfare costs along the extensive margin.

The Bailey triangle is a poor measure of the welfare cost of inflation because it fails to capture the insurance function of money (as noted and emphasized by Imrohoroglu, 1992). At a higher inflation rate, not only does the opportunity cost of holding money increase (which is the Bailey triangle), but the crucial benefit of holding money also diminishes. In particular, when demand for money declines, the portion of the liquidity-constrained population rises; consequently, the welfare cost of inflation increases sharply due to the loss of self-insurance for an increasingly larger proportion of the population. This result is reminiscent of the analysis by Aiyagari (1994) in which he shows that the welfare cost of the loss of self-insurance in an incomplete-market economy is equivalent to a 14% reduction in consumption even though his calibrated model matches only one-third of the income and wealth inequalities in the data.

This paper is also related to the work of Alvarez, Atkeson, and Edmond (2008). Both papers are based on an inventory-theoretic approach with heterogeneous money demand and can explain the short-run dynamic behavior of velocity and sticky aggregate prices under transitory monetary shocks. However, my approach differs from theirs in important aspects. Most notably, their model is based on the Baumol-Tobin inventory-theoretic framework where money is not only a store of value but also a means of payment (similar to CIA models). In their model, agents are exogenously and periodically segregated from the banking system and the CIA constraint always binds. For these reasons, the implications for the welfare cost of inflation in their model may also be very different from those in this paper. For example, Attanasio, Guiso, and Jappelli (2002) estimate the

[5] This asymmetric effect of inflation is related but different from the distributional effect emphasized by Erosa and Ventura (2002) in a heterogeneous-agent model where rich households rely more on credit transactions than low-income households.

[6] In this paper, the term "liquidity-constrained agents" is synonymous to "households with a binding liquidity constraint" or "those with zero cash balances."

welfare cost of inflation based on a simple Baumol-Tobin model and find the cost to be less than 0.1% of consumption under 10% inflation. The main reason is that this segment of the literature has relied exclusively on Bailey's triangle (or the interest elasticity of money demand) to measure the welfare cost of inflation. Hence, despite having heterogeneous money holdings across households, such research is not able to obtain significantly larger estimates of the cost of inflation than those in representative-agent models.

Bewley's (1980) model has been studied in the recent literature, but the main body of this literature focuses on an endowment economy. For example, Imrohoroglu (1992), Imrohoroglu and Prescott (1991), and Akyol (2004) study the welfare cost of inflation in the Bewley model. To the best of my knowledge, Imrohoroglu (1992) is the first in the literature to recognize that the welfare cost of inflation in a Bewley economy is larger than that suggested by the Bailey triangle. However, like Bewley's (1980) work, this segment of the literature is based mainly on an endowment economy without capital and these models are not analytically tractable.[7]

The rest of this paper is organized as follows: Section 2 presents the benchmark model on the household side and shows how to solve for individuals' decision rules analytically under borrowing constraints and idiosyncratic risks. It reveals some of the basic properties of a monetary model based on liquidity preference. Section 3 extends the model to a production economy with capital and uses the model to evaluate the welfare cost of inflation. Section 4 introduces credit and banking into the general-equilibrium model and discusses some robustness issues of our welfare results. Section 5 concludes the paper.

2 The Benchmark Model

The model features money as the only asset that can be adjusted quickly (costlessly) to buffer idiosyncratic shocks to consumption demand at any moment. Interest-bearing nonmonetary assets (such as capital) can be accumulated to support consumption but are not as useful (or liquid) as money in buffering idiosyncratic shocks. This setup captures the characteristics of households in developing countries (or poor households in rich countries) for whom money (cash and checking accounts) is the major form of household financial wealth and a vital tool of self-insurance (precautionary saving) to buffer idiosyncratic shocks.

[7] This literature tends to find higher welfare costs of inflation, but the absolute magnitude is still small. For example, Imrohoroglu (1992) shows the welfare cost of 10% inflation is slightly above 1% of consumption. The reason is that the distribution of liquidity demand in the model does not respond significantly to inflation if the support of idiosyncratic shocks is binary or does not have enough points (as typically assumed in this literature to reduce numerical computation burdens). Hence, an extensive margin would be missing and this margin is important for the welfare costs of inflation. Our study also complements the analysis of Telyukovay and Visschers (2013), who found in a Lagos-Wright model that the welfare cost of inflation is higher when agents hold money for insurance purpose in addition to transaction purpose. Specifically, idiosyncratic uncertainty can increase the welfare cost of 10% annual inflation (relative to the Friedman rule) from 0.2% to 0.5% of consumption.

We make the model analytically tractable by introducing two important features: (i) We allow an endogenous labor supply with quasi-linear preferences (as in Lagos and Wright, 2005), and (ii) we replace idiosyncratic labor income shocks typically assumed in the incomplete-market literature (e.g., Imrohoroglu, 1989, 1992; Aiyagari, 1994; Huggett, 1993) by preference shocks. Even with quasi-linear preferences, the model is not analytically tractable if wage income is subject to idiosyncratic shocks. There are two ways to overcome this difficulty. One is to place idiosyncratic shocks on preferences (i.e., to the marginal utility of consumption as in Lucas, 1980), and the other is to place them on total gross wealth, which includes labor income. This paper takes the first approach. Both approaches yield similar results for the welfare cost of inflation when they are calibrated to match some key features of aggregate money demand and the idiosyncratic liquidity risk faced by households. This is reassuring because it suggests that the source of uninsurable idiosyncratic shocks does not matter for our welfare results.[8]

Time is discrete. There is a unit mass of continuum households in the interval $[0, 1]$. Each household is subject to an idiosyncratic preference shock, θ_t, which is iid across both households and time and has the distribution $\mathbf{F}(\theta) \equiv \Pr[z \leq \theta]$ with support $[\theta_L, \theta_H]$. A household chooses sequences of consumption $\{c_t\}$, labor supply $\{n_t\}$, and nominal money balance $\{m_{t+1}\}$ to maximize lifetime utilities, taking as given the paths of aggregate real wage $\{W_t\}$, the aggregate price $\{P_t\}$, and the nominal lump-sum transfers $\{\tau_t\}$. The nominal rate of return to holding money is zero.[9]

Assume that in each period t the decisions for labor supply and holdings for interest-bearing assets (if any) must be made before observing the idiosyncratic shock θ_t in that period, and the decisions, once made, cannot be changed for the rest of the period (i.e., these markets are closed afterward for households until the beginning of the next period). Thus, if there is an urge to consume during period t after labor supply and capital investment decisions are made and the preference shock θ_t is realized, money is the only asset that can be adjusted to smooth consumption. Borrowing of liquidity (money) from other sources is not allowed.[10] These assumptions imply that households may find it optimal to carry money as self-insurance to cope with idiosyncratic uncertainty (as in Bewley, 1980), even though money is not required as a medium of exchange.

An alternative way of formulating the above information structure for decision-making is to divide each period into two subperiods, with labor supply and nonmonetary-asset investment determined in the first subperiod, the rest of the variables (consumption and money balances) deter-

[8] See an earlier version of this paper (Wen, 2010) for analyses based on the alternative approach.

[9] The zero lower bound on government bonds implies that the nominal return to money is zero. In addition, the nominal interest rate on checking accounts is essentially zero in many countries. However, since we may consider any perfectly liquid asset, including interest bearing checking accounts, as money, we can add a fixed interest rate on money. But doing so does not change our main results. This can be seen by simply redefining the time discount factor in our model as the product of β and a fixed deposit rate. However, see Section 4 in this paper for the analysis with a time varying and endogenously determined deposit rate.

[10] This assumption will be relaxed in Section 4.

mined in the second subperiod, and the idiosyncratic shocks θ_t realized only in the beginning of the second subperiod. Yet another alternative specification of the model is to have two islands, with labor supply and interest-bearing assets (if any) determined on island 1 and c_t and m_t determined on island 2 simultaneously by two spatially separated household members (e.g., a worker and a shopper), but only the shopper—who determines consumption and money balances in island 2—can observe θ_t in period t. Both members can observe aggregate shocks and the history of family decisions up to period t. At the end of each period the two members reunite and share everything perfectly (e.g., income, wealth, and information) and separate again in the beginning of the next period.[11]

2.1 Household Problem

We use lower-case letters to denote individual variables and upper-case letters to denote aggregate variables in this paper. Denote h^t as the history of an individual household up to period t, and h^t_{-1} as h^t excluding θ_t, namely, $h^t = h^t_- \cup \theta_t$. Let H^t denote the history of the aggregate state up to period t. Then the problem of a household is to solve

$$\max_{c_t(h^t,H^t),m_{t+1}(h^t,H^t),n_t(h^t_-,H^t)} E_0 \sum_{t=0}^{\infty} \beta^t \left\{\theta_t \log c_t(h^t,H^t) - a n_t(h^t_-,H^t)\right\} \tag{1}$$

subject to

$$c_t(h^t,H^t) + \frac{m_{t+1}(h^t,H^t)}{P_t(H^t)} \leq \frac{m_t(h^{t-1},H^{t-1}) + \tau_t}{P_t(H^t)} + W_t(H^t) n_t(h^t_-,H^t), \tag{2}$$

$m_{t+1}(h^t,H^t) \geq 0$, $n_t(h^t_-,H^t) \in [0,\bar{n}]$, and $m_0 \geq 0$ given; where τ_t is an exogenous, uniform, lump-sum nominal transfer (to be specified later). To save notation, we drop the history indices $\left\{h^t,H^t\right\}$ by denoting $P_t = P_t(H^t)$, $W_t = W_t(H^t)$, $c_t = c_t(h^t,H^t)$, $m_{t+1} = m_{t+1}\left(h^t,H^t\right)$, and $n_t = n_t(h^t_-,H^t_t)$, unless confusion may arise. Without loss of generality, assume $a = 1$ in the utility function.[12]

[11] Since the time period in the model can be short (e.g., t represents a month, a week, or a day), the assumption that labor supply and nonmonetary asset holdings (such as fixed capital) are predetermined and cannot be adjusted instantaneously in the second subperiod after the realization of θ_t is not as extreme as it appears. This information/timing structure amounts to creating a necessary friction for the existence of money as a liquid asset. In reality, especially in developing countries, it is costly to exchange labor and real assets (such as land and livestock) for consumption goods in spot markets (e.g., due to search frictions and other transaction costs). In developed countries, even government bonds are rarely held as a major form of liquid assets by low-income households and there are always costs involved in trading nonmonetary assets. As documented by Telyucova (2011) using household survey data, even credit cards are not as liquid as cash in meeting certain types of consumption demand.

[12] The model remains tractable if the utility function takes the more general form of $\frac{c_t^{1-\sigma}-1}{1-\sigma} - n_t$. For simplicity we set $\sigma = 1$ in this paper. Setting $\sigma > 1$ can only enhance our conclusions.

The household problem can be formulated recursively. Define

$$x_t \equiv \frac{m_t + \tau_t}{P_t} + W_t n_t \tag{3}$$

as real wealth, and denote $J_t(x_t, \theta_t)$ as the value function of the household based on the choice of c_t and m_{t+1} after the realization of θ_t. We then have

$$J_t(x_t, \theta_t) = \max_{c_t, m_{t+1}} \left\{ \theta_t \log c_t + \beta E_t V_{t+1}(\frac{m_{t+1}}{P_{t+1}}) \right\} \tag{4}$$

subject to

$$c_t + \frac{m_{t+1}}{P_t} \leq x_t \tag{5}$$

$$m_{t+1} \geq 0, \tag{6}$$

where $V_t(\frac{m_t}{P_t})$ is the value function of the household based on the choice of n_t before observing θ_t. That is,

$$V_t(\frac{m_t}{P_t}) = \max_{n_t} \left\{ -n_t + \int J_t(x_t, \theta_t) d\mathbf{F} \right\} \tag{7}$$

subject to (3) and $n_t \in [0, \bar{n}]$.

Since money is not required as a medium of exchange, nor does it provide utility, a monetary equilibrium is a belief-driven equilibrium. In what follows, we focus on the monetary equilibrium where money is accepted as a store of value and the aggregate price $P_t \in (0, \infty)$ is finite and bounded away from zero.

Proposition 1 *The decision rules for consumption, money demand, and real wealth are given, respectively, by*

$$c_t = \min \left\{ 1, \frac{\theta_t}{\theta_t^*} \right\} x_t \tag{8}$$

$$\frac{m_{t+1}}{P_t} = \max \left\{ \frac{\theta_t^* - \theta_t}{\theta_t^*}, 0 \right\} x_t \tag{9}$$

$$x_t = W_t \theta_t^* R(\theta_t^*), \tag{10}$$

where the cutoff θ_t^* is independent of individual history h^t and is determined implicitly by the following Euler equation:

$$\frac{1}{W_t} = \left[\beta E_t \frac{P_t}{P_{t+1} W_{t+1}} \right] R(\theta_t^*), \tag{11}$$

where

$$R(\theta_t^*) \equiv \int \max\left\{1, \frac{\theta_t}{\theta_t^*}\right\} d\mathbf{F} > 1. \tag{12}$$

Proof. See Appendix A1. ■

Consumption is a concave function of real wealth, with the marginal propensity to consume given by $\min\left\{1, \frac{\theta_t}{\theta_t^*}\right\}$, which is less than 1 in the case of a low urge to consume ($\theta_t < \theta_t^*$). Saving (money demand) is a buffer stock: Agents save in the low-return asset when consumption demand is low ($\frac{m_{t+1}}{P_t} > 0$ if $\theta_t < \theta_t^*$), anticipating that future consumption demand may be high ($\Pr[\theta > \theta^*] > 0$).

Equation (11) implicitly determines the optimal cutoff $\theta^*(H^t)$ as a function of the aggregate state only. The interpretation of equation (11) is straightforward. Treat $\frac{1}{W}$ as the marginal utility of consumption from wage income. The left-hand-side (LHS) of the equation is the opportunity cost of holding one more unit of real balances as inventories (as opposed to increasing consumption by one unit). The right-hand-side (RHS) is the expected gains by holding money, which take two possible values: The first term inside the integral of equation (12) reflects simply the discounted and inflation-adjusted next-period utility value of inventories (real balances) in the case of a low urge to consume (since 1 dollar is just 1 dollar if not consumed), which has probability $\int_{\theta \leq \theta^*} d\mathbf{F}(\theta)$. The second term is the marginal utility of consumption ($\left[\beta E_t \frac{P_t}{P_{t+1} W_{t+1}}\right] \frac{\theta_t}{\theta_t^*} = \frac{\theta_t}{x_t}$) in the case of a high urge to consume ($\theta > \theta^*$), which has probability $\int_{\theta \geq \theta^*} dF(\theta)$. The optimal cutoff θ_t^* (or real wealth x_t) is chosen so that the marginal cost of holding money equals the expected marginal gains.

Hence, the rate of return to money is the inflation-adjusted real interest rate ($\beta \frac{P_t}{P_{t+1}}$) compounded by a liquidity premium $R(\theta^*)$. Notice that $R(\theta_t^*) > 1$, which implies that the option value of one dollar exceeds 1 because as inventories it provides liquidity in the case of a high consumption demand. This is why money has positive value in equilibrium despite the fact that its real rate of return is negative ($\beta \frac{P_t}{P_{t+1}} < 1$) or dominated by interest-bearing assets.

