Economic Quarterly— Volume 104, Number 1— First Quarter 2018— Pages 1–33

Cyclical Properties of Bank Margins: Small versus Large Banks

Borys Grochulski, Daniel Schwam, and Yuzhe Zhang

The US banking sector is composed of few very large banks and many small ones.[1] In fact, about 95 percent of all Federal Deposit Insurance Corporation-insured depository institutions are small, community banks.[2] Although these small banks hold only about 15 percent of total bank assets by value, many observers, bank supervisors, and monetary policymakers share the view that small, local banks fulfill an important role in the intermediation of credit in the US.

Former Federal Reserve Governor (and current chairman) Jerome H. Powell expresses this view in his speech to the 2013 Community Banking Research Conference:

> My colleagues on the Board of Governors and I understand the value of having a diverse financial system that includes a large and vibrant contingent of community banks. By fostering the economic health and vitality of local communities throughout the country, community banks play a central role in our national economy. One important aspect of that role is to serve as a primary source of credit for

The authors would like to thank Erica Paulos, Nicholas Trachter, John Walter, John Weinberg, and Russell Wong for their helpful comments, and Sara Ho for excellent research assistance. The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] See Section 4 for a short summary or McCord and Prescott (2014) for detailed analysis.

[2] Traditionally, a community bank is defined as a small bank that operates in a single banking market. A single banking market is defined as a county in rural areas or as a metropolitan statistical area in urban areas; see, e.g., Powell (2016). A small bank is often defined as one with less than $1 billion in assets. Recently, however, many studies have used a $10 billion size cutoff instead. See Keeton et al. (2003) and Federal Deposit Insurance Corporation (2012).

DOI: https://doi.org/10.21144/eq1040101

> the small businesses that are responsible for creating a substantial proportion of all new jobs.[3]

More specifically, small banks are thought of as having access to granular information about local business conditions. In its 2012 Community Banking Study, the FDIC states that small banks

> [...] have specialized knowledge of their local community and their customers. Because of this expertise, community banks tend to base credit decisions on local knowledge and nonstandard data obtained through long-term relationships [...] This relationship approach to lending is particularly important to small businesses that rely on community banks for loans and other services.

Further, this granular information is thought of as containing early signals of changing business cycle conditions. In her address to the 2017 convention of the Community Bankers Association of Ohio, President Loretta J. Mester of the Federal Reserve Bank of Cleveland states:

> Because of their important work, community bankers are among the most knowledgeable about changes in conditions on the ground in local areas. Such information often takes much longer to show up in official statistical reports. So I find the insights gained from speaking with bankers to be especially valuable as part of the mosaic of information I use in formulating my views on appropriate monetary policy.

In this article, our objective is to explore US commercial banking data and look for signs consistent with these widely held views about the special role of small banks. In particular, from these views we extract two hypotheses. First, if small banks indeed have access to some unique business cycle information not available to other financial intermediaries, then one could expect to see differences in how small and large banks' profit margins react to changes in business cycle conditions. Second, if the advance information available to small banks comes from long-term relationships with local businesses, i.e., the borrowers, one could expect these differences to appear on the asset side of the banks' balance sheet rather than on the liability side.

In the empirical literature on bank profitability, the net interest margin (NIM) is perhaps the most commonly used profit margin indicator. NIM is defined as the ratio of a bank's net interest income and average earning assets. Net interest income, in turn, is the difference

[3] Similar views have been expressed by other Federal Reserve officials; see, e.g., Keeton et al. (2003) and Mester (2017).

between interest earned on assets and interest incurred on liabilities. The cyclical properties of NIM have been studied extensively in aggregate US banking sector data. Using administrative data collected by the FDIC, Aliaga-Díaz and Olivero (2010, 2011) and Beabrun-Diant and Tripier (2015) show that the sector's NIM is countercyclical. However, little is known about any heterogeneity in the cyclical properties of NIM between small and large banks.

In the first part of this article, we consider the first of our two hypotheses. We investigate the cyclical properties of NIM in the banking sector as a whole and, separately, among small and large banks. Our baseline definition of a small bank follows Kashyap and Stein (2000), where small banks are defined as all institutions below the 95th percentile of the size distribution by assets.[4] We find that there is a significant difference in the cyclical behavior of NIM at large and small banks. Among large banks, similar to the earlier estimates obtained for all banks, the cyclical component of NIM exhibits negative correlation with the cyclical component of gross domestic product (GDP). We find a point estimate of this correlation coefficient of -0.33. Our main finding in this article is that, by contrast, among small banks, i.e., 95 percent of all banks, NIM is positively correlated with the business cycle. Among small banks, the estimated correlation between the cyclical components of average NIM and GDP is 0.34.

This empirical finding documents a significant difference between small and large banks, consistent with the view that the timing of business cycle signals received by small banks is different from the aggregate statistics. It is important, however, not to overstate our result. In this article, we report a difference in the correlations of small and large banks' NIM with GDP without establishing causation.

In the second part of this article, we consider our second hypothesis, which points to the asset side of the balance sheet as the source of the discrepancy between the cyclical properties of small and large banks' NIM. To investigate this hypothesis, we decompose the correlation of the net interest margin into a weighted average of correlations of the average yield on assets (interest income over average assets) and average funding costs (interest expense over average assets).

We find that the cyclical properties of the average yield on assets are virtually identical at small and large banks. We estimate the correlation

[4] This definition of a small bank is also consistent with the definition of a community bank used recently by Federal Deposit Insurance Corporation (2012). It falls in between the traditional absolute community bank size cutoff of $1 billion in assets and the more recently used $10 billion. Our results are fairly robust to the choice of this relative cutoff, as we discuss in Section 5.

between the asset yield and GDP to be positive and virtually of the same magnitude at small and large banks.

Differences exist in the cyclical properties of the funding costs and in the weights with which the asset-yield and funding-cost correlations contribute to the cyclicality of NIM. While funding costs are procyclical among both small and large banks, the small banks' correlation is much lower. In particular, it is lower than the correlation of the average asset yield, consistent with the small banks' NIM being procyclical. Large banks' NIM, in turn, is countercyclical because their funding costs are more strongly procyclical than their average asset yield.

These findings point to the liability side of the small banks' balance sheet as the source of their procyclical profit margin. In this way, these findings seem to be at odds with the view that it is the small banks' close relationships with their borrowers that gives small banks a special role in the intermediation of credit. Rather, these correlations point to the small banks' relationships with their depositors. Consistent with this hypothesis, we show that small banks rely on deposits for their funding significantly more than large banks, while the compositions of small and large banks' asset portfolios are less dissimilar.

Further, we find that the difference in the GDP correlations of asset income and funding costs is magnified among small banks by the relatively high magnitude of the weights with which these correlations contribute to the overall correlation of NIM with GDP. We attribute the magnitude of these weights to the lower volatility of NIM among small banks, which in turn can be accounted for by stronger correlation between small banks' average asset yield and funding costs.

In sum, we view the fact that the cyclical behavior of net interest margins among small banks is very different from that of large banks as evidence supporting the view that small banks fulfill a special role in the intermediation of credit in the US. However, our decomposition of the cyclicality of NIM into cyclical correlations of the asset and liability sides of the balance sheet raises an interesting question about the role of the cost of deposit funding for the behavior of net interest margins at small banks. The relative cyclical insensitivity of the small banks' average funding costs and their strong reliance on deposit funding suggest that the "sticky" properties of deposits documented by Driscoll and Judson (2013), Drechsler et al. (2017), and others could be more pronounced among small banks.

Related literature The paper in the literature most closely related to this article is Aliaga-Díaz and Olivero (2010). They show the countercyclicality of the NIM in the aggregate US banking sector both unconditionally and after conditioning on a set of monetary policy proxy variables. Our analysis here is disaggregated to small and large

banks. We present unconditional correlations in the text and include conditional regression analysis in Appendix B. Our main finding is the procyclicality of the average NIM among small banks.

There is a large empirical literature on the profitability of banks.[5] Most of these studies, however, use aggregate country-level data or conduct cross-country comparisons, e.g., Albertazzi and Gambacorta (2009) and Borio et al. (2017). In particular, Claessens et al. (2017) document low levels of net interest margins in the current low nominal rate environment in a cross section of banks from forty-seven countries. In the US data, Covas et al. (2015) demonstrate the difference in the level of NIM between small and large banks in the last twenty years, as well as the widening of this spread since 2010. They do not systematically investigate the cyclical properties of NIM at small banks, which is our focus here.

In a recent working paper, Haubrich (2018) studies cyclical properties of banks' capital ratios. Similar to our findings on NIM, he finds differences in the cyclical properties of small and large banks' capital buffers.

Another related strand of the empirical literature on banking studies the role banks may play in the amplification of macroeconomic shocks and transmission of monetary policy shocks. Most of this literature, however, studies sector aggregates without disaggregating to small and large banks. In an influential exception, Kashyap and Stein (2000) examine the response of bank lending to changes in the stance of monetary policy. They find support for the so-called bank-lending transmission channel, where banks reduce the supply of loans to firms in the wake of a reduction in the supply of reserves. In particular, they find that this effect is stronger among small banks, whose balance sheets are relatively less liquid.

Drechsler et al. (2017, 2018) document banks' market power in deposit markets by showing weak pass-through of changes in short-term interbank funding rates to deposit rates. They demonstrate a new transmission channel of monetary policy, where deposits flow out of the banking system when bond yields rise, to which banks respond with a contraction in lending. They show that the pass-through is weaker in more concentrated local markets, which is consistent with our findings because, as shown in Meyer (2018), small banks tend to operate in more concentrated local markets.

The article is organized as follows. Section 1 introduces the data used in our analysis. Sections 2 and 3 discuss the aggregate behavior

[5] A separate literature surveyed in Hughes and Mester (2014) uses structural models to measure efficiency in banking.

of the NIM. Section 4 takes a look at the size distribution of banks. Section 5 discusses the cyclical behavior of NIM among small and large banks. Section 6 decomposes the cycliclity of NIM into cyclical properties of the asset and the liability sides of the balance sheet. Section 7 discusses the heterogeneity of small and large banks' funding sources and asset portfolios. Section 8 concludes.

1. DATA

The primary data source for our analysis is the FDIC's Statistics on Depository Institutions (SDI) dataset. The SDI contains quarterly Call Report data for all active FDIC-reporting institutions. Our sample starts in 1992:Q4 and ends in 2016:Q2. The SDI provides the NIM along with its components: interest income, interest expense, and earning assets. This level of detail allows us to compute NIM on earning assets held by small and large institutions, discuss the interest income side and the interest expense side of NIM separately, and show differences in the composition of small and large banks' balance sheets.