The optimal level of total cash reserve (x_t) is chosen such that the probability of running out of cash is strictly positive ($1 - \mathbf{F}(\theta_t^*) \in (0, 1)$) unless the real cost of holding money is zero (i.e., at the Friedman rule). Namely, the optimal cutoff θ_t^* and real wealth x_t are chosen simultaneously (as they are two sides of the same coin) so that $0 < \Pr[\theta > \theta_t^*] < 1$. This inventory-theoretic formula of money demand is akin to that derived by Wen (2011) in a optimal target inventory model based on the stockout-avoidance motive. Also note that aggregate shocks (if they exist) will affect the distribution of money holdings across households by affecting the cutoff $\theta^*(H^t)$.

Because θ_t^* is independent of h^t, the cutoff provides a sufficient statistic for the distribution

of money demand in the economy. This property facilitates aggregation and makes the model analytically tractable. Consequently, numerical solution methods (such as the method of Krusell and Smith, 1998) are not needed to solve the model's general equilibrium and aggregate dynamics.

By equation (10), real wealth is also independent of h^t. The intuition for x_t being independent of individual history is that (i) it is determined before the realization of θ_t, and all households face the same distribution of idiosyncratic shocks when making labor supply decisions; and (ii) the quasi-linear preference structure implies that labor supply can be adjusted elastically to meet any target level of real wealth ex ante. Hence, in the beginning of each period agents opt to adjust labor income so that the target wealth x_t (or the probability of a binding liquidity constraint $1-\mathbf{F}(\theta^*)$) maximizes expected utility. As a result, x_t is the same across all households regardless of their individual history and initial real balances. This result is reminiscent of the Lagos-Wright (2005) model where the distribution of cash balances is degenerate. However, here the distribution of cash holdings ($\frac{m}{P}$) is not degenerate even though x is degenerate.

2.2 Equilibrium Analysis

Aggregation. Given the sequences of $\{W_t, \tau_t\}$ and the initial distribution of m_0, using capital letters to denote aggregate variables (i.e., $C_t \equiv \int c(h^t, H^t) d\mathbf{F}$), we can integrate individual decision rules by the law of large numbers. The resulting system of equations that determines the competitive equilibrium path of $\{C_t, M_{t+1}, N_t, X_t, \theta_t^*, P_t\}_{t=0}^{\infty}$ includes

$$\frac{1}{W_t} = \beta E_t \frac{1}{W_{t+1}} \frac{P_t}{P_{t+1}} R\left(\theta_t^*\right) \tag{13}$$

$$C_t = D(\theta_t^*) X_t \tag{14}$$

$$\frac{M_{t+1}}{P_t} = H(\theta_t^*) X_t \tag{15}$$

$$X_t = W_t \theta_t^* R(\theta_t^*) \tag{16}$$

$$N_t = \frac{1}{W_t}\left(X_t - \frac{M_t + \tau_t}{P_t}\right) \tag{17}$$

$$M_{t+1} = M_t + \tau_t, \tag{18}$$

where $D(\theta^*) \equiv \int \min\left\{1, \frac{\theta}{\theta^*}\right\} d\mathbf{F}$, $H(\theta^*) \equiv \int \max\left\{0, \frac{\theta^* - \theta_t}{\theta^*}\right\} d\mathbf{F}$, and these two functions satisfy $D(\theta^*) + H(\theta^*) = 1$. Equation (18) is the money market clearing equation. These six dy-

namic equations plus standard transversality conditions uniquely solve for the equilibrium path of $\{C_t, M_{t+1}, N_t, X_t, \theta_t^*, P_t\}_{t=0}^{\infty}$, given the initial distribution of money demand.[13]

The Quantity Theory. The aggregate relationship between consumption (equation 14) and money demand (equation 15) implies the "quantity" equation,

$$P_t C_t = M_{t+1} V_t, \tag{19}$$

where $V_t \equiv \frac{D(\theta_t^*)}{H(\theta_t^*)}$ measures the aggregate consumption velocity of money. A high velocity implies a low demand for real balances relative to consumption. Given the support of θ as $[\theta_L, \theta_H]$ and the mean of θ as $E\theta = \bar{\theta}$, by the definition for the functions D and H, it is easy to see that the domain of velocity is $\left[\frac{\bar{\theta}}{\theta_H - \bar{\theta}}, \infty\right]$, which is bounded below by zero but has no finite upper bound, in sharp contrast to CIA models where velocity is typically a constant of 1.[14] A zero velocity means a liquidity trap (excessive money hoarding), and an infinite velocity means that either the value of money ($\frac{1}{P}$) is zero or nominal money demand (M) is zero. This property explains why the model can generate enough variability in velocity to match the data.

Steady-State Analysis. Assume that money supply follows a constant growth path with

$$\tau_t = \mu M_t, \tag{20}$$

where μ is the growth rate.[15] A steady state is defined as the situation without aggregate uncertainty and with time-invariant distributions of individual variables. Hence, in a steady state all real aggregate variables are constant over time, although the individual variables may be stochastic due to the iid shocks θ_t. Equation (13) implies that the steady-state cutoff θ^* is constant and determined by the relation

$$1 = \frac{\beta}{1+\pi} R(\theta^*), \tag{21}$$

where $\pi \equiv \frac{P_t - P_{t-1}}{P_{t-1}}$ denotes the inflation rate. Hence, the cutoff θ^* is constant for a given level of inflation. The quantity relation (19) implies $\frac{P_t}{P_{t-1}} = \frac{M_{t+1}}{M_t} = 1 + \mu$ in the steady state, so the steady-state inflation rate is the same as the growth rate of money.

[13] The value functions are given by

$$V\left(\frac{m_t}{P_t}\right) = V_0 + \frac{1}{W_t}\frac{m_t}{P_t}; \quad J(x_t, \theta_t) = J_0 + \begin{cases} \left[\beta E_t \frac{P_t}{P_{t+1} W_{t+1}}\right] x_t & \text{if } \theta_t \le \theta_t^* \\ \theta_t \log x_t & \text{if } \theta_t > \theta_t^* \end{cases}.$$

Note that under preference shocks and with log utility function, the demand for real balances ($\frac{m}{P}$) is bounded from above for any positive value of P. So the value function V is also bounded. Also, if one is interested only in the dynamics near the steady state, the uniqueness of equilibrium can also be easily proven (checked) by the eigenvalue method.

[14] For the analysis of heterogeneous-agent CIA models where the CIA constraint on a household binds only with positive probability, see Wen (2010b).

[15] See Wen (2009) for dynamic analysis under monetary shocks.

Since by equation (21) the return to liquidity R must increase with π, the cutoff θ^* must decrease with π (because $\frac{\partial R(\theta^*)}{\partial \theta^*} < 0$); therefore, $\frac{\partial \theta^*}{\partial \pi} < 0$. This means that when inflation rises, the required rate of return to liquidity must also increase accordingly to induce people to hold money. However, because the cost of holding money increases with π, agents opt to hold less money so that the probability of stockout $(1 - \mathbf{F}(\theta^*))$ rises, which reinforces a rise in the liquidity premium (i.e., $\frac{\partial^2 R}{\partial \theta^* \partial \pi} < 0$). Also, since the target wealth is given by $x(\theta^*) = W\theta^* R(\theta^*)$, we have $\frac{\partial x}{\partial \theta^*} = W\mathbf{F}(\theta^*) > 0$, so the target wealth decreases with π. Therefore, a higher rate of inflation has two types of effects on welfare: The intensive margin and the extensive margin. On the intensive margin, $\frac{\partial x}{\partial \pi} < 0$, so higher inflation leads to lower consumption through a negative wealth effect for all agents. In addition, liquidity-constrained agents suffer disproportionately more because they (i) do not have self-insurance ($m_{t+1} = 0$) to buffer shocks and (ii) still face the same variance of idiosyncratic shocks (σ_θ^2) when having a lower wealth level. This second aspect of the intensive margin is emphasized by Imrohoroglu (1992) and Akyol (2004). On the extensive margin, $\frac{\partial \theta^*}{\partial \pi} < 0$; thus, a high inflation rate means that a larger portion of the population will become liquidity constrained and subject to idiosyncratic shocks without the buffer stock. This extensive margin will be shown to be an important force in affecting social welfare but it has not been fully appreciated by the existing literature.

Under the Friedman rule, $1 + \pi = \beta$, we must have $R = 1$ and $\theta^* = \theta_H$ according to equation (12). Consequently, $D(\theta^*) = \frac{\bar{\theta}}{\theta_H}$ and $H(\theta^*) = 1 - \frac{\bar{\theta}}{\theta_H}$. Hence, we have $x(H^t) = W\theta_H$ and

$$c_t = \min\left\{1, \frac{\theta_t}{\theta_H}\right\} x = \theta_t W. \tag{22}$$

That is, individual consumption is perfectly adjusted ("smoothed") based on preference shocks under the Friedman rule, suggesting perfect self-insurance or the first-best allocation. The probability of being liquidity constrained (running out of cash) is zero in this case because households opt to hold the maximum amount of money when the cost of doing so is zero: $\frac{m_{t+1}}{P_t} = (\theta_H - \theta_t) W > 0$ for all θ.

However, since θ^* is bounded below by θ_L, the liquidity premium is then bounded above by $R(\theta_L) = \frac{\bar{\theta}}{\theta_L}$. This means there must exist a maximum rate of inflation $\pi_{\max}$ such that equation (21) holds: $\frac{\bar{\theta}}{\theta_L} = \frac{1+\pi_{\max}}{\beta}$. At this maximum inflation rate,

$$\pi_{\max} = \beta \frac{\bar{\theta}}{\theta_L} - 1, \tag{23}$$

we have $D(\theta_L) = 1$ and $H(\theta_L) = 0$. That is, the optimal demand for real balances from all

households goes to zero, $m\left(h^t, H^t\right) = 0$, if $\pi \geq \pi_{\max}$.[16]

When the cost of holding money is sufficiently high, agents opt not to use money as the store of value and the velocity becomes infinity: $V = \frac{D(\theta_L)}{H(\theta_L)} = \infty$. The velocity is a decreasing function of inflation because money demand drops faster than consumption as the inflation tax rises: $\frac{\partial V}{\partial \theta^*}\frac{\partial \theta^*}{\partial \pi} > 0$. This long-run implication is consistent with empirical data. For example, Chiu (2007) has found using cross-country data that countries with higher average inflation also tend to have significantly higher levels of velocity and argued that such an implication cannot be deduced from the Baumol-Tobin model with an exogenously segmented asset market.[17]

When money is no longer held as a store of value because of sufficiently high inflation ($\pi \geq \pi_{\max}$), it must be true that $c_t = \min\left\{1, \frac{\theta_t}{\theta_L}\right\} x = x = \bar{\theta} W$, so consumption is constant and completely unresponsive to preference shocks, suggesting the worst possible allocation with a significantly lower level of welfare than the case with the Friedman rule. Note that the aggregate (average) consumption under hyper inflation $\pi \geq \pi_{\max}$ is identical to the aggregate (average) consumption under the Friedman rule by equation (22): $C = W \int \theta_t d\mathbf{F} = \bar{\theta} W$. This equivalence of aggregate allocations under drastically different inflation rates reveals the danger of measuring the welfare cost of inflation (or the welfare cost of the business cycle) based on representative-agent models. This point is quantified in the next section.

3 welfare cost of Inflation

3.1 Measures of welfare cost

We measure the welfare cost of inflation as the percentage increase ($\Delta\%$) in household consumption that would make any individual household indifferent in terms of expected lifetime utilities between living in a high-inflation regime (π) and the Friedman-rule inflation regime ($\pi^o \equiv \beta - 1$). That is, Δ is the compensation required for household consumption as the inflation rate increases above the Friedman rule. By the law of large numbers, the expected momentary utility of an individual i is the same as the aggregate (or average) utility of the population (with equal weights); namely, $\int \theta(i) \log c(i) d\mathbf{F}(\theta) - \int n(i)\, d\mathbf{F}(\theta) = \int \theta(i) \log c(i) di - N$. Thus, our welfare measure also corresponds to a social planner's measure with equal welfare weights.

[16] Obviously we need to assume that $\beta > \frac{\theta_L}{\bar{\theta}}$ to support a monetary equilibrium even at low inflation rates. If β is too small, there does not exist a high enough liquidity premium to induce people to hold money. As an example, suppose θ follows the Pareto distribution, $F(\theta) = 1 - \theta^{-\sigma}$ with support $(1, \infty)$, then $\theta_L = 1$ and $\bar{\theta} = \infty$, so the lower bound on β to support a monetary equilibrium is zero.

[17] Based on a long sample of the U.S. time-series data, Lucas (2000) also shows that the inverse of the velocity is negatively related to inflation.

Hence, the welfare cost of inflation Δ solves

$$\sum_{t=0}^{\infty}\beta^{t}\left\{\int\left[\theta\log\left(1+\Delta\right)c\left(\pi\right)\right]d\mathbf{F}-N\left(\pi\right)\right\}=\sum_{t=0}^{\infty}\beta^{t}\left\{\int\theta\log c\left(\pi^{o}\right)d\mathbf{F}-N\left(\pi^{o}\right)\right\}. \tag{24}$$

By the household consumption policy, $c=\min\left\{1,\frac{\theta}{\theta^{*}}\right\}x$, and the property that $x=X$ is independent of individual history, the above equation implies

$$\Delta=\exp\left\{\frac{1}{\bar{\theta}}\left[N\left(\pi\right)-N\left(\pi^{o}\right)+\bar{\theta}\log\frac{X\left(\pi^{o}\right)}{X\left(\pi\right)}+J_{\theta}\left(\pi^{o}\right)-J_{\theta}\left(\pi\right)\right]\right\}-1, \tag{25}$$

where $J_{\theta}\left(\pi\right)\equiv\int_{\theta\leq\theta^{*}(\pi)}\left(\theta\log\frac{\theta}{\theta^{*}(\pi)}\right)d\mathbf{F}$ captures the effect of idiosyncratic risk (or the heterogeneity of consumption) on social welfare.

A perfect decomposition of the welfare cost into the intensive and extensive margins is difficult because they are both endogenous and mutually reinforce each other. But one way to show that the high welfare cost of inflation found in our model comes mainly from the loss of the buffer-stock value of money for the increased cash-poor population to self-insure against idiosyncratic risks is to compare equation (25) with an alternative (pseudo) measure ($\Delta^{\times}$) based on the lifetime utility of the average consumption (C) of a household across different states (i.e., $\log C=\log D\left(\theta^{*}\right)X$). This pseudo measure solves

$$\sum_{t=0}^{\infty}\beta^{t}\left\{\int\left[\theta\log\left(1+\Delta^{\times}\right)C\left(\pi\right)\right]d\mathbf{F}-N\left(\pi\right)\right\}=\sum_{t=0}^{\infty}\beta^{t}\left\{\int\theta\log C\left(\pi^{o}\right)d\mathbf{F}-N\left(\pi^{o}\right)\right\}, \tag{26}$$

which implies

$$\Delta^{\times}=\exp\left\{\frac{1}{\bar{\theta}}\left[N\left(\pi\right)-N\left(\pi^{o}\right)+\bar{\theta}\log\frac{C\left(\pi^{o}\right)}{C\left(\pi\right)}\right]\right\}-1. \tag{27}$$

Clearly, the utility of the average consumption C, as opposed to the average utility of individual consumption $c\left(\theta_{t}\right)$, captures mainly the impact of inflation on welfare along the intensive margin because only the average fall in consumption (under inflation) across all households is reflected in the pseudo measure $\Delta^{\times}$. So by design the individual consumption risk (or the welfare loss of the uninsured fraction of the population without buffer stock in bad states) is omitted (averaged out) by this alternative metric $\Delta^{\times}$.

Note that $\Delta^{\times}$ also pertains to the welfare measure in a representative-agent model. So the pseudo welfare measure can also shed light on the low welfare cost of inflation found in representative-agent models.

As discussed previously, when the inflation rate is either sufficiently high with $\pi \geq \pi_{\max}$ or sufficiently low with $\pi = \pi^o = \beta - 1$, we have identical aggregate allocations between the hyper-inflation regime (where people no longer hold money) and the Friedman-rule inflation regime (where people hold infinite real money balances), with $N(\pi_{\max}) = N(\pi^o)$ and $C(\pi_{\max}) = C(\pi^o)$. Consequently, the welfare cost of rising the inflation rate from π^o to $\pi_{\max}$ is zero under the incorrect measure $\Delta^{\times}$, whereas the correctly measured welfare cost of inflation (that properly takes into account of the risk along the extensive margin) is given by $\Delta = \exp\left\{\frac{1}{\theta}\left[J_\theta(\pi^o) - J_\theta(\pi_{\max})\right]\right\} - 1$, which is strictly positive because $J(\pi_{\max}) = \int_{\theta \leq \theta_L} \left(\theta \log \frac{\theta}{\theta_L}\right) d\mathbf{F} = 0$ and $J(\pi^o) = \int_{\theta \leq \theta_H} \left(\theta \log \frac{\theta}{\theta_H}\right) d\mathbf{F} > 0$. This difference is striking. It suggests just how wrong representative-agent models can be when it comes to welfare implications.

3.2 Calibrations and Predictions

The time period t is a year.[18] We set the time discount factor $\beta = 0.95$. To facilitate calibration, we assume that the idiosyncratic shock θ follows the Pareto distribution,

$$F(\theta) = 1 - \theta^{-\sigma}, \tag{28}$$

with $\sigma > 1$ and the support $[1, \infty)$. As a benchmark, we set the shape parameter $\sigma = 2.65$ to roughly match the household consumption risk (variance) in the U.S. data (see Section 3.2.2). The mean of this distribution is $\bar{\theta} = \frac{\sigma}{\sigma - 1}$. Since θ is not bounded from above, at the Friedman rule money demand goes to infinity. In the following analysis, we assume that $\pi^o = \beta - 1 + \varepsilon$, where the positive number ε is arbitrarily close to 0 but not exactly equal to 0.[19] We refer to this inflation rate π^o as the "Friedman rule". With the Pareto distribution function, we have $R(\theta^*) = 1 + \frac{1}{\sigma - 1}\theta^{*-\sigma}$, $H(\theta^*) = R(\theta^*) - \frac{\sigma}{\sigma - 1}\frac{1}{\theta^*}$, and $D(\theta^*) = 1 - H(\theta^*)$. The cutoff $\theta^*(\pi)$ can then be solved explicitly using the relation $R(\theta^*) = \frac{1+\pi}{\beta}$.

3.2.1 A Special Case without Capital

Notice that we have not explicitly modeled firm behavior so far. But to illustrate the power of our framework, assume for simplicity a representative firm with a linear production technology, $Y = N$. The competitive real wage is then a constant $W = 1$. So the steady-state aggregate allocations are given by

$$C = \theta^* R(\theta^*) D(\theta^*), \qquad X = \theta^* R(\theta^*), \qquad N = \theta^* R(\theta^*) D(\theta^*), \tag{29}$$

[18] We choose t to be a year because the aggregate money demand data used by Lucas (2000) are available only at the annual frequency.

[19] More specifically, we set $\varepsilon = 10^{-6}$.

where the cutoff is determined by the inflation rate, $R(\theta^*) = \frac{1+\pi}{\beta}$. Thus, inflation affects aggregate allocations only through affecting the cutoff (distribution) θ^*.

The welfare costs of inflation are graphed in the top panels in Figure 1 (solid red lines). The top-left panel (solid red line) shows the correct measure of welfare cost (Δ). It is monotonically increasing with inflation. Hence, the Friedman rule is clearly optimal. The maximum welfare cost is reached at the maximum inflation rate $\pi_{\max} = 52.576\%$, in which case $\Delta \approx 14\%$ of consumption. Beyond this inflation rate money is no longer valued (held) by households, so the welfare cost of inflation remains constant at about 14% for $\pi \geq \pi_{\max}$. When the inflation rate $\pi = 10\%$ (i.e., π increases from 0% to 10%), the welfare cost is 3.94% of annual consumption, a big number. The cost would be 8.9% of consumption if measured relative to the Friedman rule because of the highly concave feature of the welfare cost function $\Delta(\pi)$.

[Insert Figure 1]

In contrast, the top-right panel of Figure 1 (solid red line) shows the pseudo measure of welfare cost ($\Delta^\times$) based on the average consumption of a household (or a representative agent). This measure captures only the intensive margin and is not monotonic; it equals 0 at two extreme points—the point of the Friedman rule and the point where $\pi = \pi_{\max}$. In the first case, individual consumption level ($c_t = \theta_t W$) is the first-best—because it is costless to hold money, so agents are perfectly insured against idiosyncratic risk. The average consumption in this case is $\bar{\theta} W$. In the latter case, individuals' consumption levels become homogeneous across households at a constant level ($c_t = \bar{\theta} W$) when money is no longer held as a store of value and agents become completely uninsured. Without money, inflation no longer has any adverse liquidity effects on consumption, so $\Delta^\times$ remains at zero for $\pi \geq \pi_{\max}$. This is in sharp contrast to CIA models where agents are forced to hold money regardless of the inflation rate. The maximum cost of inflation by the pseudo measure is $\Delta^\times = 0.065\%$ of consumption at $\pi = 2.88\%$ (relative to the Friedman rule), a trivially small number. Under a 10% inflation rate, the incorrectly measured welfare cost is even smaller. This sharp contrast between Δ and $\Delta^\times$ reveals that the bulk of the welfare costs (more than 95%) comes from the extensive margin due to the inflation-induced loss of self-insurance for the cash-poor households.

3.2.2 General Equilibrium with Capital

This subsection introduces capital accumulation into our benchmark model. This extension serves at least three purposes: (i) to facilitate calibrations, as any serious calibration requires a general-equilibrium model with capital in the production and aggregate investment in the aggregate resource

constraint; (ii) to facilitate comparisons with the existing general-equilibrium literature on welfare analysis; and (iii) to investigate the sensitivity of welfare costs to capital accumulation.

Firms. A representative firm produces final output according to the production technology, $Y_t = AK_t^{\alpha}N_t^{1-\alpha}$. The firm rents capital and hires labor from households. Perfect competition implies that factor prices equal their marginal products: $W_t = (1-\alpha)\frac{Y_t}{N_t}$ and $r_t + \delta = \alpha\frac{Y_t}{K_t}$.

Households. Households accumulate illiquid capital asset through saving (s_t) and rent the capital stock to firms in a competitive rental market with the rental rate denoted by r_t. The illiquidity of the capital asset is captured by the assumption that saving decisions for fixed capital (s_{t+1}) in period t must be made in tandem with the decision of labor supply (n_t) in the first subperiod (i.e., before the idiosyncratic preference shocks are realized). The household budget constraint becomes

$$c_t + \frac{m_{t+1}}{P_t} + s_{t+1} \leq (1+r_t)\, s_t + \frac{m_t + \tau_t}{P_t} + W_t n_t, \tag{30}$$

where $s_{t+1} = s_{t+1}(h_-^t, H^t)$ in accordance with the notations in equation (1). Compared with the previous benchmark model, there is now one more first-order condition for s_{t+1} on the household side:[20]

$$\frac{1}{W_t} = \beta E_t \frac{1+r_{t+1}}{W_{t+1}}, \tag{31}$$

which in the steady state becomes

$$1 = \beta\,(1+r)\,. \tag{32}$$

Compared with equation (21), it must be true that $(1+r)\,(1+\pi) = R\,(\theta^*)$ under the no-arbitrage condition, suggesting that the nominal rate of return to the illiquid capital asset must equal (and be dictated by) the liquidity premium of money R. In addition, the model has the interesting property that household saving s_{t+1} can be made independent of the idiosyncratic preference shock and individual history.[21]

General Equilibrium. Define net real wealth as $x_t \equiv \frac{m_t+\tau_t}{P_t} + W_t n_t + (1+r_t)\, s_t - s_{t+1}$. The household decision rules of consumption, real money balances, and net wealth are identical in functional form to equations (8)-(12) except that the labor supply changes to $n_t = \frac{1}{W_t}[x_t - \frac{m_t+\tau_t}{P_t} + s_{t+1} - (1+r_t)\, s_t]$. That is, adding capital into the benchmark model does not change its basic properties, such as the fact that $\{x_t, \theta_t^*\}$ are independent of individual history h^t. In general equilibrium,

[20] See Appendix A2.

[21] To see this, simply introduce an arbitrary quadratic adjustment cost term ζs_{t+1}^2 ($\zeta \geq 0$) into the household budget constraint, then equation (31) becomes $\frac{(\zeta s_{t+1}+1)}{W_t} = \beta E_t \frac{1+r_{t+1}}{W_{t+1}}$, which implies that household saving decisions depend only on the aggregate variables. This property holds true in the limit $\zeta \to 0$. This in turn implies that household labor supply n_t is the only margin that fully adjusts to meet the target wealth x_t each period given the initial amount of cash in hand m_t.

the aggregate supply of capital equals the aggregate demand of capital, $\int s_{t+1}\left(h_{-}^{t}, H^{t}\right) d\mathbf{F} = K_{t+1}$, and the aggregate supply of labor equals the aggregate demand of labor, $\int n_{t}\left(h_{-}^{t}, H^{t}\right) d\mathbf{F} = N_{t}$.

Using upper-case letters to denote aggregate quantities, a general equilibrium is defined as the sequence $\{C_t, Y_t, N_t, K_{t+1}, M_{t+1}, P_t, W_t, r_t, \theta_t^*\}_{t=0}^{\infty}$, such that given prices $\{P_t, W_t, r_t, \}$ and monetary policies, (i) all households maximize utilities subject to their resource and borrowing constraints (and initial capital and money holdings), (ii) firms maximize profits, (iii) all markets clear, (iv) the law of large numbers holds, and (v) the set of standard transversality conditions is satisfied.[22] Because the steady state is unique and the system is saddle stable, the distribution of money demand converges to a unique time-invariant distribution in the long run for any initial distributions of capital s_0 and money holdings m_0.

Steady-State Allocations. Proposition 2 in Appendix A2 presents the full set of equations needed to solve for the dynamic path of the general equilibrium. In the steady state, the capital-to-output and consumption-to-output ratios are given by $\frac{K}{Y} = \frac{\beta\alpha}{1-\beta(1-\delta)}$ and $\frac{C}{Y} = 1 - \frac{\delta\beta\alpha}{1-\beta(1-\delta)}$, respectively, which are the same as in standard RBC models without money. Since $r + \delta = \alpha\frac{Y}{K}$ and $w = (1-\alpha)\frac{Y}{N}$, the factor prices are given by $r = \frac{1}{\beta} - 1$ and $W = (1-\alpha)\left(\frac{\beta\alpha}{1-\beta(1-\delta)}\right)^{\frac{\alpha}{1-\alpha}}$, respectively. Hence, the existence of money in this model does not alter the steady-state aggregate saving rate, the great ratios, and the real factor prices in the neoclassical growth model, in contrast to typical CIA models. However, the levels of aggregate income, consumption, employment, and capital stock are affected by money growth. These levels are given by

$$C = W\theta^* R(\theta^*) D(\theta^*), \qquad Y = \frac{1-\beta(1-\delta)}{1-\beta(1-\delta)-\delta\beta\alpha} C, \qquad K = \frac{\beta\alpha}{1-\beta(1-\delta)} Y, \qquad N = \frac{1-\alpha}{W} Y, \tag{33}$$

where the cutoff is determined solely by the inflation rate as in the benchmark model without capital, $R(\theta^*) = \frac{1+\pi}{\beta}$.

Calibration of Key Parameters. The remaining free parameters include σ in the Pareto distribution function and A in the quantity relationship:

$$\frac{M}{PY} = A\frac{H(\theta^*)}{D(\theta^*)}, \tag{34}$$

where P denotes aggregate price level, Y aggregate output, M aggregate money supply, $\frac{PY}{M}$ is the empirical measure of the income velocity of money, and $A\frac{H(\theta^*)}{D(\theta^*)}$ is the theoretical counterpart of

[22] Such transversality conditions include $\lim_{t\to\infty} \beta^t E_t \frac{K_{t+1}}{W_t} = 0$ and $\lim_{t\to\infty} \beta^t E_t \frac{M_{t+1}}{P_t W_t} = 0$, where $\frac{1}{W}$ is the shadow value of capital and $\frac{1}{P}$ is the value of money.

income velocity implied by our model (where the scale parameter A is a function of the consumption-to-income ratio). Since (i) the definition of money in the empirical data varies greatly (such as M1, M2, and M3), and (ii) the definition of the consumption-to-output ratio changes depending on whether capital investment, government spending, net exports, and durable goods consumption are included in GDP, the measured velocity of money also varies accordingly. Hence, we introduce the scaling parameter A to reflect these variations in the measurement bias in the mean of velocity when calibrating our model to match the data.

These two free parameters $\{\sigma, A\}$ are calibrated by three independent methods, called Method 1, Method 2, and Method 3. Method 1 is our benchmark and the other two serve as robustness checks because Method 1 does not target the idiosyncratic risk facing consumers. Nonetheless, all three methods imply a consumption variance at the household level consistent with the U.S. data.

Under Method 1, we set $A = 1$ and use a least squares criterion to estimate the value of σ that enables our model to best match the empirical aggregate money demand curve ($\frac{M}{PY}$) suggested by Lucas (2000). This is also the calibration strategy of Lagos and Wright (2005). Under Method 2, we choose the values of $\{A, \sigma\}$ to jointly match the (i) the empirical money demand curve of Lucas (2000) and (ii) the probability of running out of cash (the likelihood of a binding liquidity constraint $\Pr[\theta > \theta^*]$) implied by the household survey data. Under Method 3, we choose the values of $\{A, \sigma\}$ to jointly match (i) the empirical money demand curve and (ii) the household consumption volatility implied by household data. Appendix B provides details of these calibration procedures.[23]

The calibrated parameter values are summarized in Table 1. Notice that the values of σ under various calibration methods imply that the variance of log consumption ($\log c_t$) in the model is in the range of $0.03 \sim 0.13$ (see the last column in Table 1). This range of household consumption volatility is consistent with the empirical estimates of Telyukova (2011, Table 9), who reports a range of $0.056 \sim 0.113$ for the variance of various types of household consumption. Under Method 1, the model-implied variance of household consumption is 0.055, which is on the lower bound of Telyukova's estimates and hence more conservative than the other two calibration methods. Thus, we use Method 1 as our benchmark calibration for $\{\sigma, A\}$ in this paper.[24]

[23] Because of the lack of long time-series panel data that can track the consumption expenditure and money demand of the same households for more than one year, we borrow information from cross-section data to infer consumption volatility. This is not entirely unreasonable. For example, if we survey households from the same villages with similar living standards and consumption needs, then cross-section variations may very well indicate over-time consumption risk of a typical household in the village.