To measure the business cycle, we use GDP reported by the Bureau of Economic Analysis (BEA). We convert GDP, and all FDIC data, into 2009:Q1 dollars using the BEA's implicit price deflator.[6]

2. NIM AT THE BANK AND SECTOR LEVEL

At the bank level, NIM in quarter t is defined as

$$NIM_{i,t} := \frac{TII_{i,t} - TIE_{i,t}}{AEA_{i,t}},$$

where $TII_{i,t}$ is the total interest income and $TIE_{i,t}$ is the total interest expense for bank i in period t.[7] Average earning assets, $AEA_{i,t}$, is the average of quarter t's and quarter $t-1$'s total earning assets for bank i. As we see, the NIM represents an average spread between the rates at which a bank earns income on its assets and the rates at which it funds itself.

For the banking sector as a whole, we compute NIM as an average of individual banks' NIM weighted by average earning assets:

$$NIM_t := \sum_i NIM_{i,t} \cdot \frac{AEA_{i,t}}{\Sigma_i AEA_{i,t}} = \frac{\Sigma_i TII_{i,t} - \Sigma_i TIE_{i,t}}{\Sigma_i AEA_{i,t}}. \qquad (1)$$

[6] In Appendix B, we also use the following covariates to check robustness of the correlations we find: the slope of the yield curve, the effective federal funds rate, and the yield on 3-month commercial financial paper. These series are provided by the Federal Reserve Board.

[7] The difference $TII_{i,t} - TIE_{i,t}$ is referred to as net interest income.

Figure 1 NIM for All Banks

Note: Annualized percentage rate calculated by multiplying quarterly NIM by 400.

The second equality above shows that this calculation is equivalent to computing aggregate TII and TIE and dividing their difference by aggregate AEA.

Figure 1 plots NIM for the US banking sector from 1992:Q4 through 2016:Q2. Aggregate NIM has clearly been declining since the beginning of our sample, but with a significant amount of volatility, especially around the time of the Great Recession. The negative trend in NIM seen in our sample is persistent but not permanent. Using annual data that go back to 1970, Ennis (2004) shows a trough of the ratio of aggregate net interest income to bank assets in 1975 and an upward trend between 1975 and 1992, where our sample starts.

Table 1 Average Assets and Interest Income by Source

	Domestic Office Loans	Foreign Office Loans	Trading Accounts	Securities	Financing Receivables	Federal Funds Sold	Balance due from Dep Inst	Other
% of Total Assets	53.18	3.67	4.76	19.18	1.17	3.76	7.17	7.11
% of Interest Income	70.66	5.33	2.43	16.47	1.40	2.07	1.23	0.40

Table 2 Average Liabilities and Interest Expenses by Source

	Domestic Office Deposits	Foreign Office Deposits	Federal Funds Purchased	Demand Notes and Other	Subordinated Notes and Debentures	Other
% of Total Liabilities	66.76	10.40	6.25	9.80	1.14	5.65
% of Interest Expense	55.64	12.10	8.76	20.70	2.77	0.03

Components of the NIM

To provide some insight into the structure of the NIM, in Tables 1 and 2 we present average interest income and interest expense by source, as reported by banks to the FDIC. In addition, we provide the corresponding average assets and liabilities as a fraction of total earning assets and total liabilities. The percentages reported are computed as sample time averages weighted by the total size of the banking sector interest income (TII_t).[8]

As we see in Table 1, loans (domestic and foreign office loans combined) are the largest asset class, followed by securities. In the aggregate, banks in our sample earned 76 percent of their interest income from loans. Loans represented, on average, 57 percent of total assets.[9] Securities generated on average 16 percent of interest income and represented 19 percent of total assets.

As for liabilities, we see in Table 2 that 68 percent of interest expenses arose from deposits (domestic and foreign office combined), which represented 77 percent of total average liabilities.

Next, we turn to the cyclical properties of the NIM.

3. CYCLICAL PROPERTIES OF THE NIM: AGGREGATE

Figure 2 plots the cyclical components of NIM and of log GDP.[10] The NIM deviation from trend ranges from negative 67 basis points in 2009:Q1 to positive 32 basis points in 2010:Q1. The negative correlation between NIM and GDP is rather clearly visible in this picture. Indeed, the estimated Pearson correlation coefficient, presented in Table 3, is -0.30. This estimate is statistically significant with a p-value of 0.003 indicating a countercyclical relationship between NIM and GDP. This finding is similar to Aliaga-Díaz and Olivero (2010, 2011) and Beaubrun-Diant and Tripier (2015), who also find a negative and statistically significant correlation between NIM and GDP.[11]

---

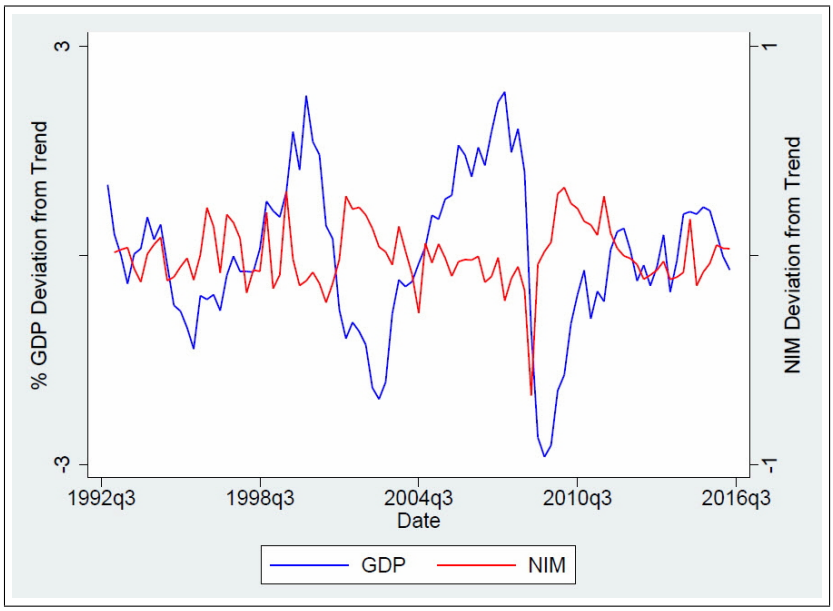
[8] The weighted time average of income from a particular asset class is calculated as $\Sigma_t \frac{II_{c,t}}{TII_t} \cdot \frac{TII_t}{\Sigma_s TII_s}$, where c refers to the asset class considered, e.g., domestic office loans, and t is the calendar-time quarter.

[9] Table C1 provides a glossary of all variables used in the calculations.

[10] We take the Hodrick-Prescott trend out from the level of aggregate NIM and from the log of real GDP. Our results do not depend on this method of detrending. In Appendix A, we discuss robustness to the alternative technique for removing trend that follows Hamilton (2016).

[11] It is worth pointing out that the Great Recession does not drive this result. In the sample ending in 2005, Aliaga-Díaz and Olivero (2010) estimate this correlation at -0.237.

Figure 2 Cyclical Components of GDP and NIM for All Banks



Note: Left scale: log deviations of GDP from trend. Right scale: percentage point deviations of NIM from trend.

In addition, Table 3 presents the correlation between NIM and GDP forwarded and lagged by one, two, four, and eight quarters. These lead-lag correlations show that, in this sample, GDP predicts NIM at durations up to one year, while NIM is not a significant predictor of GDP.

4. THE BANK SIZE DISTRIBUTION

As noted by Cuciniello and Signoretti (2015) and Ennis et al. (2016), among others, the aggregate statistics for the banking sector are primarily driven by large banks. The strong and growing concentration of assets in a small number of large banks is a well-known feature of the bank size distribution (see, e.g., McCord and Prescott [2014]).

Table 3 Correlation of NIM and GDP: All Banks 1992:Q4 through 2016:Q2

	L8Q	L4Q	L2Q	L1Q	GDP	F1Q	F2Q	F4Q	F8Q
NIM	0.0288	-0.4812	-0.5667	-0.4855	-0.3031	-0.1671	-0.1236	0.0173	0.1478
	0.7913	*0.0000*	*0.0000*	*0.0000*	*0.0030*	*0.1095*	*0.2405*	*0.8718*	*0.1743*

Note: Lead-lag correlations reported for one, two, four, and eight quarters. P-value reported in italics.

Figure 3 Fraction of Total Banking Assets Held by the Largest Banks

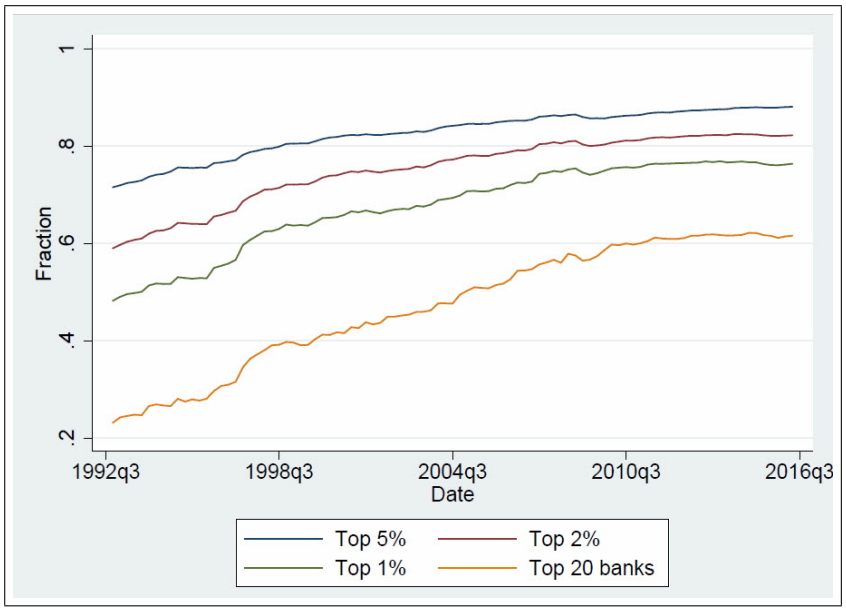


Figure 3 displays the fraction of total assets in the US banking sector held by the largest 5, 2, and 1 percent of banks, as well as by the twenty largest institutions.[12] The growing bank asset concentration is visible by all four of these measures. In 2016:Q2, 88 percent of all sector assets were held by the top 5 percent of institutions, up from 71 percent in 1992:Q4. The largest twenty banks held 62 percent of assets in 2016:Q2, up from 24 percent in 1992.

In this article, we define small banks as those outside the top 5 percent of the asset size distribution in any given time period (i.e., calendar quarter) in our dataset. Large banks are defined as those inside the fifth percentile.[13]

[12] The number of banks reporting to the FDIC decreased from approximately 14,000 in 1992:Q4 to 6,000 in 2016:Q2. Thus, the top 5 percent of banks in 2016 equals about 280 banks. The asset size cutoff thresholds, also at the end of the sample, are approximately \$2.9B, \$9.4B, \$27.3B, and \$123B for, respectively, the top 5 percent, 2 percent, 1 percent, and top twenty banks, all in current dollars.

[13] We also report our correlation results for the top 2 and 1 percentile cutoffs, as well as for the top twenty banks.