[24] Telyukova's (2011) estimates are based on monthly data. However, she also reported similar estimates for the variance of household consumption based on quarterly data in an earlier 2009 version of her paper. Annual data are not available since the SCF data keep tract of the same households for only one year. Following Telyukova, we compute in our model the variance of the logarithm of consumption, $\log c_t = \log\left[\min\left\{1, \frac{\theta_t}{\theta^*}\right\} x\right]$, based on simulated sample of θ_t with a sample size of 10^6. Keeping the variance of preference shocks constant, the model would generate larger welfare costs of inflation if the time interval becomes shorter. Hence, using an annual model is conservative because calibrating our model at a quarterly or monthly interval would only enhance our results.

[Insert Table 1]

Therefore, all the calibration methods amount to rationalizing the empirical money demand curve emphasized by Lucas (2000) and various measures of consumption risks. Using historical data for GDP, money stock (M1), and the nominal interest rate, Lucas (2000) showed that the ratio of M1 to nominal GDP is downward sloping against the nominal interest rate. Lucas interpreted this downward relationship as a "money demand" curve and argued that it can be best rationalized by the Sidrauski (1967) model of money-in-utility (MIU). Lucas estimated that the empirical money demand curve can be best captured by an ad hoc power function of the form

$$\frac{M}{PY} = Ar^{-\eta}, \tag{35}$$

where A is a scale parameter, r the nominal interest rate, and η the interest elasticity of money demand. He showed that $\eta = 0.5$ gives the best fit. Because the money demand defined by Lucas is identical to the inverted velocity, a downward-sloping money demand curve is the same as an upward-sloping velocity curve (namely, velocity is positively related to the nominal interest rate or inflation). Similar to Lucas, the money demand curve implied by our model takes the form in equation (34). Figure 2 shows the fit of the theoretical model to the U.S. data under calibration Method 1.[25]

[Insert Figure 2]

The model's welfare implications under Calibration Method 1 are graphed in Figure 1 (dashed blue lines). The top-left panel (dashed blue line) shows that the welfare cost function (Δ) is identical to that in the benchmark model without capital (solid red line) either toward π^o or toward $\pi_{\max}$, but lies slight above in between. When the inflation rate increases from the Friedman rule (π^o) to $\pi = 10\%$, the welfare cost is 9.6% of annual consumption, as opposed to 8.9% in the benchmark model without capital. When inflation increases from 0% to 10%, the welfare cost is 3.89% of annual consumption, which is slightly smaller than before because the slope of the welfare function is now flatter in the range of moderate inflation. The cost would be even higher if we calibrate σ to match the variance of consumption in developing countries.[26]

[25] The circles in Figure 1 show plots of annual time series of a short-term nominal interest rate (the commercial paper rate) against the ratio of M1 to nominal GDP, for the United States for the period 1900–1997, the data sample used by Lucas (2000). The data are downloaded from the online Historical Statistics of the United States–Millennium Edition. The solid line with crosses is the model's prediction.

[26] Wen (2011b) in the Appendix uses statistics based on medical spending, traffic incidents, and work-related injuries to argue that consumption risk in China is at least one order of magnitude larger than that in the U.S. For the sake of argument, suppose the variance of household consumption in China is twice of that in the U.S., then the implied welfare cost of 10% inflation (compared to 0% inflation) would be more than 6 percent of consumption.

In contrast, the top-right panel of Figure 1 (dashed blue line) shows the pseudo measure of welfare cost ($\Delta^{\times}$) based on the average consumption of a household (or a representative agent). This pseudo measure is now significantly higher than the counterpart in the benchmark model without capital (solid red line), but the absolute magnitude is still small, more than a order of magnitude smaller than the correct measure shown in the left panel. The maximum cost of inflation by the incorrect measure is $\Delta^{\times} = 0.78\%$ of consumption when $\pi = 2.7\%$, relative to the Friedman rule. It becomes even smaller under a 10% inflation rate. These values are similar in magnitudes to those obtained in the existing literature based on representative-agent models (e.g., Lucas, 2000; Cooley and Hansen, 1989). Similar to the benchmark model without capital, the bulk of the welfare costs (more than 85%) comes from the extensive margin.

The bottom-left panel of Figure 1 shows the level of aggregate money demand ($\frac{M}{P}$), and the bottom-right panel shows the consumption velocity of money ($\frac{D(\theta^*)}{H(\theta^*)}$), where the solid red line pertains to the benchmark model without capital and the dashed blue line to the model with capital. The money demand function is convex and decreases monotonically with inflation, whereas the velocity is also convex and increases monotonically with inflation. Near the Friedman rule, the demand for money is close to infinity and the velocity is close to zero. In contrast, the demand for real balances becomes zero for $\pi \geq \pi_{\max}$ and the velocity of money becomes infinity at $\pi_{\max}$.[27] The velocity of money is identical in the two models (regardless of capital) because it is independent of the real wage. The velocity is close to zero near the Friedman rule because households opt to hoard as much money as they can when its real rate of return equals the inverse of the time discount factor (i.e., the demand for money approaches infinity as the opportunity cost of holding money goes to zero). In this case, the aggregate price level is close to zero and the borrowing constraint ceases to bind for all households (or across all possible states). The velocity of money becomes infinity as $\pi \to \pi_{\max}$ because people want to divest their holdings of money as quickly as possible to avoid the inflation tax despite the need for a store of value to self-insure against idiosyncratic shocks. But since the cost of holding money is so high as $\frac{1}{P_t} \to \infty$ and the insurance value of money is so low (fully destroyed), the demand for money becomes zero. This pertains to the "hot potato effect" of inflation found in hyper-inflation countries where people try to get rid of money as fast as they can to avoid further destruction of the liquidity value of money in hand.

These implications for money demand and velocity are quite different from standard CIA models, which imply a constant velocity and a strictly positive lower bound on money demand, because agents under the CIA constraint must hold money even with an infinite inflation rate at $\pi = \infty$. In the real world, people often stop accepting domestic currency as a store of value or means of pay-

[27] The graph shows the velocity only for $\pi < \pi_{\max}$.

ment when the inflation rate is too high (but before reaching infinity), consistent with our model's prediction.

Under Calibration Method 2, the implied welfare costs of 10% inflation range from 3.76% to 4.95%, depending on whether π is increased from 4% to 14% or from 0% to 10%, whereas the implied variance of household consumption is in the rage of $0.03 \sim 0.11$. Under Method 3, the implied welfare costs of 10% inflation range from 4.94% to 6.13%, depending on whether π is increased from 4% to 14% or from 0% to 10%. The implied variance of household consumption is about $0.07 \sim 0.13$, roughly consistent with independent empirical work (e.g., Telyukova, 2011, Table 9). These welfare results are summarized in Table 2.

[Insert Table 2]

In a heterogeneous-agent economy with incomplete markets, the larger the variance of idiosyncratic shocks (or smaller value of σ), the stronger the precautionary motive for holding money. This raises the inflation tax at a given inflation rate. More importantly, higher inflation shifts the mass of the distribution of money demand toward zero balances by reducing cash holdings across agents, resulting in a larger portion of the cash-constrained population (or a higher probability of running out of cash for each individual) without self-insurance against idiosyncratic shocks. This shift of the distribution of money demand in response to inflation is most critical in generating the large welfare cost, as evident in the difference between the welfare cost function Δ and the pseudo welfare cost function $\Delta^{\times}$ shown in Figure 1.

To see the shift of the distribution of money demand, note that $\frac{\partial R(\theta^*)}{\partial \pi} > 0$ and $\frac{\partial \theta^*}{\partial \pi} < 0$ by equation (21). The probability of running out of cash is given by $1 - \mathbf{F}(\theta^*) = (\sigma - 1)\left(\frac{1+\pi-\beta}{\beta}\right)$, which increases with inflation. As inflation rises, the portion of the population holding zero cash balances increases rapidly. For example, given the parameter values under Method 1, when inflation increases from 0% to 10%, an additional 17% of the entire population is left without cash (thus without self-insurance), raising the total number of cashless agents to about 26% of the population. When holding money is too costly, the demand for real balances becomes so low that the probability of running out of cash is extremely high. The loss of buffer stocks or self insurance amounts to large welfare costs. A comparison of Δ and $\Delta^{\times}$ in Figure 1 suggests that this extensive margin accounts for more than 90% of the welfare costs under moderate inflation (say for $\pi \geq 5\%$), regardless of capital.

Another way to see the importance of the extensive margin in contributing to the large welfare cost of inflation is to introduce an insurance market to the model so that cash-poor agents can borrow from cash-rich agents to mitigate the adverse liquidity effect of inflation. This analysis is carried out in the next section.

Here we emphasize a point made by Lagos and Wright (2005). That is, we notice that our three different calibration methods (with quite different ranges of parameter values) can all match the empirical money demand curve in Figure 2 almost equally well, but the implied welfare costs are nonetheless quite different. In particular, the value of A is crucial for matching the Lucas money demand curve but plays no role in computing our welfare results (see equation (25)).[28] Hence, as noted by Lagos and Wright (2005), simply computing the area underneath the money demand curve as a measure of the welfare cost of inflation, as proposed by Bailey (1956) and favored by Lucas (2000), is not good enough. What is really needed is an explicit model of the micro foundations, especially the motives behind money demand, in order to properly estimate the welfare cost of inflation. Consistent with this spirit, here we offer a different model of money demand with micro foundations alternative to those of Lagos and Wright, and we obtain different welfare results (as expected) because we emphasize different functions of money in our models (medium of exchange versus store of value). One thing in common between our two approaches is that both models obtain substantially larger welfare costs of inflation than in the existing representative-agent monetary literature, because in both models money is essential yet agents can opt not to hold money if the cost of doing so is sufficiently high. However, in the Lagos-Wright model the main source of welfare cost under inflation is the loss of the medium of exchange and matching efficiency. In our model, the main source of the welfare cost of inflation is the loss of the store of value for self-protection (insurance) against idiosyncratic consumption (or income) risk, which is not captured by the standard search-and-matching monetary models.

4 Welfare Implications with Credit and Banking

There are at least two potential objections to the large welfare cost of inflation in the benchmark model. First, the model posits uninsurable risk and assumes that money is the only liquid asset to help self-insure against such risk. This setup rules out other types of store of value or insurance devices and does not take into account the role of credit and banking (such as consumer credit or credit cards) in mitigating the idiosyncratic risk through borrowing and lending. Second, it is a common belief in the existing literature that inflation benefits debtors by redistributing the burden of inflation toward creditors. For these reasons, the welfare costs of inflation may be overstated.[29]

This section confronts these issues by extending the general-equilibrium Bewley model to a setting with "narrow banking," where cash-rich agents can deposit their idle cash into a community

[28] That is, A affects the estimated value of σ but does not directly enter the welfare function (25).

[29] Kehoe, Levine, and Woodford (1992) study the welfare effects of inflation in a Bewley model with aggregate uncertainty. They show that lump-sum nominal transfers can redistribute wealth from cash-rich agents to cash-poor agents, because the latter receive disproportionately more transfers than the former and thereby benefit from inflation. Consequently, inflation may improve social welfare. However, this positive effect on social welfare is quantitatively quite small and requires extreme parameter values in their model.

bank, and cash-poor agents with a binding liquidity constraint can borrow from the bank by paying a nominal interest. The nominal interest rate of loans can be endogenously determined by the supply and demand of funds. This type of risk-sharing arrangement captures the reality that consumers in developing countries (or poor households in rich countries) can often meet their liquidity demand (in a limited way) by borrowing money from relatives, friends, or local community banks.

More specifically, the key friction in the benchmark model is the nonnegativity constraint ($m \geq 0$) on nominal balances. With this constraint, there is always an ex post inefficiency on money holdings since ex post some agents end up holding idle balances while others end up liquidity constrained with zero balances. This creates a need for risk sharing, as suggested by Lucas (1980). However, without the necessary information- and record-keeping technologies, households cannot lend and borrow among themselves. In this section, we assume that a community bank (or credit union) emerges to resolve the risk-sharing problem by developing the required information technologies. The function of the bank is to accept nominal deposits from households and make nominal loans to bank members. Because of limited contract enforcement, we assume that a household can borrow only up to a limit proportional to its average bank deposits in the past ($\int m_t d\mathbf{F}$) plus an additional fixed amount $\bar{b} \geq 0$.[30]

For simplicity, we assume that (i) all deposits are withdrawn at the end of the same period (100-percent reserve banking) and (ii) all loans are one-period loans that charge the competitive nominal interest rate $1 + \tilde{\imath}$, which is determined by the demand and supply of funds in the community. The nominal interest rate on deposits is denoted by $1 + i^d$, with $i^d = \psi\tilde{\imath}$ and $\psi \in (0, 1)$. Any profits earned by the bank are distributed back to community members as lump-sum transfers.

Similar banking arrangements have been studied recently by Berentsen, Camera, and Waller (2006) and others. This literature shows that financial intermediation improves welfare.[31] However, these authors study the issue in the Lagos-Wright (2005) framework, which focuses on the medium-of-exchange function of money and has no capital accumulation. In addition, in their model the welfare gains of financial intermediation come solely from the payment of interest on deposits and not from relaxing borrowers' liquidity constraints. In contrast, this paper focuses on welfare gains that derive mainly from risk sharing or relaxing borrowers' liquidity constraints.

The time line of events is as follows: In the beginning of each period, each household makes decisions on labor supply and capital investment, taking as given the initial wealth from last period. This is the first subperiod. In the second subperiod, idiosyncratic preference shocks are realized and each household chooses consumption, nominal balances, and the amount of new loans.

[30] The credit card is a similar type of institutional arrangement.

[31] However, Chiu and Meh (2008) show that banking may reduce welfare under moderate inflation rates if there exist transaction costs for using financial intermediation. For alternative approaches to money and banking, see Williamson (1986) and Andolfatto and Nosal (2003).

4.1 The Household Problem

As in the previous model, hours worked and nonmonetary asset investment in each period must be determined before the idiosyncratic preference shock is realized; the remaining decisions are all made after observing θ_t. Each household takes the bank's profit income (T_t) and government money transfers (τ_t) as given and chooses consumption, capital investment, labor supply, money demand, and credit borrowing (b_{t+1}) to maximize the objective function in equation (1) subject to

$$c_t + s_{t+1} + \frac{m_{t+1}}{P_t} + (1+\tilde{\imath}_t)\frac{b_t}{P_t} \leq (1+r_t)\, s_t + \left(1+i_t^d\right)\frac{m_t}{P_t} + \frac{b_{t+1}}{P_t} + W_t n_t + \frac{T_t+\tau_t}{P_t} \tag{36}$$

$$m_{t+1} \geq 0 \tag{37}$$

$$b_{t+1} \geq 0 \tag{38}$$

$$b_{t+1} \leq \gamma \int m_t d\mathbf{F} + \bar{b}, \tag{39}$$

where $\tilde{\imath}$ denotes the nominal loan rate, i^d the nominal deposit rate, and r the real rental rate of capital. The nonnegativity constraints on nominal balances (m_{t+1}) and loans (b_{t+1}) capture the idea that households cannot borrow or lend outside the banking system. Equation (39) imposes a borrowing constraint on credit limits, where $\gamma \geq 0$.

Proposition 2 *Denoting real net wealth by* $x_t \equiv \frac{m_t}{P_t} + W_t n_t + \frac{T_t+\tau_t}{P_t} - (1+\tilde{\imath}_t)\frac{b_t}{P_t} + (1+r_t)\, s_t - s_{t+1}$, *and* $\underline{\theta}_t^* \leq \bar{\theta}_t^* \leq \widetilde{\theta}_t^*$ *as the cutoffs, the decision rules of net wealth, consumption, money holdings, and loan demand are given, respectively, by*

$$x_t = \underline{\theta}_t^* W_t R_t, \tag{40}$$

$$c_t = \begin{cases} \frac{\theta}{\underline{\theta}^*} x_t & \text{if } \theta \leq \underline{\theta}_t^* \\ x_t & \text{if } \underline{\theta}_t^* < \theta \leq \bar{\theta}_t^* \\ \frac{\theta}{\bar{\theta}^*} x_t & \text{if } \bar{\theta}_t^* < \theta \leq \widetilde{\theta}_t^* \\ x_t + \frac{\gamma M_t + \bar{b}}{P_t} & \text{if } \theta > \widetilde{\theta}_t^* \end{cases} \tag{41}$$

$$\frac{m_{t+1}}{P_t} = \begin{cases} \left(1 - \frac{\theta}{\underline{\theta}^*}\right) x_t & \text{if } \theta \leq \underline{\theta}_t^* \\ 0 & \text{if } \theta > \underline{\theta}_t^* \end{cases} \tag{42}$$

$$\frac{b_{t+1}}{P_t} = \begin{cases} 0 & \text{if } \theta \leq \bar{\theta}_t \\ \left(\frac{\theta}{\bar{\theta}^*} - 1\right) x_t & \text{if } \bar{\theta}_t^* < \theta \leq \widetilde{\theta}_t^* \\ \frac{\gamma M_t + \bar{b}}{P_t} & \text{if } \theta > \widetilde{\theta}_t^* \end{cases} \tag{43}$$

where the cutoffs $\left\{\underline{\theta}_t^*, \bar{\theta}_t^*, \widetilde{\theta}_t^*\right\}$ *are determined jointly and uniquely by the following three equations:*

$$\widetilde{\theta}_t^* = \bar{\theta}_t^* \left[1 + \frac{\gamma M_t + \bar{b}}{P_t x_t}\right] \tag{44}$$

$$\frac{\overline{\theta}_t^*}{\underline{\theta}_t^*} = \frac{E_t\left[(1+\tilde{\imath}_{t+1})\frac{P_t}{P_{t+1}w_{t+1}}\right]}{E_t\left[\left(1+i_{t+1}^d\right)\frac{P_t}{P_{t+1}w_{t+1}}\right]} \tag{45}$$

$$1 = \beta E_t\left[\left(1+i_{t+1}^d\right)\frac{P_t W_t}{P_{t+1}W_{t+1}}\right] R(\underline{\theta}_t^*, \bar{\theta}_t^*, \widetilde{\theta}_t^*), \tag{46}$$

where

$$R_t \equiv \int_{\theta<\underline{\theta}^*} d\mathbf{F}(\theta) + \int_{\underline{\theta}^*\le\theta\le\bar{\theta}^*} \frac{\theta}{\underline{\theta}^*} d\mathbf{F}(\theta) + \int_{\bar{\theta}^*<\theta<\widetilde{\theta}^*} \frac{\overline{\theta}^*}{\underline{\theta}^*} d\mathbf{F}(\theta) + \int_{\theta>\widetilde{\theta}^*} \frac{\overline{\theta}^*}{\underline{\theta}^*}\frac{\theta}{\widetilde{\theta}^*} d\mathbf{F}(\theta) \tag{47}$$

measures the liquidity premium of money.

Proof. See Appendix A3. ■

The decision rules for illiquid capital assets and labor supply are similar to the previous models. But here there are four possible cases for money-credit demand: (i) If $m_{t+1} > 0$, then $b_{t+1} = 0$; namely, a household has no incentive to take a loan if it has idle cash in hand. (ii) If $b_{t+1} > 0$, then $m_{t+1} = 0$; namely, a household will take a loan only if it runs out of cash. (iii) It is possible that a household has no cash in hand but does not want to borrow money from the bank because the interest rate is too high; namely, $m_{t+1} = b_{t+1} = 0$. (iv) Finally, the optimal demand for credit may exceed the credit limit and in this case, $b_{t+1} = \gamma M_t + \bar{b}$ and $m_{t+1} = 0$. Which of these situations prevails in each period depends on the realized preference shock θ_t. So there exist three cutoff values with $\underline{\theta}^* \le \bar{\theta}^* \le \widetilde{\theta}^*$ and these cutoffs divide the domain of θ into four regions.

Hence, the consumption function is easy to interpret. If the urge to consume is low ($\theta < \underline{\theta}^*$), then case (i) prevails and $c = \frac{\theta}{\underline{\theta}^*}x < x$. If the urge to consume is high ($\theta > \bar{\theta}^*$), then case (ii) prevails and $c = \frac{\theta}{\bar{\theta}^*}x > x$. In between (if $\underline{\theta}^* \le \theta \le \overline{\theta}^*$), consumption simply equals household net wealth, $c = x$, so case (iii) prevails. Finanlly, if the urge to consume is too high ($\theta > \widetilde{\theta}^*$), then the household opts to hit its credit limit with $c = x + \frac{\gamma M + \bar{b}}{P}$ (case iv). The household in cases (ii) and (iv) is able to consume more than its real net wealth because of the possibility of borrowing.

In the money market, the aggregate supply of credit is $\int m_t d\mathbf{F} = M_t$ and the aggregate demand is $\int b_t d\mathbf{F} = B(\tilde{\imath}_t)$. Note that credit demand can never exceed supply because the loan rate $\tilde{\imath}_t$ will always rise to clear the market. The nominal loan rate cannot be negative because people have

the option not to deposit. Hence, the credit market-clearing conditions are characterized by the following complementarity conditions:

$$(M_t - B_t)\,\tilde{\imath}_t = 0; \quad M_t \geq B_t \text{ and } \tilde{\imath}_t \geq 0. \tag{48}$$

That is, the nominal loan rate is bounded below by zero. The market-clearing condition $(M - B)\,\tilde{\imath} = 0$ determines the nominal interest rate of money. When the supply of funds exceeds credit demand ($M_t > B_t$), the equilibrium interest rate is zero, $\tilde{\imath} = 0$. Otherwise, $\tilde{\imath}$ is determined by the equation $M = B\,(\tilde{\imath})$. Notice that the bank does not accumulate reserves because all reserves are distributed back to bank members by the end of each period. The bank's balance sheet is given by

$$\underbrace{M_t}_{\text{deposit}} + \underbrace{(1+\tilde{\imath}_t)B_t}_{\text{loan payment}} \quad \Longrightarrow \quad \underbrace{\left(1+i_t^d\right)M_t}_{\text{withdraw+interest}} + \underbrace{B_t}_{\text{loan}} + \underbrace{T_t}_{\text{profit income}}, \tag{49}$$

where the LHS is total inflow of funds in period t and the RHS is total outflow of funds in period t. That is, at the beginning of period $t-1$ (more precisely, the second subperiod of $t-1$), the bank accepts deposit M_t and makes new loans B_t, and at the end of period $t-1$ it receives loan payment $(1+\tilde{\imath}_t)\,B_t$, faces withdrawal of M_t, and makes interest payments to depositors. Any profits are distributed back to households in lump sums at the end of period $t-1$ in the amount $T_t = \left(\tilde{\imath}_t - i_t^d\right)B_t$, which becomes household income in the beginning of the next period.

4.2 Welfare Costs of Inflation with Financial Intermediation

Aggregating the household decision rules gives the following relationships linking aggregate consumption, aggregate money demand, and aggregate credit demand, respectively, to household net wealth (x) in the steady state:

$$C = D\left(\underline{\theta}^*, \bar{\theta}^*, \tilde{\theta}^*\right)x \tag{50}$$

$$\frac{M'}{P} = H\left(\underline{\theta}^*, \bar{\theta}^*, \tilde{\theta}^*\right)x \tag{51}$$

$$\frac{B'}{P} = G\left(\underline{\theta}^*, \bar{\theta}^*, \tilde{\theta}^*\right)x, \tag{52}$$

where $D \equiv \int_{\theta<\underline{\theta}^*} \frac{\theta}{\underline{\theta}^*} d\mathbf{F} + \int_{\underline{\theta}^*<\theta<\bar{\theta}^*} d\mathbf{F} + \int_{\bar{\theta}^*<\theta<\tilde{\theta}^*} \frac{\theta}{\bar{\theta}^*} d\mathbf{F} + \int_{\theta>\tilde{\theta}^*} \left[1 + \frac{\gamma M + \bar{b}}{Px}\right] d\mathbf{F}$, $H \equiv \int_{\theta<\underline{\theta}^*} \left(1 - \frac{\theta}{\underline{\theta}^*}\right) d\mathbf{F}$, and $G \equiv \int_{\bar{\theta}^*<\theta<\tilde{\theta}^*} \left(\frac{\theta}{\bar{\theta}^*} - 1\right) d\mathbf{F} + \int_{\theta>\tilde{\theta}^*} \frac{\gamma M + \bar{b}}{Px} d\mathbf{F}$. Notice that $D + H - G = 1$.

The model is closed by adding a representative firm as in the previous section. Hence, the general equilibrium of the model can be solved in the same way as in the previous section. The model has

a unique steady state in which the following relationships hold: $\frac{K}{Y}=\frac{\alpha\beta}{1-\beta(1-\delta)}$, $\frac{C}{Y}=1-\frac{\delta\alpha\beta}{1-\beta(1-\delta)}$, $W=(1-\alpha)\left(\frac{\beta\alpha}{1-\beta(1-\delta)}\right)^{\frac{\alpha}{1-\alpha}}$, and $1+r=\frac{1}{\beta}$. In addition, we also have $R\left(\underline{\theta}^*,\bar{\theta}^*,\widetilde{\theta}^*\right)=\frac{1+\pi}{\beta\left(1+i^d\right)}$, $X=\underline{\theta}^*WR$, and $N=(1-\alpha)\frac{Y}{X}\underline{\theta}^*R$. The welfare cost function looks similar to equation (25), except the values of $\{X,N\}$ are different and the welfare term $J_\theta(\pi)$ in Equation (25) is now given by

$$J_\theta(\pi)=\int_{\theta<\underline{\theta}^*}\theta\log\frac{\theta}{\underline{\theta}^*}d\mathbf{F}+\int_{\bar{\theta}^*<\theta<\widetilde{\theta}^*}\theta\log\frac{\theta}{\bar{\theta}^*}d\mathbf{F}+\int_{\theta>\widetilde{\theta}^*}\theta\log\left[1+\frac{\gamma M+\bar{b}}{Px}\right]d\mathbf{F}. \tag{53}$$

Proposition 3 *Suppose the deposit rate i^d is bounded above by the lending rate $\tilde{\imath}$; if the total supply of funds exceeds the total credit demand ($M_t>B_t$) at inflation rate π for some finite values of $\{\gamma,\bar{b}\}$, then the welfare cost function with banking is identical to that without banking. In other words, financial intermediation does not improve welfare whenever the money-market interest rate is at the zero lower bound.*

Proof. See Appendix A4. ■

Since bank lending has an upper limit when $\{\gamma,\bar{b}\}$ are finite, and since the optimal probability of running out of cash is an increasing function of inflation, under low inflation rates the demand for credit can be too low to exhaust the supply of funds. In this case, Proposition 4 states that there is no welfare gain from financial intermediation. In other words, at sufficiently low inflation rates, self-insurance can achieve identical allocations (distributions) to those with financial intermediation.

To understand this result, imagine first that the credit limit is infinitely small. Then by continuity the welfare cost function in the banking model is identical to that in the benchmark model for all inflation rates. Also, since the supply of funds exceeds the demand, the nominal lending rate is zero for all inflation rates. Second, imagine that the credit limit is unbounded (infinity). Then the welfare cost functions in the two models shall remain identical (cross each other) at the Friedman-rule inflation rate because at this point the supply of funds exceeds credit demand and the nominal lending rate is zero. Therefore, for any finite credit limits the welfare cost functions in the two models must overlap at low inflation rates toward the Friedman rule as long as $M_t>B_t$ or $\tilde{\imath}=0$. In this overlapping interval, financial intermediation does not improve welfare.

We calibrate the credit limit in three ways. First, we set $\gamma=\bar{b}=\infty$, so that there are no borrowing limits, but with the deposit rate $i^d=0$ as in the benchmark model. Second, we keep $i^d=0$ and set $\gamma=0$ and $\bar{b}=0.5x$. Third, we set $\gamma=1,\bar{b}=0.05x$, and the deposit rate equals half of the lending rate: $i^d=\frac{1}{2}\tilde{\imath}$. Namely, in the first case (called Model 1 in Table 3), agents are free to borrow without limits at the market-determined lending rate $\tilde{\imath}$. In the second case (called

Model 2), there exists a fixed credit limit $\bar{b}$ that is 50% of the household's wealth (x).[32] In the third case (called Model 3) we set the credit limit to the household average deposits ($\gamma = 1$) plus an allowance worth 5% of net wealth, and set the deposit rate equal to half of the lending rate.[33]. The rest of the parameters are identical to Calibration 1 in Table 1.

[Insert Figure 3]

The welfare cost functions with different borrowing limits are graphed in Figure 3. The solid line in Figure 3 shows the welfare cost function of the previous benchmark model with capital. It forms an upper envelope on the other three cost functions in the figure. The dashed line represents the cost function in the banking model without borrowing limits (Model 1 in Table 3). It is clear that the welfare cost of inflation can be reduced significantly when banking credit (risk-sharing across households) is available (except at the Friedman rule and the maximum inflation rate $\pi_{\max}$). Nonetheless, the welfare cost still increases (almost linearly) with inflation, such that in the limit when inflation reaches $\pi_{\max}$, the welfare cost of inflation is identical to that without banking or credit.

The dotted line in Figure 3 represents the cost function in the second model with parameters $\left\{\gamma = i^d = 0, \bar{b} = x/2\right\}$, and the dot-dashed line represents the cost function in the third model with parameters $\left\{\gamma = 1, i^d = 0.5\tilde{\imath}, \bar{b} = 0.05x\right\}$. Clearly, under either forms of credit limits (i.e., with finite values of γ and $\bar{b}$), financial intermediation does not improve welfare at sufficiently low inflation rates despite that the cost of borrowing may be zero ($\tilde{\imath} = 0$), confirming Proposition 4. The intuition is as follows. First, with relatively low inflation, households opt to hold a sufficient amount of real balances to buffer consumption shocks since the cost of holding cash is relatively small. Second, the demand for credit is low because of borrowing limits. Hence, households are forced to self-insure against idiosyncratic shocks at relatively low inflation rates. Third, since households take the availability of credit into account when determining the optimal level of net wealth, they opt to reduce their net wealth one for one with the increased credit limit if the borrowing cost is zero. Consequently, the consumption level is not affected and remains the same across all states of nature if $\tilde{\imath} = 0$ (see the proof in Appendix A4); so financial intermediation has no effects on welfare whenever $M_t > B_t$.