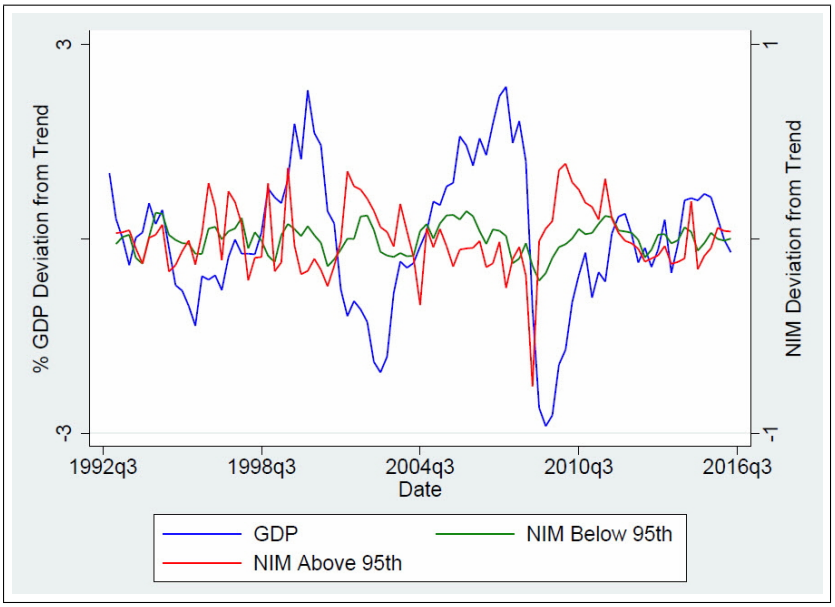
Figure 4 NIM for Large and Small Banks (95 percent cutoff)

Since large banks hold the bulk of the sector's assets, it is clear that aggregate banking sector statistics mirror those of large banks. In particular, we should expect the cyclical properties of NIM for large banks to be similar to those previously reported for the whole banking sector. In the next section, we examine cyclical properties of NIM for small banks and show that they are quite different from the aggregate.

5. CYCLICAL PROPERTIES OF NIM: SMALL VERSUS LARGE BANKS

We start out by presenting the time series of the level of NIM among small and large banks in our sample period. Figure 4 plots NIM levels for banks below and above the 95th percentile of the asset distribution. Note that the level of NIM at small banks is greater than the NIM at large banks in almost all quarters in the sample. This feature may reflect some local monopoly power of small banks in their local areas,

Figure 5 Cyclical Component of GDP and NIM by Bank Size



which could be offset by higher fixed cost, per dollar of assets, of smaller institutions.[14]

Covas et al. (2015) attribute the recent (since 2010) compression of NIM at large banks to smaller declines in funding costs relative to small banks and to declining interest income from short-term trading accounts, which makes up a greater percentage of interest income for large banks than for small banks.

We now turn to the cyclical component of the NIM at small and large banks.[15] Figure 5 plots the cyclical component of the NIM separately for the banks in the bottom 95 percent of the size distribution and in the top 5 percent. As we see in that figure, the time series of the NIM at large banks tracks very closely the aggregate NIM series shown in Figure 2. This indicates that the aggregate negative correlation between the NIM and GDP also holds for large banks. However, the NIM at small banks appears to be much less volatile and its co-movement with GDP is unclear from the graph.

[14] See Drechsler et al. (2018).

[15] For subgroups of banks, the cyclical component on NIM is computed the same way as the aggregate NIM series, i.e., by removing the Hodrick-Prescott trend.

Table 4 Correlation of NIM and GDP by Varying Size Cutoff

	Size	
Percentile Cutoff	Below	Above
95	0.3363	-0.3328
	0.0009	*0.0010*
98	0.2549	-0.3361
	0.0132	*0.0009*
99	0.2287	-0.3346
	0.0266	*0.0010*
Top 20	-0.1139	-0.2948
	0.2742	*0.0076*

Note: P-value reported in italics.

Table 4 presents the estimated Pearson correlation of the NIM with GDP at large and small banks, where small banks are defined by four different size cutoffs. As expected, large banks have a statistically significant and negative correlation with GDP at all four cutoffs. The point estimates of the correlation coefficient are close to -0.3, i.e., similar to the aggregate correlation presented in Table 3.

The correlation of small banks' NIM with GDP, however, is very different. As we see in Table 4, the average NIM in the group of all banks below the top 5, top 2, or top 1 percentile of the size distribution exhibits a statistically significant and positive correlation with GDP. That is, the sign of the correlation is opposite of that of the large banks and the aggregate. If the size cutoff is chosen to be just the twenty largest institutions, the correlation between NIM and GDP for small banks is no longer statistically different from zero.

The results presented in Table 4 highlight our main finding: in contrast to large banks and the aggregate, small banks' NIM is procyclical.

The correlations presented here are unconditional. In Appendix B, we run several regressions to test the robustness of our findings by conditioning on several standard covariates. After controlling for the slope of the yield curve, the level of the federal funds rate, the commercial paper spread, and an indicator of the general stance of monetary policy, our result is unchanged: the NIM at small, i.e., the majority of, banks is procyclical while at large banks, and in the aggregate series, it is countercyclical.

In the remainder of this article, we explore the sources of this difference in the cyclical properties of NIM among small and large banks. In particular, we ask if this difference can be attributed to the cyclical properties of the asset side or the liability side of the banks' balance sheet.

6. CYCLICAL PROPERTIES OF INTEREST INCOME AND INTEREST EXPENSE

In this section, we discuss the cyclicality of NIM in terms of the cyclicality of interest income on the asset side and the cyclicality of the interest expense on the liability side of the balance sheet.

For a given quarter t, we define average asset yield (AAY) among banks as $AAY_t := \frac{\Sigma_i TII_{i,t}}{\Sigma_i AEA_{i,t}}$ and average funding cost (AFC) as $AFC_t := \frac{\Sigma_i TIE_{i,t}}{\Sigma_i AEA_{i,t}}$. From equation (1), we have

$$NIM_t = AAY_t - AFC_t$$

in every quarter t. We want to use this decomposition of the level of NIM to decompose the correlation of NIM with GDP into the corresponding correlations of AAY and AFC with GDP.

Denoting these correlations by, respectively, ρ_N, ρ_A, and ρ_L, we have

$$\rho_N = \frac{\sigma_A}{\sigma_N}\rho_A - \frac{\sigma_L}{\sigma_N}\rho_L, \tag{2}$$

where σ_N, σ_A, and σ_L denote the respective standard deviations.[16] As we see, the NIM correlation is the difference between the correlations of AAY and AFC weighted by standard deviations of AAY and AFC relative to the standard deviation of NIM.

Cyclicality of AAY and AFC among large and small banks

Table 5 presents the components of this decomposition estimated for all banks in the sample as well as separately for large and small banks. We start with the following three observations. First, AAY and AFC are strongly procyclical among both large and small banks. Second, the magnitude of the correlation between AAY and GDP is virtually

[16] This formula follows from $\rho_N = \frac{cov(NIM,GDP)}{\sigma_N\sigma_Y} = \frac{cov(AAY,GDP)}{\sigma_N\sigma_Y} - \frac{cov(AEY,GDP)}{\sigma_N\sigma_Y} = \frac{\rho_A\sigma_A\sigma_Y}{\sigma_N\sigma_Y} - \frac{\rho_L\sigma_L\sigma_Y}{\sigma_N\sigma_Y} = \frac{\sigma_A}{\sigma_N}\rho_A - \frac{\sigma_L}{\sigma_N}\rho_L$, where σ_Y is the standard deviation of GDP.

Table 5 Decomposition of the NIM Correlation for All, Top 5 Percent, and Bottom 95 Percent of Banks

	ρ_N	$\frac{\sigma_A}{\sigma_N}$	ρ_A	$\frac{\sigma_L}{\sigma_N}$	ρ_L
All banks	-0.30	3.25	0.58	3.31	0.66
Large banks	-0.33	2.94	0.57	3.00	0.68
Small banks	0.33	5.11	0.55	5.02	0.49

Note: All p-values (not reported) are smaller than 0.001

the same among large and small banks. Third, AFC is significantly less cyclical at small banks than at large ones.

From these observations, we can conclude that it is the liability side of the balance sheet that drives our main result concerning the difference in the cyclical properties of NIM at large and small banks. With the correlation of the average yield on assets being virtually the same, small banks' NIM is less cyclical than large banks' because small banks' funding costs, AFC, are more stable, i.e, their comovement with the business cycle is weaker. In particular, small banks' NIM is procyclical because their average interest expenses are less strongly correlated with the business cycle than their average asset yields. Among large banks, in contrast, the funding costs are more strongly procyclical than the asset yields, implying countercyclical NIM.

Table 5 presents the decomposition of the NIM correlation for small and large banks defined by the 95 percent cutoff in the bank size distribution. Our conclusion about the relative importance of AFC for explaining the differences in the cyclical properties of NIM between small and large banks holds also for other cutoff thresholds.

Table 6 presents the correlations of NIM, AAY, and AFC at small banks defined by four size cutoffs. The first column, similar to Table 4, shows that the procyclicality of NIM among small banks becomes weaker as the definition of a small bank encompasses larger and larger institutions. The second and third columns show that the correlation of AAY does not change across the four definitions while the correlation of AFC becomes stronger, accounting for the procyclicality of small banks' NIM becoming weaker and eventually, among all but the top twenty banks, becoming insignificant.[17]

[17] The ratios $\frac{\sigma_A}{\sigma_N}$ and $\frac{\sigma_L}{\sigma_N}$, which we do not report here for all size cutoffs, do not depend strongly on the size cutoff.

Table 6 Correlations of NIM, AAY, and AFC Among Small Banks Defined by Four Size Cutoffs

Percentile Cutoff	ρ_N	ρ_A	ρ_L
95	0.33	0.55	0.49
98	0.25	0.56	0.51
99	0.22	0.57	0.53
top 20	-0.11	0.58	0.59

Note: The p-value for ρ_N at top twenty banks is 0.27. All other estimates are significantly different from zero.

Volatility of AAY and AFC among large and small banks

In the decomposition given in Equation (2), the NIM correlation ρ_H depends on the correlations of asset yields and funding costs, but also on the weights $\frac{\sigma_A}{\sigma_N}$ and $\frac{\sigma_L}{\sigma_N}$ that multiply these correlations. The estimated decomposition presented in Table 5 lets us make the following two observations about these weights. First, in each decomposition (i.e., for all, large, and small banks), the weights on the asset yields and funding costs are similar. Clearly, this means the standard deviations of AAY and AFC, σ_A and σ_L, are similar in each group of institutions. Second, in the decomposition for large banks, the weights are slightly smaller than in the decomposition for all banks, while in the decomposition for small banks, the weights are much larger.

The size of these weights, clearly, affects the magnitude of the NIM correlation. In particular, the differences in the small and large banks' weights account for the large difference in the correlation of small and large banks' NIM despite the difference in the correlation of AFC being relatively small (and AAY virtually nonexistent). In this section, we ask if the large weights in the decomposition for small banks are due to high volatility of asset yields and funding costs or due to low volatility of small banks' NIM.