However, things change dramatically when the inflation rates are high enough. With sufficiently high inflation, the supply of funds shrinks and the demand for credit rises to a point such that the

[32] Since the average consumption $C = Dx \leq x$, the second calibration for credit limits is quite generous, as it allows a household to borrow more than 50% of its average consumption in the money market.

[33] In the real world the deposit rates are in general significantly lower than the lending rates, especially in developing countries. For example, in China the average interest rate on demand deposits is at most 20% of the lending rate. Telyukova (2009) assumes that $i^d = 0.3\bar{\imath}$ in the U.S. Setting a relatively higher deposit rate implies a lower welfare cost of inflation in our model, so it goes against our welfare results.

loan market clears ($M_t = B_t$). In this case, self-insurance is no longer sufficient and outside credit becomes beneficial even though it is now costly to borrow ($\tilde{\imath} > 0$). So for inflation rate larger than a critical value, the welfare cost of inflation with financial intermediation becomes lower than that in the benchmark model.

In particular, the two cost functions with limited borrowing (the dotted line and the dot-dashed line in Figure 3) become essentially linear (instead of hump-shaped) when π is high enough. Under the credit-limit calibration with $\{\gamma = 1, i^d = 0.5\tilde{\imath}, \bar{b} = 0.05x\}$, the cost function becomes essentially flat for $\pi > 21\%$ per year. This means two things: First, the welfare cost of inflation with banking can be significantly reduced only at relatively high inflation rates. Second, since the cost increases only slowly with inflation for $\pi > 21\%$, the maximum tolerable rate of inflation for households to stop holding money is now much higher than in the benchmark model. This second feature of the banking model arises because the nominal deposit rate increases with inflation whenever $\tilde{\imath} > 0$, which can significantly reduce the inflation tax and the adverse liquidity effect of inflation on money holdings. However, the welfare cost of inflation at moderate inflation rates ($\pi \leq 20\%$) remains just as high as that in the no-banking model (see Table 3).

[Insert Table 3]

On the other hand, under the credit-limit calibration with $\{\gamma = i^d = 0, \bar{b} = x/2\}$, even though deposits (checking accounts) do not pay interest, the welfare cost function starts to deviate from the benchmark model at a much lower inflation rate (around $\pi = 2\%$ or $1 + \pi = 1.02$). Beyond this point, the cost function increases almost linearly and reaches the same maximum cost of 14% at $\pi_{\max} = 52.576\%$. At the moderate inflation rate (π increases from 0% to 10%), the welfare cost is only 1.82% of consumption, more than 2 percentage points lower than in the benchmark model (see Table 3).

Therefore, the welfare gains of credit and banking depend on the form of credit limits and the inflation rate. With reasonable credit limits, there is little gain at low inflation rates because agents can self-insure against consumption risk when the cost of holding money is low, regardless of the form of credit limits. For very high inflation rates close to $\pi_{\max}$, the insurance value of money is low and the cost of borrowing ($\tilde{\imath}$) is high, so redistributing idle cash balances through financial intermediation may not significantly improve welfare (such as in Model 2 under the second credit-limit calibration), unless the nominal interest rate on demand deposits is close to the lending rate, or closely indexed to inflation (such as in Model 3 under the third credit-limit calibration).[34]

How about the redistributive effects of inflation? Because inflation reduces the incentives for

[34] Most types of interest-bearing checking accounts in the U.S. pay very low interest compared with the lending rate (such as the interest rate on credit cards), and the deposit interest rate is often sluggish to reflect inflation changes.

holding money, it thus decreases credit supply (deposits) and increases the costs of borrowing in the credit market. Consequently, the nominal interest of loans in the money market can increase with inflation more than one for one in our model. It is precisely the higher interest cost of loans that may make debtors worse off (instead of better off) in the face of positive inflation, offsetting the redistributive effects noted by Kehoe, Levine, and Woodford (1992). Consequently, inflation can never be optimal in our model despite financial intermediation.

5 Conclusion

This paper provides a tractable general-equilibrium Bewley model of money demand as a new framework for studying important monetary issues, such as the business-cycle properties of velocity,[35] the welfare cost of inflation, monetary credit arrangement and narrow banking. In particular, the framework is used to rationalize the practice of a low inflation target in both developed and developing economies. The model shows that inflation can be very costly—at least about 3% to 4% of consumption under 10% inflation, and that most of the costs are borne by the liquidity-constrained cash-poor agents.

The primary reason for the astonishingly high welfare costs is that inflation erodes the buffer-stock-insurance value of money, thus leading to increased liquidity risk (under preference shocks) or consumption volatility (under income shocks) at the micro level. Such an inflation-induced increase in consumption risk at the household level cannot be captured by the Bailey triangle or representative-agent models or even the standard money-search models.[36]

Our benchmark model deliberately ignores other financial assets (such as stocks and government bonds) as a liquid store of value to meet individual liquidity needs. The empirical facts about the composition of household financial wealth (presented in the Introduction) show that households in developing countries (or a significant number of them in rich countries) hold only cash as the major form of liquidity. Hence, it is only by taking an extreme approach can we hope to find potential upper bounds on the welfare cost of inflation. It would be difficult to rationalize any central bank's inflation targets without knowledge of such upper bounds.

Yet, when we allow the possibility of credit through bank lending, or permit the use of credit cards, the welfare cost of inflation can be reduced significantly, suggesting large welfare gains from financial intermediation or financial development in poor countries. However, with realistic credit limits, the welfare cost of inflation still remains much higher than predicted by representative-agent

[35] See Wen (2009, 2010a).

[36] This result is the opposite of that found by Aiyagari (1994). In a general-equilibrium Bewley model Aiyagari shows that taxing capital is optimal because precautionary saving under uninsurable risks induces dynamic inefficiency at the aggregate level; hence, taxing the rate of return on household saving can improve welfare by increasing the marginal product of capital (the interest rate) towards the Golden rule. In our model, however, savings do not contribute to capital accumulation; thus, taxing nominal balances via inflation does not improve welfare.

models.

Two simplifying strategies allow our general-equilibrium Bewley model to be analytically tractable despite the existence of capital, financial intermediation, and possible aggregate uncertainty. First, the idiosyncratic shocks come from preferences (as in Lucas, 1980) or wealth income (as in Wen, 2010) rather than from labor income (as in Bewley, 1980). Second, the utility function is linear in leisure. These simplifying strategies make the expected marginal utility of an individual's consumption and the target wealth independent of idiosyncratic shocks and individual histories. With these properties, closed-form decision rules for individuals' consumption and money demand can be derived explicitly, and exact aggregation becomes possible under the law of large numbers. The aggregate variables form a system of nonlinear dynamic stochastic equations as in a representative-agent RBC model and can thus be solved by standard methods for business cycle analysis.

These simplifying strategies come at some costs. First, the assumption of preference shocks as the sole source of idiosyncratic risk rules out any positive correlation between the distributions of consumption and money demand. However, Wen (2010a) shows that this correlation problem can be overcome by assuming wealth shocks—this alternative approach preserves closed-form solutions and generates similar welfare costs of inflation. Another cost is that the elasticity of labor supply is not a free parameter. This imposes some limits on the model's ability to study labor supply behavior and labor market dynamics within this framework. Nonetheless, the payoff of the simplifying assumptions is obvious: They not only make the generalized Bewley model analytically tractable regardless of aggregate uncertainty and capital accumulation and financial intermediation, but they also reduce the computational costs for a heterogeneous-agent model to the level of solving a representative-agent RBC model. Because of these advantages, the model may prove useful in applied work and serve as an alternative framework to the Baumol-Tobin model and the Lagos-Wright (2005) model for monetary policy analysis.[37]

[37] The Lagos-Wright model is not suitable for dynamic business-cycle analysis, whereas the model presented in this paper does not suffer from this weakness.

References

[1] Aiyagari, S.R., 1994. Uninsured idiosyncratic risk and aggregate saving. Quarterly Journal of Economics 109, 659-684.

[2] Aiyagari, S.R., Williamson, S.D., 2000. Money and dynamic credit arrangements under private information. Journal of Economic Theory 91, 248-279.

[3] Akyol, A., 2004. Optimal monetary policy in an economy with incomplete markets and idiosyncratic risk. Journal of Monetary Economics 51, 1245-1269.

[4] Alvarez, F., Atkeson, A., Edmond, C.,2008. Sluggish responses of prices and inflation to monetary shocks in an inventory model of money demand. Quarterly Journal of Economics 124, 911-967.

[5] Andolfatto, D., Nosal, E., 2003. A theory of money and banking. Federal Reserve Bank of Cleveland Working Paper 03-10.

[6] Attanasio, O., Guiso, L., Jappelli, T., 2002. The demand for money, financial innovation, and the welfare cost of inflation: An analysis with household data. Journal of Political Economy 110, 317-351.

[7] Bailey, M., 1956. The welfare cost of inflationary finance. Journal of Political Economy 64, 93-110.

[8] Baumol, W. J., 1952. The transactions demand for cash: An inventory theoretic approach. Quarterly Journal of Economics 66, 545-556.

[9] Bewley, T., 1980. The optimum quantity of money. In: Kareken, J., Wallace, N. (Eds.), Models of Monetary Economies. Federal Reserve Bank of Minneapolis, Minneapolis.

[10] Bewley, T., 1983. A difficulty with the optimum quantity of money. Econometrica 51, 1485-1504.

[11] Cartwright, P., Delorme, C., Wood, N., 1985. The by-product theory of revolution: Some empirical evidence. Public Choice 46, 265-274.

[12] Chiu, J., 2007. Endogenously segmented asset market in an inventory theoretic model of money demand. Bank of Canada Working Paper 2007-46.

[13] Cooley, T., Hansen, G., 1989. The inflation tax in a real business cycle model. American Economic Review 79, 733-748.

[14] Dotsey, M., Ireland, P., 1996. The welfare cost of inflation in general equilibrium. Journal of Monetary Economics 37, 29-47.

[15] Erosaa, Andrés and Gustavo Ventura, 2002, On inflation as a regressive consumption tax, Journal of Monetary Economics 49(4), 761-795.

[16] Guvenen, F., Smith, A., 2010. Inferring labor income risk from economic choices: An indirect inference approach. NBER Working Paper 16327.

[17] Gross, D., Souleles, N., 2002. Do liquidity constraints and interest rates matter for consumer behavior? Evidence from credit card data. Quarterly Journal of Economics 117, 149-185.

[18] Heathcote, J., Storesletten, K., Violante, G., 2008. Quantitative macroeconomics with heterogeneous households. Federal Reserve Bank of Minneapolis Research Department Staff Report.

[19] Henriksen, E., Kydland, F., 2010. Endogenous money, inflation, and welfare. Review of Economic Dynamics 13, 470–486.

[20] Huggett, M., 1993. The risk-free rate in heterogeneous-agent incomplete-insurance economies. Journal of Economic Dynamics and Control 17, 953-969.

[21] Imrohoroglu, A., 1989. Cost of business cycles with indivisibilities and liquidity constraints. Journal of Political Economy 97, 1364-1383.

[22] Imrohoroglu, A., 1992. The welfare cost of inflation under imperfect insurance. Journal of Economic Dynamics and Control 16, 79-91.

[23] Jappelli, T., 1990. Who is credit constrained in the U.S. economy? Quarterly Journal of Economics 105, 219-234.

[24] Kehoe, T., Levine, D., Woodford, M., 1992. The optimum quantity of money revisited. In Dasgupta, P., Gale, D., Hart, O, Maskin, E. (Eds.), Economic Analysis of Markets and Games. M.I.T. Press, Cambridge.

[25] Krusell, P., Smith, A., 1998. Income and wealth heterogeneity in the macroeconomy. Journal of Political Economy 106, 867-96.

[26] Lagos, R., Wright, R., 2005. A unified framework for monetary theory and policy analysis. Journal of Political Economy 113, 463-484.

[27] Looney, R., 1985. The Inflationary process in prerevolutionary Iran. Journal of Developing Areas 19, 329-350.

[28] Lucas, R.E. Jr., 1980. Equilibrium in a pure currency economy. Economic Inquiry 18, 203-220.

[29] Lucas, R.E. Jr., 1990. Liquidity and interest rates. Journal of Economic Theory 50, 237-264.

[30] Lucas, R.E. Jr., 1990b. Supply-side economics: An analytical review. Oxford Economic Papers 42, 293-316.

[31] Lucas, R.E. Jr., 2000. Inflation and welfare. Econometrica 68, 247-274.

[32] Makinen, M. et al., 2007. Inequalities in health care use and expenditures: Empirical data from eight developing countries. Bulletin of the World Health Organization 78, 55-65.

[33] Ragot, X., 2009. The case for a financial approach to money demand. Banque de France Working Paper 300.

[34] Stokey, N.L., Lucas, R.E. Jr., 1989. Recursive Methods in Economic Dynamics. Harvard University Press, Cambridge.

[35] Svensson, L., 1985. Money and asset prices in a cash-in-advance economy. Journal of Political Economy 93, 919-944.

[36] Telyukova, Irina, 2011, Household Need for Liquidity and the Credit Card Debt Puzzle, Working Paper, University of California, San Diego.

[37] Telyukovay, Irina and Ludo Visschers, 2013, Precautionary Demand for Money in a Monetary Business Cycle Model, Journal of Monetary Economics (forthcoming).

[38] Tobin, J., 1956. The interest-elasticity of the transactions demand for cash. Review of Economics and Statistics 38, 241-247.

[39] Townsend, R., 1995. Consumption insurance: An evaluation of risk-bearing systems in low-income economies. Journal of Economic Perspectives 9, 83-102.

[40] Wen, Y., 2009. Liquidity and welfare in a heterogeneous-agent economy. Federal Reserve Bank of St. Louis Working Paper 2009-019B.

[41] Wen, Y., 2010a. Liquidity demand and welfare in a heterogeneous-agent economy. Federal Reserve Bank of St. Louis Working Paper 2010-009A.

[42] Wen, Y., 2010b, Lucas meets Baumol and Tobin, Federal Reserve Bank of St. Louis Working Paper 2010-014B.

[43] Wen, Y., 2011, Input and output inventory dynamics, American Economic Journal: Macroeconomics 3 (October), 181-212.

Table 1. Calibrated Parameter Values

	α	β	δ	σ	A	Implied $\sigma^2_{\log c}$
Method 1	0.42	0.95	0.1	2.65	1	0.055
Method 2	0.42	0.95	0.1	2.1∼3.1	0.54∼1.45	0.03∼0.11
Method 3	0.42	0.95	0.1	2.0∼2.5	0.48∼0.86	0.07∼0.13

Note: α denotes capital share, β time discount factor, δ capital depreciation, σ shape parameter in Pareto distribution, A the scaling factor in equation (34), and $\sigma^2_{\log c}$ the variance of log household consumption.

Table 2. Welfare Costs (Δ% of Consumption)

	Raising π from 0% to 10%	Raising π from 4% to 14%
Calibration 1	3.89	3.05
Calibration 2	4.95	3.76
Calibration 3	6.13	4.94

Note: 3rd and 4th rows report the average cost under the two values of σ in Table 1.

Table 3. Welfare Costs with Banking (Δ% of Consumption)

	Raising π from 0% to 10%	Raising π from 4% to 14%
Benchmark	3.89	3.05
Model 1	2.32	2.47
Model 2	1.82	1.01
Model 3	3.89	3.05

Note: Model 1 is the banking model without borrowing limits, Model 2 features a fixed credit limit of $\bar{b} = 0.5x$, and Model 3 features a credit limit proportional to the average deposits $\frac{M}{P}$ and a deposit rate of $i^d_t = 0.5\tilde{i}_t$.

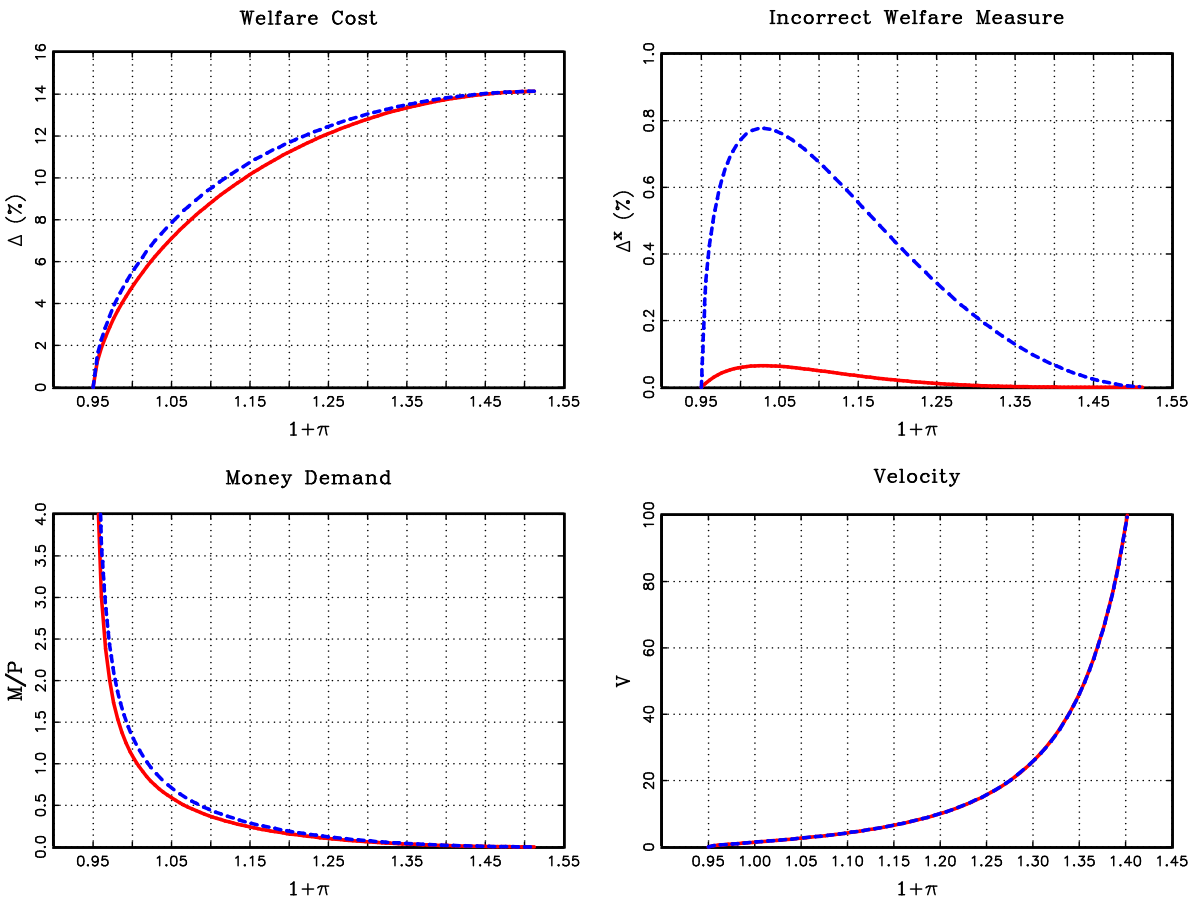

Figure 1. Welfare Costs, Money Demand, and Velocity.

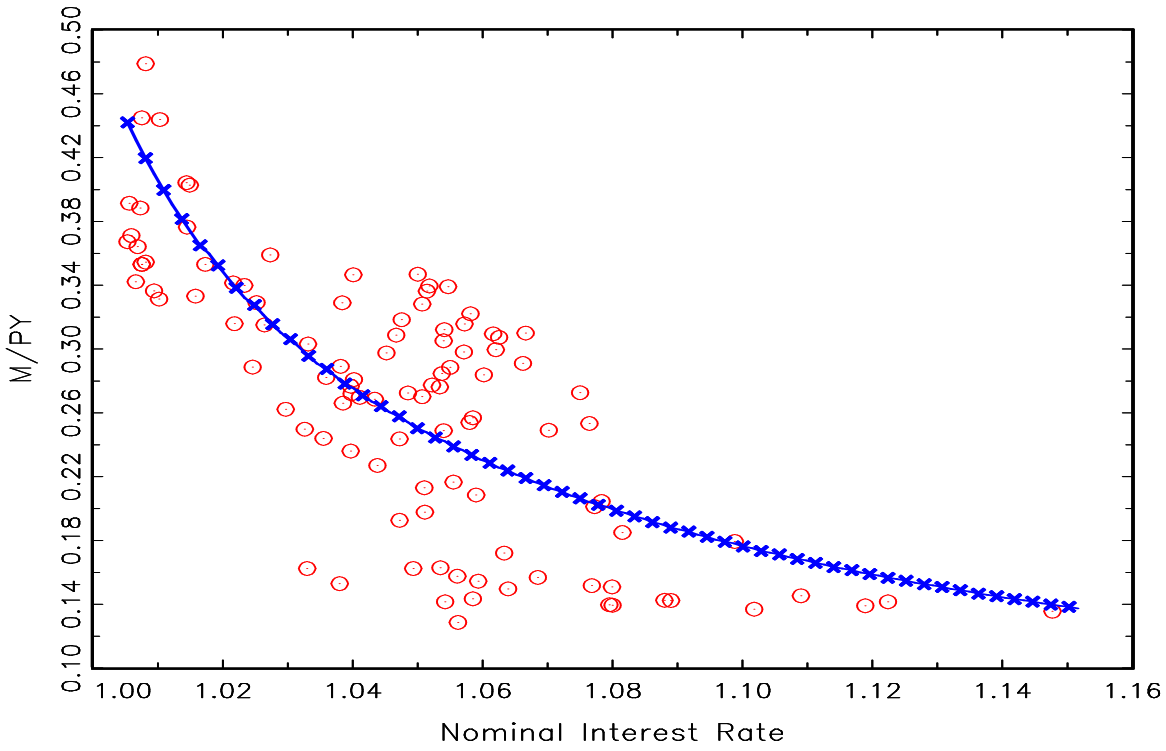

Figure 2. Aggregate Money Demand Curve in the Model ($\times - \times - \times$) and Data (o o o).

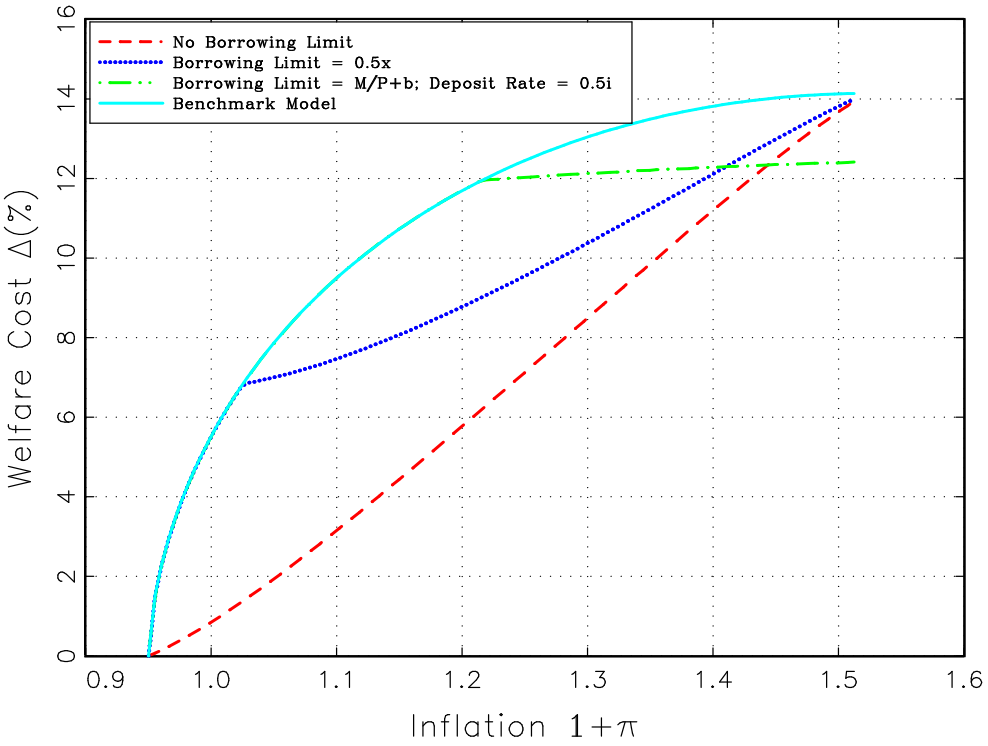

Figure 3. Welfare Costs with Credit and Banking.

Appendix

A1. Proof of Proposition 1

Proof. Denoting $\{\lambda_t, v_t\}$ as the Lagrangian multipliers for constraints (5) and (6), respectively, and assuming interior solution for n_t, the first-order conditions for $\{c_t, m_{t+1}, n_t\}$ are given, respectively, by

$$\frac{\theta_t}{c_t} = \lambda_t \tag{54}$$

$$\lambda_t = \beta E_t \left[\frac{\partial V_{t+1}}{\partial \tilde{m}_{t+1}} \frac{P_t}{P_{t+1}} \right] + P_t v_t \tag{55}$$

$$1 = \int \frac{\partial J_t}{\partial x_t} W_t d\mathbf{F}, \tag{56}$$

where $\tilde{m}_t \equiv \frac{m_t}{P_t}$ denotes real balances. The envelop theory implies

$$\frac{\partial J_t}{\partial x_t} = \lambda_t \tag{57}$$

$$\frac{\partial V_t}{\partial \tilde{m}_t} = \int \frac{\partial J_t}{\partial x_t} d\mathbf{F}. \tag{58}$$

Equation (56) reflects the assumption that decision for labor supply n_t must be made before the idiosyncratic preference shock θ_t (and hence the value of $J(x_t, \theta_t)$) are realized.

By the law of iterated expectations and the orthogonality assumption of aggregate and idiosyncratic shocks, equations (56) and (55) can be rewritten, respectively, as

$$\frac{1}{W_t} = \int \lambda_t d\mathbf{F} \tag{59}$$

$$\lambda_t = \beta E_t \frac{P_t}{P_{t+1} W_{t+1}} + v_t, \tag{60}$$

where $\frac{1}{W}$ pertains to the expected marginal utility of consumption in terms of labor income. The decision rules for consumption and money demand are characterized by a cutoff strategy, taking as given the aggregate environment. Denoting the cutoff by θ_t^*, we consider two possible cases below.

Case A. $\theta_t \leq \theta_t^*$. In this case, the urge to consume is low relative to a target. It is hence optimal to hold money as a store of value (to prevent possible liquidity constraints in the future). So $m_{t+1} \geq 0$, $v_t = 0$, and the shadow value of good (marginal utility of consumption) $\lambda_t = \beta E_t \frac{P_t}{W_{t+1} P_{t+1}}$. Thus, $c_t = \theta_t \left[\beta E_t \frac{P_t}{W_{t+1} P_{t+1}} \right]^{-1}$.

Case B. $\theta_t(i) > \theta_t^*$. In this case, the urge to consume is high relative to a target. It is then optimal to spend all money in hand, so $v_t > 0$ and $m_{t+1} = 0$. By the resource constraint (5), we have $c_t = x_t$. Equation (54) then implies that the marginal utility of consumption is given by $\lambda_t = \frac{\theta_t}{x_t}$.

The above considerations imply

$$\lambda_t = \max\left\{\beta E_t \frac{P_t}{W_{t+1}P_{t+1}}, \frac{\theta_t}{x_t}\right\}, \tag{61}$$

which determines the cutoff:

$$\beta E_t \frac{P_t}{W_{t+1}P_{t+1}} \equiv \frac{\theta_t^*}{x_t}. \tag{62}$$

Equation (59) then implies

$$\frac{1}{W_t} = \int \max\left\{\left[\beta E_t \frac{P_t}{W_{t+1}P_{t+1}}\right], \frac{\theta_t}{x_t}\right\} d\mathbf{F}, \tag{63}$$

which implicitly determines the optimal wealth $x\left(H^t\right)$ as a function of aggregate state only (i.e., x_t is independent of individual history h^t). Using equation (62) to substitute real wealth x_t, equation (63) implies the Euler equation (11) for money demand.