Table 7 provides estimates of the standard deviation, i.e., volatility, of NIM, AAY, and AFC for all banks in the sample as well as for large and small banks separately.[18] Clearly, small banks have lower volatility on all three measures. This leads us to conclude that the weights for

[18] In this section, we present the results for largest 5 percent and smallest 95 percent of institutions. The estimates using the top 2 percent and the top 1 percent size cutoffs are similar.

Table 7 Standard Deviations of NIM and its Components for All, Top 5 Percent, and Bottom 95 Percent of Banks

	σ_N	σ_A	σ_L
All banks	0.37	1.21	1.24
Large banks	0.44	1.28	1.31
Small banks	0.19	0.97	0.95

Note: Values reported multiplied by 10^3. All p-values (not reported) are smaller than 0.001

the small banks are large because their NIM is less volatile, despite the fact that the standard deviations of their asset yields and funding costs are lower.

Further, Table 7 shows that the standard deviation of NIM equals about a third of the standard deviation of asset yields or funding costs among large banks but only about a fifth among small banks. This suggests that asset yields and funding are more strongly correlated with each other for small banks than they are for large ones. Using the identity

$$\sigma_N^2 = \sigma_A^2 + \sigma_L^2 - 2\rho_{AL}\sigma_A\sigma_L,$$

we can compute the correlation between average asset yields and average funding costs, ρ_{AL}, from the estimates of the standard deviations σ_N, σ_A, and σ_L given in Table 7. We find that, indeed, AAY and AFC are more closely correlated at small banks. The computed correlation is $\rho_{AL} = 0.955$ for all banks, 0.942 for large banks, and 0.981 for small banks.

7. BALANCE SHEET COMPOSITION AT SMALL AND LARGE BANKS

Having shown differences in the large and small banks' cyclical properties of NIM, AAY, and AFC, we now ask if these differences are reflected in the composition of the average balance sheet of small and large banks.[19]

Section 2 introduced the main classes of assets and liabilities held by banks. On the one hand, it is possible that small and large banks'

[19] Debbaut and Ennis (2014) provide a detailed description of the average balance sheet of large US bank-holding companies between 2005 and 2011.

Table 8 Breakdown of Assets by Class as a Percent of Total Assets

Asset Cutoff		Domestic Office Loans	Foreign Office Loans	Trading Accounts	Securities	Other
Aggregate		53.18	3.67	4.76	19.18	19.21
95	Below	64.01	0.05	0.04	23.37	12.53
	Above	51.10	4.37	5.66	18.37	20.50
98	Below	64.16	0.11	0.08	23.11	12.54
	Above	49.89	4.74	6.16	18.00	21.21
99	Below	63.75	0.16	0.14	22.86	13.09
	Above	48.60	5.19	6.75	17.59	21.87
Top 20	Below	63.00	0.43	0.55	21.25	14.77
	Above	43.53	6.86	8.89	17.15	23.57

Note: Numbers reported are percents, averaged over time.

NIM, AAY, and AFC differ in their business cycle correlations because assets and liabilities held by large and small banks have inherently different business cycle properties. On the other hand, it is possible that assets and liabilities of small and large banks are homogeneous but are simply held by small and large banks in different proportions. Although a full investigation of these alternatives is beyond the scope of this article, we take a step toward it by presenting the average balance sheet composition of small and large banks.

Table 8 presents average asset portfolios for large and small banks over our sample period, where small banks are defined by four separate size cutoffs.[20] The asset classes shown are domestic office loans, foreign office loans, trading accounts, securities, and other. The other class consists of balances from depository institutions, federal funds sold, and lease financing receivables.[21] We observe that small banks hold a larger proportion of domestic office loans and securities than large banks, which in turn hold relatively more foreign office loans, trading accounts, and other. The differences in the asset composition, however, are not large. The average bank allocates 53 percent of assets to domestic office loans, the largest asset class across the board, while the average small bank (below the 95 percent cutoff) and the average large bank (top 5

[20] The small/large banks' average portfolio is a size-weighted sample-period average of the portfolios of all banks below/above the respective cutoff.

[21] In Table 1, these subclasses are broken out separately but only reported for the aggregate (i.e., all banks) average portfolio.

percent) allocate, respectively, 64 and 51 percent of assets to domestic office loans. Further, these allocations do not vary with the size cutoff by much.

Differences between large and small banks are more pronounced on the liabilities side of the balance sheet, presented in Table 9. Looking at the 95 percent size cutoff, domestic office deposits account for 91 percent of total liabilities for small banks and only 62 percent for large banks. Large banks hold significant amounts of foreign office deposits, which are virtually unheld by small banks. Still, total deposits constitute a much smaller fraction of liabilities at large banks than at small. Further, the differences in the composition of liabilities between small and large banks depend strongly on the size cutoff. The average bank in the top percentile of the size distribution allocates less than 59 percent of liabilities to the domestic deposit class.

These observations lead us to conclude that balance sheet composition is an important factor behind the differences in the cyclical properties of asset yields, liability costs, and, consequently, net margins at small and large banks. Section 6 shows that small and large banks differ in the cyclical properties of their funding costs but are virtually homogeneous with respect to the cyclical properties of their asset returns. Correspondingly, Table 9 shows significant differences between small and large banks in the composition of liabilities, while Table 8 shows that the differences in the composition of assets are much smaller. While composition does not explain everything, these observations suggest it is important. In particular, the high proportion of funding obtained from domestic deposits seems to be important for the lower cyclicality of small banks' cost of funding and, therefore, also for the observed procyclicality of their NIM.

Table 9 Breakdown of Liabilities by Source as a Percent of Total Liabilities

Asset Cutoff		Domestic Office Deposits	Foreign Office Deposits	Federal Funds Purchased	Demand Notes	Subordinated Notes and Debentures	Other
Aggregate		66.76	10.40	6.25	9.80	1.14	5.65
95	Below	91.09	0.14	2.20	5.44	0.04	1.09
	Above	62.10	12.37	7.03	10.63	1.35	6.52
98	Below	88.23	0.30	3.33	6.67	0.10	1.37
	Above	60.35	13.41	7.12	10.73	1.45	6.94
99	Below	85.14	0.53	4.44	8.08	0.20	1.61
	Above	58.82	14.66	7.03	10.54	1.55	7.40
Top 20	Below	78.56	2.09	5.89	10.58	0.58	2.30
	Above	55.25	18.50	6.60	9.03	1.69	8.93

Note: Numbers reported are percents, averaged over time.

8. CONCLUSIONS

In this article, we analyze the comovement of bank NIMs with the business cycle both at the sector level and disaggregated for large and small banks. We find that the cyclical component of NIM among large and small banks responds differently to business cycle fluctuations. Specifically, while the average NIM at large banks (the top 5 percent of the size distribution by assets) is negatively correlated with the business cycle, the average NIM at small banks (the bottom 95 percent of the size distribution) is positively correlated with GDP. Due to the high degree of concentration of asset holdings in the banking sector, the aggregate, sector-wide correlation is nearly the same as that of the largest 5 percent of institutions, standing at about -0.3. The correlation computed for the bottom 95 percent banks is of the opposite sign and nearly the same magnitude, i.e., it stands at about +0.3. To our knowledge, our findings on small banks are novel to the literature. In an Appendix, we show our results to be robust to the detrending method and introducing controls for the stance of monetary policy.

We consider this finding to be broadly supportive of the widely held view that small banks occupy a special role in the intermediation of credit. However, when we decompose the cyclical properties of NIM into the asset and liability sides of the balance sheet, we find the liability side to be the driver. This points us to attribute the small banks'special role to their ability to keep their funding costs relatively insensitive to the business cycle rather than their ability to extract business-cycle-relevant information from their long-term relationships with borrowers.

APPENDIX

A: DETRENDING METHOD

In this article, we detrend our time series data with the Hodrick-Prescott (HP) filter. In this section of the Appendix, we consider an alternative detrending method proposed in Hamilton (2016), which defines the cyclical component as the deviation of actual data from their predicted values based on a linear projection from their own lags. For quarterly data, we follow Hamilton (2016) in using an eight-quarter ahead projection from a regression including four lags. Our results are robust to using this definition of the cyclical component of GDP and NIM. The estimated aggregate correlation coeffi cient is -0.4484 with the p-value of 0.0000, which shows a stronger negative correlation than the one we report in Table 3.

Similarly, when using the Hamilton method of detrending and recomputing the correlation between the cyclical components of GDP and NIM for small and large banks separately, i.e., the correlations presented in Table 4, we get estimates consistent with those obtained using the HP filter up to the 98th percentile asset cutoff, which indicates robustness of our results to the method of detrending. Likewise, our results concerning the asset- and liability-adjusted measures of NIM are robust. Detailed estimates of these correlations with the Hamilton detrending method are available upon request.

B: COVARIATES

In this section, we use a multivariate regression to check robustness of our correlation results to the stance of monetary policy. We consider the following four control variables: the slope of the yield curve, measured as the difference in the yields on 10-year and 3-month Treasury securities; the level of the effective federal funds rate; the spread on 3-month financial commercial paper over the 3-month Treasury yield;

and a simple indicator variable marking periods of contractionary, expansionary, and neutral monetary policy.

The policy-stance indicator is constructed, similar to Ennis et al. (2016), via two dummy variables: one indicating periods of increasing rates and the other periods of decreasing rates. In addition, the post-2007 period of zero nominal rates and large reserves outstanding is classified as a loose monetary policy stance. The effective federal funds rate path is presented in Figure B1 along with color-coded periods of contractionary, expansionary, and neutral monetary policy stances.

We run two regression specifications, with each specification consisting of independent regressions for large and small banks. In the first specification, presented in Table B1, we regress the cyclical component of the NIM on the cyclical component of GDP, one of our four monetary policy proxy variables, and an interaction term.[22] In all cases, the coefficient on the GDP variable is negative for large banks and positive for small banks, confirming our unconditional correlation results. When using the effective federal funds rate or the commercial paper spread, these coefficients are statistically significant at the 5 percent level. When using the slope of the yield curve or our dummy variables as proxies for the stance of monetary policy, some coefficients become statistically insignificant at the 5 percent level.

In our second specification, presented in Table B2, we regress NIM on GDP and four lags of the proxy variable of interest. In all cases for small banks, the coefficient on GDP is positive and statistically significant at the 1 percent level, indicating that the positive correlation seen earlier is robust to controlling for monetary policy. Regarding large banks, the coefficients on GDP are all negative and at least statistically significant at the 10 percent level.[23] The coefficient is significant at the 5 percent level when using the dummy variables for periods of increasing (decreasing) federal funds rate and is significant at the 1 percent level when using the effective federal funds rate and the spread on commercial financial paper.

Overall, these results indicate to us that the conclusions reached in the main body of the paper are statistically robust to including controls for the stance of monetary policy.