The above analyses and the first-order conditions imply the decision rules for consumption, money demand, real wealth, and labor supply summarized in the equations in Proposition 1. Notice that labor supply, $n_t = \frac{1}{W_t}\left[x_t - \frac{m_t+\tau_t}{P_t}\right]$, may be negative if the existing real balances are too high. To ensure that we have an interior solution ($n > 0$), consider the worst possible case where money demand takes its maximum possible value, $\frac{m_t}{P_{t-1}} = \max\left\{0, \frac{\theta^*-\theta}{\theta^*}\right\} x = \frac{\theta^*-\theta_L}{\theta^*} x$. Suppose $M_{t+1} = M_t + \tau_t = (1+\mu) M_t$ and $\frac{P_t}{P_{t-1}} = 1 + \pi = 1 + \mu$ in the steady state, then $\frac{\tau_t}{P_t} = \frac{\pi M_t}{P_t} = \frac{\pi}{1+\pi}\frac{M_{t+1}}{P_t} = \frac{\pi}{1+\pi}\int \frac{m_{t+1}}{P_t} d\mathbf{F} = \frac{\pi}{1+\pi} H(\theta^*) x$ (by the decision rule), where $H(\theta^*) \equiv \int \max\left\{0, \frac{\theta^*-\theta}{\theta^*}\right\} d\mathbf{F}$. According to the definition of x in equation (3), $n > 0$ even in the worst possible case (i.e., the minimum value of n is greater than 0) if $x_t - \frac{m_t+\tau_t}{P_t} = \left[1 - \frac{\theta^*-\theta_L}{\theta^*} - \frac{\pi}{1+\pi} H(\theta^*)\right] x = \left[\frac{\theta_L}{\theta^*} - \frac{\pi}{1+\pi} H(\theta^*)\right] x = \left[\frac{\theta_L}{\theta^*} - \frac{\pi}{\beta}\frac{H(\theta^*)}{R(\theta^*)}\right] x > 0$. This condition is clearly satisfied when $\pi = 0$. It is also satisfied in our model for $0 < \pi \leq \pi_{\max}$ since the ratio $\frac{H(\theta^*)}{R(\theta^*)}$ is monotonically decreasing in π and approaches zero as $\pi \to \pi_{\max}$, whereas θ^* is decreasing in π but approaches θ_L, so the first term inside the bracket always exceeds the second term for any $\pi \in [\beta - 1, \pi_{\max}]$ under our parameter calibrations for β. The intuition is simple: Since consumption is bounded below by zero under log utility and it

depends on labor income, under strictly positive preference shocks cash holdings can never become too large to render hours worked negative. Thus, labor supply is always positive. It is also easy to see that n is bounded away from above by a sufficiently large constant $\bar{n}$ because cash holdings are bounded below by zero. ■

A2. Proposition 2 and the Proof

Proposition 4 *The general equilibrium of the model with capital can be characterized by the dynamic paths of ten aggregate variables,* $\{C_t, K_{t+1}, M_{t+1}, N_t, X_t, Y_t, \theta_t^*, P_t, W_t, r_t\}$*, which can be uniquely solved by the following ten aggregate equations:*

$$C_t = D(\theta_t^*)X_t \tag{64}$$

$$\frac{M_{t+1}}{P_t} = H(\theta_t^*)X_t \tag{65}$$

$$X_t = W_t\theta_t^* R(\theta_t^*) \tag{66}$$

$$X_t = \frac{M_t + \tau_t}{P_t} + (1 + r_t)K_t - K_{t+1} + W_t N_t \tag{67}$$

$$\frac{1}{W_t} = \beta E_t \frac{1}{W_{t+1}} \frac{P_t}{P_{t+1}} R\left(\theta_t^*\right) \tag{68}$$

$$\frac{1}{W_t} = \beta E_t \frac{1 + r_{t+1}}{W_{t+1}} \tag{69}$$

$$W_t = (1 - \alpha)\frac{Y_t}{N_t} \tag{70}$$

$$r_t + \delta = \alpha \frac{Y_t}{K_t} \tag{71}$$

$$C_t + K_{t+1} - (1 - \delta)K_t = Y_t \tag{72}$$

$$Y_t = A_t K_t^{\alpha} N_t^{1-\alpha}, \tag{73}$$

where $D(\theta^*) \equiv \int \min\left\{1, \frac{\theta}{\theta^*}\right\} d\mathbf{F}$*,* $H(\theta^*) \equiv \int \max\left\{0, \frac{\theta^* - \theta_t}{\theta^*}\right\} d\mathbf{F}$*, and* $R(\theta_t^*) \equiv \int \max\left\{1, \frac{\theta_t}{\theta_t^*}\right\} d\mathbf{F}$*.*

The aggregate dynamics of the model (such as transitional dynamics or impulse responses to aggregate shocks) can be solved by standard methods popular in the representative-agent RBC literature, such as the log-linearization method.

Proof. We provide only a briefly sketch of the proof here. Details can be found in Wen (2009, 2010a). Define

$$\tilde{x}_t \equiv \frac{m_t + \tau_t}{P_t} + W_t n_t + (1 + r_t)\, s_t \tag{74}$$

as net real wealth, and denote $J(x_t, \theta_t)$ as the value function of the household based on the choice of c_t, m_{t+1}, and s_{t+1}. We then have

$$J(\tilde{x}_t, \theta_t) = \max_{c_t, m_{t+1}} \left\{ \theta_t \log c_t + \beta E_t V(\frac{m_{t+1}}{P_{t+1}}, s_{t+1}) \right\} \tag{75}$$

subject to

$$c_t + \frac{m_{t+1}}{P_t} + s_{t+1} \leq \tilde{x}_t \tag{76}$$

$$m_{t+1} \geq 0, \tag{77}$$

where $V(\frac{m_t}{P_t}, s_t)$ is the value function of the household based on the choices of n_t. That is,

$$V(\frac{m_t}{P_t}, s_t) = \max_{n_t} \left\{ -n_t + \int J(\tilde{x}_t, \theta_t) d\mathbf{F} \right\} \tag{78}$$

subject to (74) and $n_t \in [0, \bar{n}]$. Denoting $\{\lambda_t, v_t\}$ as the Lagrangian multipliers for constraints (76) and (77), respectively, and assuming interior solution for n_t, the first-order conditions for $\{c_t, m_{t+1}, s_{t+1}, n_t\}$ are given, respectively, by

$$\frac{\theta_t}{c_t} = \lambda_t \tag{79}$$

$$\lambda_t = \beta E_t \left[\frac{\partial V_{t+1}}{\partial \tilde{m}_{t+1}} \frac{P_t}{P_{t+1}} \right] + P_t v_t \tag{80}$$

$$\int \lambda_t d\mathbf{F} = \beta E_t \left[\frac{\partial V_{t+1}}{\partial s_{t+1}} \right] \tag{81}$$

$$1 = W_t \int \frac{\partial J_t}{\partial \tilde{x}_t} d\mathbf{F}, \tag{82}$$

where $\tilde{m}_t \equiv \frac{m_t}{P_t}$ denotes real balances and equation (81) is derived by differentiating both sides of equation (75) with respect to s_{t+1}. The envelop theory implies

$$\frac{\partial J_t}{\partial \tilde{x}_t} = \lambda_t \tag{83}$$

$$\frac{\partial V_t}{\partial \tilde{m}_t} = \int \frac{\partial J_t}{\partial \tilde{x}_t} d\mathbf{F} \tag{84}$$

$$\frac{\partial V_t}{\partial s_t} = (1 + r_t) \int \frac{\partial J_t}{\partial \tilde{x}_t} d\mathbf{F}. \tag{85}$$

Equation (81) reflects the assumption that decision for s_{t+1} is made before the idiosyncratic preference shock θ_t (and hence the value of λ_t) is realized. Equations (81)-(85) together imply equation (31). The first-order conditions for $\{c_t, \frac{m_{t+1}}{P_t}, n_t\}$ are identical to those in the benchmark model. After redefining cash in hand as $x_t = \tilde{x}_t - s_{t+1}$, the rest of the proof is similar to that in Appendix A1 for Proposition 1. ■

A3. Proof of Proposition 3

Proof. The proof is analogous to that in Appendix A1 or Appendix A2. The only important difference is to realize that we have three cutoffs here, $\left(\underline{\theta}^*, \overline{\theta}^*, \widetilde{\theta}^*\right)$. Hence, we have four possible cases to consider as elaborated in the main text. ■

A4. Proof of Proposition 4

Proof. It suffices to show that if $M_t > B_t$, then both the individual consumption level and the labor supply in the banking model are identical, respectively, to those in the benchmark model. First, notice that the real wage W in the banking model is identical to that in the benchmark model because the capital-to-output ratio is not affected by financial intermediation. Second, when $M_t > B_t$, equation (48) implies that $\tilde{\imath}_t = 0$. Hence, we also have $i_t^d = 0$. Equation (45) then implies $\bar{\theta} = \underline{\theta}$. Thus, equations (46) and (47) imply

$$\frac{1+\pi}{\beta} = R\left(\widetilde{\theta}^*\right) = \int_{\theta < \widetilde{\theta}^*} d\mathbf{F}(\theta) + \int_{\theta > \widetilde{\theta}^*} \frac{\theta}{\widetilde{\theta}^*} d\mathbf{F}(\theta), \tag{86}$$

which is identical to equations (12) and (21) in the benchmark model. Hence, the liquidity premium function R is identical in the two models with $\widetilde{\theta}^* = \theta^*$. In addition, equation (44) implies $\left[1 + \frac{\gamma M_t + \bar{b}}{P_t x_t}\right] = \frac{\widetilde{\theta}^*}{\bar{\theta}^*}$, so the decision rules in equations (40) and (41) imply that household consumption in the banking model is given by

$$\begin{aligned} c_t &= \begin{cases} \theta W R & \text{if } \theta \leq \widetilde{\theta}^* \\ \widetilde{\theta}^* W R & \text{if } \theta > \widetilde{\theta}^* \end{cases} \\ &= \min\left\{1, \frac{\theta}{\widetilde{\theta}^*}\right\} \widetilde{\theta}^* W R, \end{aligned} \tag{87}$$

which is identical to equation (8) in the benchmark model after substituting out x by equation (10). As a result, the aggregate consumption level C and aggregate output Y are also identical across the two models, respectively. Finally, labor supply in the banking model is given by $N = (1-\alpha)\frac{Y}{W}$, which is identical to that in the benchmark model because the aggregate output level is. Therefore, the welfare cost function in the banking model is identical to that in the benchmark model whenever the credit market for loanable funds does not clear (or the lending rate is at the zero lower bound) because of excess liquidity on the supply side and borrowing constraints on the credit demand side. ■

Appendix B. Calibration Methods (Not for Publication)

Method 1. We choose the value of σ to minimize the distance between the model and the data using the following least square function,

$$\Gamma = \sum_{i=1}^{T} \left(\left(x_i^m - x_i^d \right)^2 + \left(y_i^m - y_i^d \right)^2 \right) + 2 \left(\max\left\{ x_i^m \right\} - \max\left\{ x_i^d \right\} \right) + 2 \left(\max\left\{ y_i^m \right\} - \max\left\{ y_i^d \right\} \right), \tag{88}$$

where T denotes sample size, x_i^d denotes the ith observation of the nominal interest rate in the Lucas data, x_i^m its model counterpart, y_i^d the ith observation of the money demand in the Lucas data, and y_i^m its model counterpart. The last two terms in the above function serve to put more weight on the two end points of the demand curve.

Method 2. In addition to minimizing the moment condition in (88), we add an additional constraint that the model-implied likelihood of running out of cash is consistent with the household survey data. This alternative calibration method is to choose the parameters $\{\sigma, A\}$ so that the probability of running out of cash $1 - \mathbf{F}(\theta^*)$ in our model equals the proportion of liquidity-constrained population in the United States. According to the Survey of Consumer Finances (SCF), the portion of households having zero balances in checking accounts is 19.3% of the population based on surveys in the years between 1989 and 2007 (with standard deviation of 1.3%), the portion of households having less than \$10 in checking accounts is 20% (with standard deviation of 1.4%), and that having less than \$20 is 20.6% (with standard deviation 1.3%).[38] Households with little balances in their checking accounts also tend to have very little balances in other types of accounts, such as saving accounts and money-market accounts. Hence, if we define a household with zero balances in checking accounts as those facing a binding liquidity constraint in our model, we can choose $\{\sigma, A\}$ such that (i) $1 - \mathbf{F}(\theta^*) = 0.2$ when the inflation rate is 4% per year and (ii) the

[38] On the other hand, the portion of households with balances greater than $\$3,000$ in checking accounts is larger than 32% with standard deviation of 2.2%. See Wen (2010a) for more details.

value of Γ in equation (88) is minimized.[39] One problem with this approach is that the fraction of the population with zero balances at any moment may overstate the likelihood of running out of cash for a typical individual. These two statistics are identical only under the assumption of iid idiosyncratic shocks. However, based on the rule of thumb that over-time risk is roughly half of the cross-household dispersion in income and consumption, we can set the probability of running out of cash in the model to the interval $(0.1, 0.2)$. To generate a $10\% \sim 20\%$ probability of running out of cash in the model when the inflation rate is 4% a year (in addition to minimize Γ), it requires $\sigma \in (2.1, 3.1)$ and $A \in (0.54, 1.45)$, respectively.

Method 3. Similar to Method 2, we require the model to match the consumption risk under unexpected shocks to consumption demand. One measure of consumption risk is household expenditure volatility. It is arguable that aside from income uncertainty, a perhaps more important source of consumption volatility (especially in developing countries) is expenditure uncertainty, such as unexpected spending for housing, education, and health care, or unpredictable expenditures related to accidents, property damages, and volatile fluctuations in consumption goods prices. Such expenditure uncertainty is especially large and highly uninsurable in developing countries than developed countries because of the lack of insurance markets. An ideal proxy of spending risk would be the frequency of illness and the associated costs or accessibility of medical services, but such data are either unavailable or highly inadequate. Wen (2011b) reports that the risk related to car accidents in China is 24 to 35 times higher than in the U.S. Also, the risk of work-related injuries in China is two orders of magnitude higher than in the U.S. For example, the average annual incidence rate of fatal injuries in the U.S. mining industry is 0.026% (or 2.6 individuals per 1000 persons for the period 2005-09). The comparable incidence rate in China (for the period 1981-94) is about 15%. Alternatively, if the accident rate is measured by the number of fatal injuries (death) per millions of tons of coal output, the value is 0.02% in the U.S. and 4% in China.

Since there do not exist long enough time-series panel data for household expenditures (SCF do not keep track of the same households for more than one year), an alternative way to gauge expenditure uncertainty is the Gini coefficient for households with similar income and living standards. The following table shows the Gini coefficients for consumption expenditure and health care expenditure across households in villages of developing countries. Based on the rule of thumb that over-time risk is roughly half of the cross-household dispersion in consumption expenditure,[40] we

[39] A large body of empirical literature suggests that 19% of the U.S. population is liquidity constrained. For example, Hall and Mishkin (1982) use the Panel Study of Income Dynamics and find that 20% of American families are liquidity constrained. Mariger (1986) uses a life-cycle model to estimate this fraction to be 19.4%. Hubbard and Judd (1986) simulate a model with a constraint on net worth and find that about 19.0% of United States consumers are liquidity constrained. Jappelli (1990) uses information on individuals whose request for credit has been rejected by financial intermediaries and estimates through a Tobit model that 19.0% of families are liquidity constrained. Therefore, the emerging consensus points to a fraction of approximately 20% of the population to be liquidity constrained.

[40] See Wen (2011b) listed below for references.

can infer from the table the approximate consumption spending risk faced by households in these countries.

Expenditure Inequality for Developing Countries

	Burkina Faso	Guate-mala	Kazakh-stan	Kyrgyz-stan	Para–guay	South Africa	Thailand
Consumption Gini	0.43	0.39	0.37	0.45	0.47	0.54	0.39
	Burkina Faso	Guate-mala	Kazakh-stan	South Africa	Para–guay	Zambia	Thailand
Health care Gini	0.43	0.42	NA	0.67	0.18	0.32	0.38

The average consumption Gini across these developing countries is 0.43 and the average health-care Gini is 0.4; these values are both significantly larger than the consumption Gini (0.3) in the United States. So we calibrate the model to generate a consumption Gini of 0.15 for the U.S. and 0.2 for developing countries. This calibration yields the range of parameter values of σ in the last row in Table 1 under Method 3. The implied consumption volatility under these calibrated values of σ turn out consistent with the empirical estimates provided by Telyukova (2011, Table 9) for U.S. households. Hence, we believe that our calibration provides a reasonable benchmark value on the consumption risk in developing countries.

References

[1] Wen, Y., 2011b. Explaining China's Trade Imbalance Puzzle. Federal Reserve Bank of St. Louis Working Paper 2011-018A.