---

[22] As noted by Borio et al. (2017), among others, the impact of interest rates on the NIM can be nonlinear. By including the interaction term, we can capture potential changes in the slope coefficient.

[23] This conclusion is not overturned by the positive coefficient on the interaction term with the credit spread variable in specification (6) of the regression because the credit spread variable contains only the cyclical component of the spread, and, thus, it is very close to zero on average with small dispersion.

Table B1 Regression of Cyclical NIM on Cyclical GDP and other Covariates with Interaction Terms

Variables	(1) Small Banks	(2) Large Banks	(3) Small Banks	(4) Large Banks	(5) Small Banks	(6) Large Banks	(7) Small Banks	(8) Large Banks
GDP	1.820*	-0.490	3.124**	-5.721**	2.291***	-5.084***	0.430	-6.520**
	(0.970)	(2.183)	(1.276)	(2.498)	(0.741)	(1.368)	(0.837)	(2.883)
Yield Curve Slope	-0.000466	0.0341**						
	(0.00831)	(0.0161)						
Yield Curve Slop #GDP	0.264	-1.567						
	(0.468)	(1.081)						
FFR			0.00121	-0.00520				
			(0.00299)	(0.00755)				
FFR #GDP			-0.328	0.307				
			(0.267)	(0.613)				
ComPprFFRSpreadcyc					-0.00281	-0.183		
					(0.0415)	(0.123)		
ComPprFFRSpreadcyc #GDP					4.186	20.62**		
					(3.342)	(8.274)		
1.FFR increase							0.00239	-0.0509
							(0.0232)	(0.0552)
2.FFR decrease							-0.0361**	-0.00832
							(0.0169)	(0.0428)
1.FFR increase #GDP							3.668**	5.867
							(1.790)	(4.572)
2.FFR decrease #GDP							0.629	0.857
							(1.460)	(3.534)
Constant	0.00298	-0.0732**	0.000308	0.0106	-0.000734	-0.00571	0.0154	0.00956
	(0.0165)	(0.0330)	(0.0112)	(0.0260)	(0.00753)	(0.0160)	(0.0130)	(0.0356)
Observations	94	94	94	94	94	94	94	94
R-squared	0.112	0.148	0.122	0.116	0.130	0.251	0.211	0.124

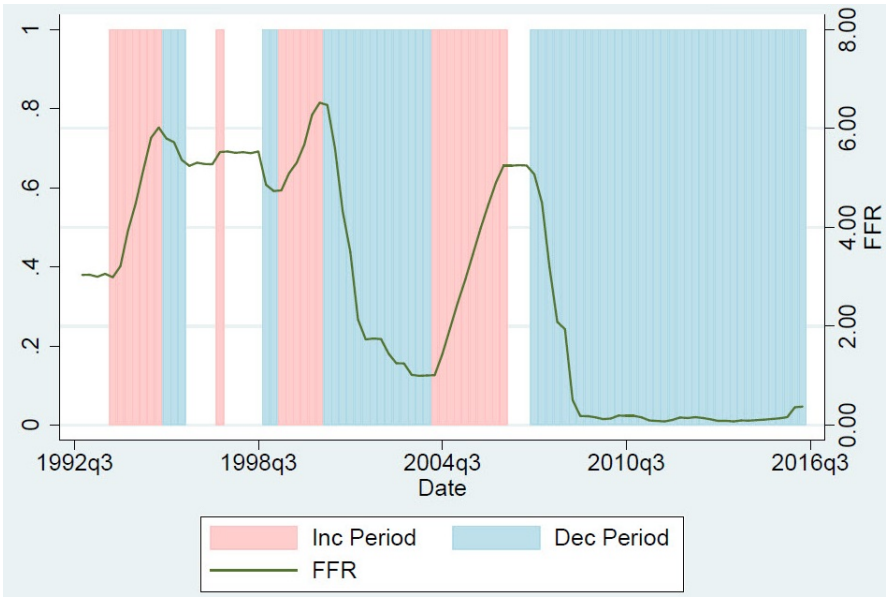
Note: Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table B2 Regression of Cyclical NIM on Cyclical GDP and other Lagged Covariates

VARIABLES	(1) Small Banks	(2) Large Banks	(3) Small Banks	(4) Large Banks	(5) Small Banks	(6) Large Banks	(7) Small Banks	(8) Large Banks
GDP	3.445***	-3.129*	2.041***	-4.354***	2.177***	-6.039***	2.110**	-4.035**
	(0.826)	(1.679)	(0.558)	(1.487)	(0.715)	(1.448)	(0.806)	(1.614)
L.Yield Curve Slope	-0.0218	0.0486						
	(0.0192)	(0.0457)						
L2.Yield Curve Slope	0.0638**	0.0544						
	(0.0307)	(0.0768)						
L3.Yield Curve Slope	-0.0374	-0.109						
	(0.0304)	(0.0659)						
L4.Yield Curve Slope	0.0175	0.0424						
	(0.0198)	(0.0498)						
L.FFR			0.104***	-0.000508				
			(0.0270)	(0.0877)				
L2.FFR			-0.118**	0.0302				
			(0.0578)	(0.165)				
L3.FFR			-0.0175	-0.0947				
			(0.0581)	(0.0986)				
L4.FFR			0.0297	0.0584				
			(0.0325)	(0.0916)				
L.ComPprFFRSpreadcyc					-0.0266	-0.229		
					(0.0662)	(0.141)		
L2.ComPprFFRSpreadcyc					-0.00719	-0.0833		
					(0.0583)	(0.106)		
L3.ComPprFFRSpreadcyc					0.00869	0.0656		
					(0.0422)	(0.218)		
L4.ComPprFFRSpreadcyc					-0.0424	-0.101		
					(0.0433)	(0.182)		
L.FFR_increase							0.000900	-0.000568
							(0.0374)	(0.0780)
L.FFR_decrease							-0.0903***	-0.0860
							(0.0260)	(0.0831)
L2.FFR_increase							0.0643	0.0264
							(0.0402)	(0.0644)
L2.FFR_decrease							0.0761**	0.0153
							(0.0327)	(0.0868)
L3.FFR_increase							-0.0181	-0.116
							(0.0298)	(0.0912)
L3.FFR_decrease							0.00583	0.0139
							(0.0325)	(0.179)
L4.FFR_increase							0.0234	0.0445
							(0.0292)	(0.113)
L4.FFR_decrease							0.0432	0.0894
							(0.0278)	(0.193)
Constant	-0.0396**	-0.0651	0.00587	0.0160	1.02e-05	-0.00243	-0.0358*	-0.00474
	(0.0171)	(0.0436)	(0.00971)	(0.0252)	(0.00775)	(0.0168)	(0.0205)	(0.0549)
Observations	91	91	91	91	91	91	91	91
R-squared	0.222	0.181	0.342	0.129	0.127	0.207	0.337	0.207

Note: Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Figure B1 Effective Federal Funds Rate



C: GLOSSARY

Table C1 Glossary of Variables

	Variable Name	FDIC Variable ID	Formula
Assets	Total Loans and Leases	lnlsgr	
	Total Loans & Leases held in Foreign Offices	lnlsgrf	
	Lease Financing Receivables	ls	
	Lease Fin. Receiv. held in Foreign Offices	lsfor	
	Total Securities	sc	
	Trading Account Assets	trade	
	Cash Balances due from Depository Instit.	chbal	
	Federal Funds Sold and Reverse Repo	frepo	
	Total Assets	asset	
	Earning Assets	ernast	
	Domestic Office Loans		lnlsgr-ls-(lnlsgrf-lsfor)
	Foreign Office Loans		lnlsgrf-lsfor
Liabilities	Deposits held in Domestic Offices	depdom	
	Deposits held in Foreign Offices	depfor	
	Subordinated Debt	subnd	
	Other Borrowed Funds	idobrmtg	
	Federal Funds Purchased and Repo	frepp	
Interest Income (II)	II from Domestic Office Loans	ilndom	
	II from Foreign Office Loans	ilnfor	
	II from Lease Financing Receivables	ils	
	II from Securities	isc	
	II from Trading Account Assets	itrade	
	II From Cash Balances due from Dep Insts	ichbal	
	II from Federal Funds Sold & Reverse Repo	ifrepo	
	Other Interest Income	iothii	
	Total Interest Income	intinc	
Interest Expense (IE)	IE from Deposits held in Domestic Offices	edepdom	
	IE from Deposits held in Foreign Offices	edepfor	
	IE from Subordinated Notes & Debentures	esubnd	
	IE from Other Borrowed Funds	ettlotmg	
	IE from Federal Funds Purchased & Repo	efrepp	
	Total Interest Expense	eintexp	
	Net Interest Income	nim	intinc-eintexp
	Net Interest Margin	nimy	$\frac{\text{nim}}{\text{ernast}}$

REFERENCES

Albertazzi, Ugo, and Leonardo Gambacorta. 2009. "Bank Profitability and the Business Cycle." *Journal of Financial Stability* 5 (December): 393–409.

Aliaga-Díaz, Roger, and María Pía Olivero. 2010. "Is There a Financial Accelerator in US Banking? Evidence from the Cyclicality of Banks' Price-Cost Margins." *Economics Letters* 108 (August): 167–171.

Aliaga-Díaz, Roger, and María Pía Olivero. 2011. "The Cyclicality of Price-Cost Margins in Banking: An Empirical Analysis of Its Determinants." *Economic Inquiry* 49 (January): 26–46.

Beaubrun-Diant, Kevin E., and Fabian Tripier. 2015. "Search Frictions, Credit Market Liquidity and Net Interest Margin Cyclicality." *Economica* 82 (January): 79–102.

Borio, Claudio, Leonardo Gambacorta, and Boris Hofmann. 2017. "The Influence of Monetary Policy on Bank Profitability." *International Finance* 20 (Spring): 48–63.

Claessens, Stijn, Nicholas Coleman, and Michael Donnelly. 2017. "'Low-for-Long' Interest Rates and Banks' Interest Margins and Profitability: Cross-Country Evidence." Board of Governors of the Federal Reserve System International Finance Discussion Papers 1197 (February).

Covas, Francisco B., Marcelo Rezende, and Cindy M. Vojtech. 2015. "Why are Net Interest Margins of Large Banks so Compressed?" Board of Governors of the Federal Reserve System FEDS Notes (October 5).

Cuciniello, Vincenzo, and Federico M. Signoretti. 2015. "Large Banks, Loan Rate Markup, and Monetary Policy." *International Journal of Central Banking* 11 (June): 141–77.

Debbaut, Peter, and Huberto M. Ennis. 2014. "Large U.S. Bank Holding Companies During the 2007-09 Financial Crisis: An Overview of the Data." Federal Reserve Bank of Richmond *Economic Quarterly* 100 (Second Quarter): 113–57.

Drechsler, Itamar, Alexi Savov, and Philipp Schnabl. 2017. "The Deposits Channel of Monetary Policy." *Quarterly Journal of Economics* 132 (November): 1819–76.

Drechsler, Itamar, Alexi Savov, and Philipp Schnabl. 2018. "Banking on Deposits: Maturity Transformation Without Interest Rate Risk." Working Paper (April).

Driscoll, John C., and Ruth A. Judson. 2013. "Sticky Deposit Rates." Federal Reserve Board Finance and Economics Discussion Series 2013-80 (October).

Ennis, Huberto M. 2004. "Some Recent Trends in Commercial Banking." Federal Reserve Bank of Richmond *Economic Quarterly* 90 (Spring): 41–61.

Ennis, Huberto M., Helen Fessenden, and John R. Walter. 2016. "Do Net Interest Margins and Interest Rates Move Together?" Federal Reserve Bank of Richmond *Economic Brief* 16-05 (May).

Federal Deposit Insurance Corporation. 2012. "FDIC Community Banking Study" (December).

Hamilton, James D. 2016. "Why You Should Never Use the Hodrick-Prescott Filter." UCSD Working Paper (June).

Haubrich, Joseph G. 2018. "How Cyclical Is Bank Capital?" Federal Reserve Bank of Cleveland Working Paper 15-04R (February).

Hughes, Joe, and Loretta J. Mester. 2014. "Measuring the Performance of Banks: Theory, Practice, Evidence, and Some Policy Implications." In *Oxford Handbook of Banking*, Second Edition (2 ed.), edited by Allen N. Berger, Philip Molyneux, and John O.S. Wilson. Oxford: Oxford University Press.

Kashyap, Anil K., and Jeremy C. Stein. 2000. "What Do a Million Observations on Banks Say About the Transmission of Monetary Policy?" *American Economic Review* 90 (June): 407–28.

Keeton, William, George A. Kahn, Linda Schroeder, and Stuart Weiner. 2003. "The Role of Community Banks in the U.S. Economy." Federal Reserve Bank of Kansas City *Economic Review* (Second Quarter): 15–43.

McCord, Roisin, and Edward S. Prescott. 2014. "The Financial Crisis, the Collapse of Bank Entry, and Changes in the Size Distribution of Banks." Federal Reserve Bank of Richmond *Economic Quarterly* 100 (First Quarter): 23–50.

Mester, Loretta J. 2017. "Perspectives on the Economic Outlook and Banking Supervision and Regulation." Speech at the Community Bankers Association of Ohio Annual Convention, Cincinnati, Ohio, August 2.

Meyer, Andrew P. 2018. "Market Concentration and Its Impact on Community Banks." Federal Reserve Bank of St. Louis *Regional Economist* 26 (First Quarter).

Powell, Jerome H. 2016. "Trends in Community Bank Performance Over the Past 20 Years." Speech at the "Community Banking in the 21st Century" Fourth Annual Community Banking Research and Policy Conference, St. Louis, Missouri, September 29.

Economic Quarterly—Volume 104, Number 1—First Quarter 2018—Pages 35–52

Self-Insurance and the Risk-Sharing Role of Money

Tsz-Nga Wong

"The inherent vice of capitalism is the unequal sharing of blessings."
-*Winston Churchill address to the House of Commons, 1945*

Money is well-acknowledged as a social construct to overcome the lack of coincidence of wants in a society. On top of its transactional role, the literature of monetary theory has also pointed out the role of self-insurance following the fact that money can be a vehicle for precautionary saving.[1] Less is mentioned about money as a social construct to promote risk sharing among individuals. In this review, I will provide a simple mathematical model to illustrate this new insight. The model can be solved in closed form with paper and pencil. The material here is borrowed from some recent studies.[2]

In my "toy model" there are a lot of agents; each receives a constant flow of endowment but also faces uncertainty about the timing and the number of "liquidity shocks": when the shock hits, the agent needs to spend a big chunk of endowment. In real life, the liquidity shock

I have benefited from the comments of John Weinberg, Felix Ackon, Felipe Schwartzman, and Nicholas Trachter. The views expressed in this article are those of the author and do not necessarily represent the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] See the seminal paper of Bewley (1980) and its corrigendum Bewley (1983) for the error in the existence proof pointed out by Hellwig (1982). It is well-known that precautionary motive can lead to efficiency loss due to excessive saving; see Aiyagari (1994) and Davila et al. (2012).

[2] I select and simplify some results from Rocheteau et al. (forthcoming); see therein for a general treatment. The discrete time setting in the introduction comes from Rocheteau et al. (2015). The discussion of perfect self-insurance in a monetary economy borrows from Wong (2016). See Lagos et al. (2017) for a recent survey on the literature of monetary theory; see Rocheteau and Nosal (2017) for a textbook introduction. The general welfare property of money with the risk-sharing role is still an open question; see Wallace (2014) for a conjecture.

DOI: https://doi.org/10.21144/eq1040102

captures unexpected expenditures like car accidents and medical expenses (a top reason for bankruptcy in America). Liquidity shock is idiosyncratic, so it comes early for some agents but late for others; some agents experience only a few shocks over a long period of time, but some agents are hit often.[3] Agents can store their endowments, thus in principle, agents always can self-insure against their liquidity shocks with private storage. In the design of the toy model, I do not force agents to use money—agents choose between money and storage—so the model has the potential to explain the fundamental reason for the emergence of money rather than merely begging the question. Furthermore, there is always coincidence of wants: every agent owns and consumes the same goods. If money is socially useful in this model, then it must be due to other reasons—in particular, I will show mathematically that when individuals hold money, it helps share others' liquidity risks.

1. MODEL

As a benchmark, I begin an economy without money. The economy is populated by a unit measure of risk-neutral agents. I first consider the time discrete and the horizon infinite. I need an infinite horizon for money to circulate in the latter section; otherwise, if the economy will terminate at a fixed date, then no agent will exchange his goods for money on the date before (since money has no future use), and hence on the date before before—by backward induction, no one will ever hold money in a finite horizon.[4]

Timing. The timing follows the literature of banking theory (e.g. Diamond and Dybvig [1983]). Each period has two stages. In the first stage, some agents are hit by liquidity shocks: when the shock hits, the agent needs to cover an expense of $\bar{y} > 0$ units of goods, otherwise she is subject to an increasing loss for the amount falling short. Mathematically, it is conveniently captured by a quadratic loss function, $L(y) = -0.5A \min(y - \bar{y}, 0)^2$, where $A > 0$ captures the marginal loss and y is the amount of goods the agent can raise to cover the liquidity shock. I simply refer to y as early consumption as in the literature. Liquidity shocks are i.i.d. and occur with probability $\alpha\Delta > 0$. In the second stage, each agent receives $h\Delta$ units of goods and consumes

[3] In a neat setting, Scheinkman and Weiss (1986) cleverly assume shocks are stochastic but alternate between two types of agent. This captures the redistribution effect, still the model remains highly tractable. Lippi et al. (2015) found that with this redistribution channel, the optimal monetary policy can be countercyclical.

[4] The backward induction argument hinges on the discreteness of time. In the continuous time, money can circulate in a finite horizon, but the monetary equilibrium, if it exists, is typically nonstationary. It is out of the scope of this review.

$c\Delta$, pro rata to the length of period. In the absence of technologies enforcing and monitoring actions, debt contracts, either across stages or across periods, are not incentive feasible. However, agents can self-insure against the liquidity shocks by storing any unconsumed goods in the second stage, subject to the depreciation rate δ. In sum, the period-utility function is $\varepsilon L(y) + c\Delta$, where $\varepsilon \in \{0, 1\}$ is the indicator of the liquidity shock with $\Pr(\varepsilon = 1) = \alpha\Delta$. The discount factor across periods is $\exp(-r\Delta)$. The economy starts in the second stage of period 0. The order of events does not quite matter when time becomes continuous.

Value. The choices of action are made with the following consideration. Contingent on the history of liquidity shocks, $\mathcal{H}^t \equiv \{\varepsilon_1, \varepsilon_2, ...\varepsilon_t\}$, agents choose the level of consumption and storage prepared for potential early consumption. Denote such a contingent plan as the functions, C and Y, of history such that $c_t = C\left(\mathcal{H}^t\right)$ and $y_t = Y\left(\mathcal{H}^t\right)$. Also, denote a_t as the level of storage brought to the second stage of t. Starting in the second stage of t, the value of storage is the expected discounted sum of the future utility and loss, which is given by

$$\begin{aligned} V_t &= \max_{C,Y} \left\{ C\left(\mathcal{H}^t\right)\Delta + \mathbb{E}_t \sum_{s=t+1}^{\infty} e^{-r\Delta s} \left\{ \varepsilon_s L\left[Y\left(\mathcal{H}^s\right)\right] + C\left(\mathcal{H}^s\right)\Delta \right\} \right\}, \\ &\quad \text{given } a_t. \end{aligned}$$

With bounded endowments and depreciating storage, the level of storage is always bounded. Together with the fact that the loss function, $L(y)$, is bounded, the value of storage, V_t, cannot explode to infinity (positive or negative), and hence it satisfies the asymptotic boundary condition that $\lim_{s\to\infty} \mathbb{E}_t \exp(-r\Delta s) V_s = 0$ almost surely. Then there exists a value function, $V(a)$, such that the value of storage, V_t, can be recursively expressed by the following Bellman equation

$$V_t = V(a) = \max_{c,a',y} \left\{ c\Delta + e^{-r\Delta} \left\{ \alpha\Delta \left[L(y) + V\left(a' - y\right) \right] + (1 - \alpha\Delta) V\left(a'\right) \right\} \right\} \tag{1}$$

$$\begin{aligned} \text{s.t. } a_t &= a, \\ a' &= (1 - \delta\Delta)(h\Delta - c\Delta + a), \\ c &\geq 0, \\ y &\in \left[0, a'\right]. \end{aligned}$$

Denote the solutions to (1) as $c(a)$ and $y(a)$. Instead of making use of the entire history, $\mathcal{H}^t$, the current level of storage, $a_t = a$, is sufficient information for decision-making such that the agent's choices are given by $c_t = c(a)$ and $y_t = y(a)$. This recursive structure will be useful for analyzing the agent's infinite-horizon problem.

The economic meaning of the Bellman equation (1) is as follows. According to (1), the typical agent chooses her consumption, c, the next-period storage (after depreciation applies), a', and the early consumption, y, in order to maximize her expected discounted continuation value in the next period. The budget identity specifies that the next-period storage is equal to the current income net of consumption multiplied by the gross depreciation factor. With probability $\alpha\Delta$, the agent receives a liquidity shock for early consumption. The maximal early consumption the agent can draw on is her entire level of storage, i.e. $y \leq a'$. The choice of y balances the trade-off between early consumption and leaving some storage for the future. With probability $1 - \alpha\Delta$, the agent is not hit by a liquidity shock and enters the second stage with a'. In sum, to self-insure against the liquidity shocks, the agent wants to maintain a sufficient level of storage at the cost of giving up current consumption and wasting resources to depreciation.

Self-insurance. Agents who are frequently hit by liquidity shocks consume less, hold less storage, and, in a vicious cycle, can become more vulnerable to future liquidity shocks. Should they share the liquidity risks, if possible? Actually, when their endowment, $h\Delta$, is sufficiently large, agents can achieve perfect self-insurance with constant storage. In this case, the marginal value of storage is constant at $V'(a) = 1$ for all a, and there is no need to share liquidity risks across agents in the economy. This case is well-studied in the literature after Lagos and Wright (2005).[5] Here, I am interested in the case otherwise, but then the solution to (1) will feature occasionally binding constraints. To ease the analysis, I take the period length Δ to zero and the time becomes continuous. In this case, the flow of endowment, $h\Delta$, is so small compared to the magnitude of the liquidity shock, $\bar{y}$, that agents can never achieve perfect self-insurance. The rate of storage in the continuous time is given by

$$\dot{a} \equiv \lim_{\Delta \to 0} \frac{a' - a}{\Delta} = h - c - \delta a.$$

[5] The appearance of perfect self-insurance does not necessarily depend on the risk-neutral assumption. Wong (2016) illustrates examples with strictly concave utility functions. See the discussion therein for details. The key to perfect self-insurance is that agents can reach the target level of storage immediately after shocks, either by adjusting labor supply, consumption, portfolio, or a combination of these. Of course, achieving perfect self-insurance does not mean the first best.

In the continuous time, the value function solves the following Hamilton-Jacob-Bellman (HJB) equation instead[6]:

$$rV(a) = \max_{c\geq 0, y\in[0,a]} \left\{ c + V'(a)(h - c - \delta a) + \alpha \left[L(y) + V(a-y) - V(a) \right] \right\}, \tag{2}$$

where $V'(a)$ denotes the first derivative of $V(a)$. Unlike the case of perfect self-insurance, generically $V'(a)$ is varying in a.[7] The economic meaning of the HJB equation is as follows. When agents maximize their utility, the flow of the agent's value, $rV(a)$, is equal to the flow of the consumption utility, c, the rate the value changes due to storage, $V'(a)\dot{a}$, and the expected change in the value due to the liquidity shocks, $\alpha\left[L(y) + V(a-y) - V(a)\right]$.

For the later derivation of the distribution, define $\Phi(a)$ as the maximal level of storage such that after a liquidity shock the agent will keep a units of storage, given by

$$\Phi(a) \equiv \max d \text{ s.t. } d = y(d) + a, \tag{3}$$

which means that the preshock storage, Φ, is equal to the sum of the early consumption, $y(\Phi)$, and the postshock storage, a. In general, there can be multiple levels of preshock storage that lead to the same postshock storage; for example, when the agent always draws her entire storage for the early consumption, any level of preshock storage will lead to zero postshock storage. That is why in the definition of Φ, I always pick the maximal one. I adopt the convention that $\Phi(a) = \infty$ if no solution to (3) exists.

Closed-form solutions. In general, the value function, $V(a)$, is the solution to the delayed differential equation (DDE), (2), satisfying the asymptotic boundary condition. Here, it is straightforward to verify

---

[6] Heuristically, the HJB equation can be derived as follows. Rearranging (1), I have

$$-\frac{V(a') - V(a)}{a' - a}\frac{a' - a}{\Delta} = \max\left\{ c + e^{-r\Delta}\alpha\left[L(y) + V(a' - y) - V(a')\right] + \frac{e^{-r\Delta} - 1}{\Delta}V(a') \right\}.$$

When $\Delta \to 0$, I have $a' \to a$. The right side above becomes $\max\{c + \alpha\left[L(y) + V(a-y) - V(a)\right]\} - rV(a)$. The first term on the left side becomes $-V'(a)$ at the limit. The second term converges to $h - c - \delta a$. Collecting these terms, I have the HJB equation (2). It is only a heuristic derivation because it begs the question of proving the existence of $V(a)$ and, more challenging, that $V(z)$ is twice differentiable for the HJB to be well-defined. The complete proof is given by Rocheteau et al. (forthcoming), which utilizes the techniques of the viscosity solution.

[7] Suppose $V'(a) = v$ for all a. The HJB equation becomes

$$r(va + \text{constant}) = v(h - \delta a) + \max_{c\geq 0}(1 - v)c + \max_{y\in[0,a]} \alpha\left[L(y) - vy\right].$$

The right side is linear in a only if the constraint $y \leq a$ never binds. Then the first-order condition implies that the agent can always finance a constant early consumption $y(a) = L'^{-1}(v)$ after any history, which is impossible.

that the value function admits the following closed-form solution:

$$V(a) = -\frac{v_2}{2}a^2 + v_1 a + v_0, \text{ for } a \leq a^*, \tag{4}$$

where v_i, $i = 0, 1, 2$, are constant given by

$$\begin{aligned} v_2 &= \frac{\alpha A}{r + \alpha + 2\delta}, \\ v_1 &= \frac{\alpha A \bar{y} - v_2 h}{r + \alpha + \delta}, \\ v_0 &= \frac{v_1 h}{r} - \frac{\alpha A}{2r}\bar{y}^2, \\ a^* &= (v_1 - 1)/v_2. \end{aligned}$$

The optimal choices for consumption and early consumption are functions of storage given by[8]

$$c(a) = \begin{cases} 0, \text{ for } a < a^*, \\ h - \delta a^*, \text{ for } a = a^*. \end{cases} \tag{5}$$

$$y(a) = a, \text{ for } a \leq a^*. \tag{6}$$

The rate of storage, $\dot{a} = s_a(a)$, is given by

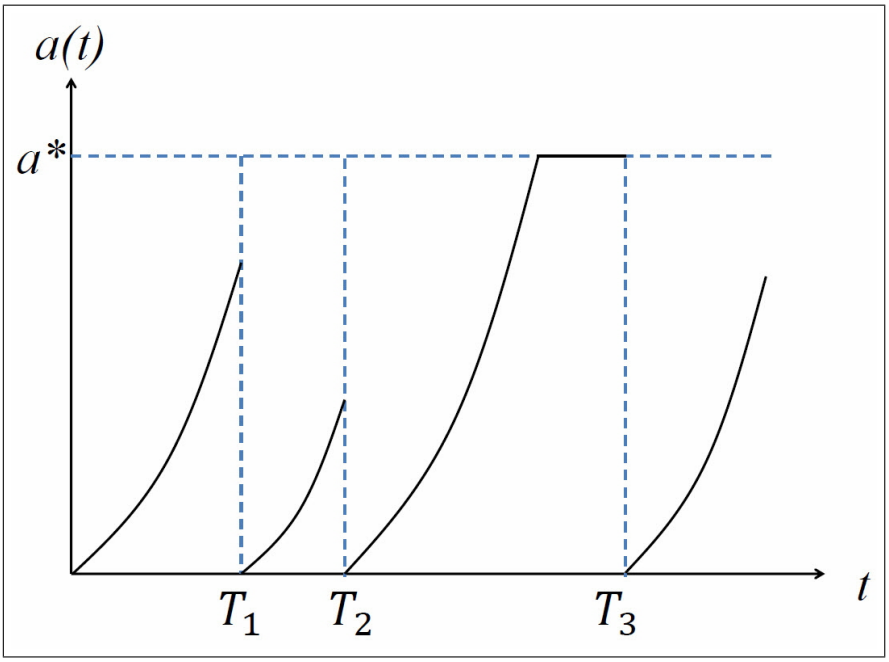
$$s_a(a) = h - c(a) - \delta a. \tag{7}$$

The typical pattern of storage is illustrated by Figure 1. Whenever a liquidity shock hits, the agent will draw all her storage available for early consumption and then "restart" the accumulation. The agent accumulates her storage toward the target level a^*. During this process, she stores all her endowment and consumes nothing. By doing so, she can build up the target level of storage the quickest. Once she reaches the target, she starts consumption at the rate that maintains her storage at the target, i.e., the endowment after deducting depreciation. The target level of storage balances current consumption and self-insurance for future liquidity shocks: maintaining a higher target gives better protection against the liquidity shocks, but it wastes more resources to depreciation and leads to lower consumption.

Remark. The model has a closed-form solution because it is designed to make use of the following properties. Suppose the value function, $V(a)$, is a polynomial of degree n and the consumption function, $c(a)$, is of degree n_c. Notice that the second term on the right side of HJB equation (2) involves a product of a term of degree $n-1$ and a sum of degrees of 0, 1, and n_c, so in general the resulting product is a term

[8] For these to hold, I have implicitly assumed the parameterization such that $a^* \in [0, \bar{y}]$, $c(a^*) \geq 0$, and $L'(a^*) \geq v_1$. It requires, for example, the marginal loss, A, to be sufficiently high such that an agent prefers to fully deplete all of her storage rather than leave some for the future. See Rocheteau et al. (forthcoming) for details.

Figure 1 A Time Path of Storage



of degrees of $\max(n, n-1+n_c)$. If the consumption utility is linear, then the consumption is bang-bang so $n_c = 0$ almost everywhere. If the equilibrium features full depletion, i.e., $y(a) = a$, then the last term of (2) is of degree $\max(n_L, n)$. Setting $n_L = 2$, a closed-form solution of V with $N = 2$ will match the degree on the both sides of (2).

Distribution. In this economy, agents have different histories of liquidity shocks, so they are different from one another in the level of storage. Are there a lot of agents poor in storage and hence severely exposed to the liquidity shocks? If so, then there will be potential social gain from risk sharing by having an alternative market structure. To answer this question, I need to know the distribution of storage in this economy. Denote $F_a(a) \in [0,1]$ the share of agents with weakly less than a units of storage, also known as the distribution function. In the continuous time, the distribution function simply solves the following Kolmorgorov forward equation (KFE):

$$s_a(a) F_a'(a) = \alpha [F_a[\Phi(a)] - F_a(a)], \text{ for all interior } a, \qquad (8)$$

where $F_a'(a)$ denotes the first derivative of $F(a)$—the density function. The mechanical meaning of the KFE is as follows. Consider the group

of agents with storage weakly less than a. Let's call this group of agents $\mathcal{A}$. The size of $\mathcal{A}$ is $F_a(a)$ by definition. I want to check the change in the size of $\mathcal{A}$ after a very short amount of time, $\Delta \cong 0$. For any member of $\mathcal{A}$ with storage strictly less than a, i.e., those with a' where $a' < a$, their storage will increase by $s_a(a')\Delta$ units if they do not receive any liquidity shock during the spell of Δ. When Δ is very small, the level of storage after Δ is given by $a' + s_a(a')\Delta \leq a$. If they do receive a liquidity shock, then according to the solution (6) they draw all their storage, $y(a') = a'$, for the early consumption, and hence their postshock storage is zero. In either case, the level of storage after Δ is still weakly less than a, so they remain in $\mathcal{A}$ and the size of $\mathcal{A}$ does not change. Now, consider the "border" agents to $\mathcal{A}$ with exactly $a' = a$. The size of the border agents is, roughly speaking, given by the density function, $F_a'(a)\partial a$. In the continuous time, the outflow of the border agents is $s_a(a)\Delta F_a'(a)$: the rate they are leaving the border group, $s_a(a)\Delta/\partial a$, multiplied by the size of the border group, $F_a'(a)\partial a$. On the other hand, there are also agents who did not belong to $\mathcal{A}$ until the liquidity shocks. These agents must have storage strictly greater than a before the liquidity shock but less than a after the shock. The size of this group of potential "immigrants" is $F_a[\Phi(a)] - F_a(a)$. In the continuous time, the inflow to $\mathcal{A}$ from the potential immigrants is $\alpha\Delta[F_a[\Phi(a)] - F_a(a)]$: by the law of large numbers, there is a share $\alpha\Delta$ of the potential immigrants hit by liquidity shocks. According to the KFE (8), the distribution of a is stationary when the outflow is equal to the inflow for any $\mathcal{A}$.

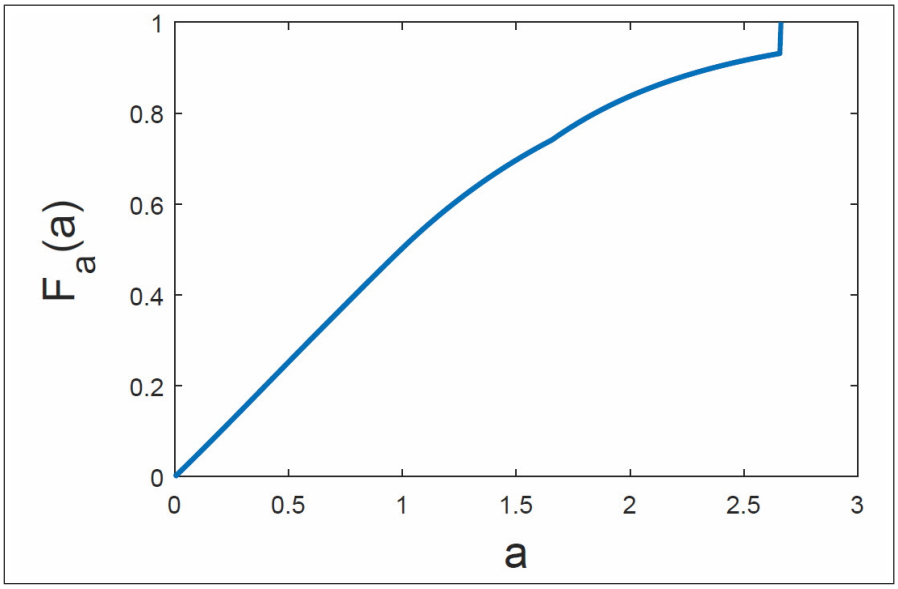
In general, the distribution function, $F_a(a)$, is the solution to the DDE (8) satisfying the boundary conditions $F_a(0) = 0$ and $F_a(a) = 1$ for all $a \geq a^*$: there is no agent with storage less than 0 or strictly greater than a^*. Since agents always draw all the storage for early consumption, i.e., $y(a) = a$, the real balances before a shock are given by $\Phi(a) = \infty$ for all $a > 0$. It is straightforward to verify that the distribution function follows a truncated beta distribution:

$$F_a(a) = \begin{cases} 1 - \left(1 - \frac{\delta}{h}a\right)^{\alpha/\delta}, \text{ if } a < a^*, \\ 1, \text{ if } a \geq a^*. \end{cases}$$

Figure 2 illustrates the distribution $F_a(a)$. The distribution is smooth everywhere except at $a = a^*$ with the point mass $1 - F_a(a^*)$. Once they reach the target a^*, agents stop accumulating further storage so there is a positive measure of agents with exactly a^* units of storage. The aggregate storage is given by

$$\mathbb{E}(a) = \int_0^{a^*} [1 - F_a(a)]\, da = \frac{h}{\delta + \alpha}\left[1 - \left(1 - \frac{\delta}{h}a^*\right)^{\alpha/\delta+1}\right].$$

Figure 2 The Distribution Function



The following proposition summarizes how the fundamentals change the aggregate storage.

Proposition 1 *The aggregate storage, $\mathbb{E}(a)$, is increasing in the magnitude of the liquidity shock, $\bar{y}$.*

Proof. It follows the fact that a^* is increasing in $\bar{y}$. ■

Proposition 1 states that liquidity shock raises the aggregate storage. This is because when agents expect bigger liquidity shocks, they want to prepare a higher level of storage on average for self-insurance.

Welfare formula. The (utilitarian) welfare of this economy is the total sum of the utility and loss across agents, which is given by

$$\mathcal{W}_a = \int_0^\infty [c(a) + \alpha L[y(a)]] \, dF_a(a).$$

The first term is given by

$$\int_0^\infty c(a) \, dF_a(a) = c(a^*)[1 - F_a(a^*)] = h\left(1 - \frac{\delta}{h}a^*\right)^{\alpha/\delta+1} = h - (\delta + \alpha)\mathbb{E}(a).$$

Since the economy features full depletion, the second term is given by

$$\int_0^\infty L[y(a)] \, dF_a(a) = -\frac{A}{2}\int_0^{a^*} (\bar{y} - a)^2 \, dF_a(a) = -\frac{A}{2}\left[[\bar{y} - \mathbb{E}(a)]^2 + VAR(a)\right].$$

Collecting these terms, the welfare can be written as the following mean-variance formula:

$$\mathcal{W}_a = h - (\delta+\alpha)\,\bar{y} + \frac{(\delta+\alpha)^2}{2\alpha A} - \frac{\alpha A}{2}\left[\bar{y} - \frac{\delta+\alpha}{\alpha A} - \mathbb{E}(a)\right]^2 - \frac{\alpha A}{2} VAR(a). \tag{9}$$

The following proposition summarizes how the welfare depends on the distribution.

Proposition 2 *The welfare of the self-insurance economy, $\mathcal{W}_a$, is*

(a) negatively related to the dispersion of individual storage, $VAR(a)$, and

(b) positively related to the aggregate storage, $\mathbb{E}(a)$, if and only if $\mathbb{E}(a) \leq \bar{y} - (\delta+\alpha)/(\alpha A)$.

Proposition 2 states that dispersion and aggregate level of storage are sufficient statistics for welfare. Inequality in storage reduces welfare because it means there are a lot of storage-poor agents exposed to the liquidity shocks. The effect of the aggregate storage on the welfare is not monotone. When the aggregate storage is small, such that $\mathbb{E}(a) \leq \bar{y} - (\delta+\alpha)/(\alpha A)$, an increase in the aggregate storage allows better protection against the liquidity shocks on average and raises the welfare. However, when the aggregate storage is large, such that $\mathbb{E}(a) \geq \bar{y} - (\delta+\alpha)/(\alpha A)$, an increase in the aggregate storage diverts too many resources from consumption on average, which reduces the welfare. This threshold is increasing in the marginal loss of the liquidity shock, A, because agents need more protection against the liquidity shock when its loss becomes more costly. For a similar reason, the threshold is increasing in the probability of the liquidity shock, α, but decreasing in the depreciation rate, δ.

2. EFFICIENCY LOSS TO SELF-INSURANCE

In this economy there is unit measure of uncountably many agents. One advantage to a mass society is that it has the critical mass to eliminate individual shocks by pooling resources. What should agents do collectively if they can coordinate for the best of themselves? The situation is the same as when there is one "representative" agent, the planner, solving the following problem

$$\max_{c^*, y^*} \{c^* + \alpha L(y^*)\} \text{ s.t.}$$

$$h \geq c^* + \alpha y^*.$$

The planner maximizes the total flow of the consumption utility, c, and minimizes the total flow of the loss to liquidity shocks, $\alpha L(y)$. The

resource constraint faced by the planner is that the total consumption and early consumption are not greater than the total endowments pooled by agents. The planner's solution, also known as the first best, is given by

$$\begin{aligned} y^* &= \bar{y} - A^{-1} \\ c^* &= h - \alpha\left(\bar{y} - A^{-1}\right). \end{aligned}$$

In the first best, agents pay a premium αy^* to insure the liquidity shocks by guaranteeing y^* units of early consumption. Compared with the first best, there are two sources of efficiency loss in the self-insurance economy. On one hand, when agents self-insure with storage, the average consumption is lower, $\mathbb{E}\left[c\left(a\right)\right] < c^*$, because some resources are wasted to depreciation. On the other hand, self-insurance by building up storage is a slow process due to depreciation and limited endowment. Compared with the first best, the self-insurance economy features less protection against the liquidity shocks on average, $\mathbb{E}\left[a\right] < y^*$, especially to the agents with less storage. It echoes Proposition 2 that the welfare is increasing in the inequality of storage. In the first best, every agent can perfectly share the liquidity risk and there is no inequality.

3. SHARING RISKS WITH MONEY

Now consider an intrinsically useless object called money. Maintained by a central bank, the stock of money, M_t, grows at the rate π, where $\pi \in [0, \delta]$. The central bank does not consume anything or withhold any resources, so the simplest way of injecting new money to the economy is the helicopter drop: the central bank creates and transfers a lump sum πM_t of money to every agent. It is also the same as the policy where the central bank purchases endowments from agents with newly printed money and then transfers all the purchased endowments to agents.

In other words, the central bank keeps printing for agents increasing amounts of paper, so-called money. How can it change the economy? Potentially, there is a market where agents can buy or sell endowment with money (they are not forced to do so). Denote as ϕ_t the real price of money in terms of goods, i.e., each unit of money can buy ϕ_t units of goods and its negative growth rate, $-\dot{\phi}_t/\phi_t$, is simply the inflation rate: the loss rate of the real purchasing power of money. The real price of money is determined by agents' net demand and the central bank's supply. The situation where money cannot be exchanged for anything is captured by $\phi_t = 0$. I first guess (and verify later) that agents will no longer store any endowment but will hold money instead. In the

continuous time, the budget constraint is given by

$$\dot{m}_t = \frac{h + \pi M_t - c_t}{\phi_t}.$$

That is, the change in money holding, $\dot{m}_t$, is equal to the monetary amount of income not consumed, $(h + \pi M_t - c_t)/\phi_t$. In the stationary equilibrium, if it exists, the total purchasing power of money should be constant such that $\phi_t M_t$ remains the same over time. It implies

$$-\frac{\dot{\phi}_t}{\phi_t} = \pi.$$

In other words, inflation is always a monetary phenomenon in the sense that the change in the real price of money is driven by the increase of the money supply. Denote the individual real balances as $z_t = \phi_t m_t$ and the aggregate real balances as $Z = \phi_t M_t$, then the budget constraint in the stationary equilibrium is given by

$$\dot{z}_t = h + \pi Z - c_t - \pi z_t. \tag{10}$$

That is, the change in the real balances of money is the income not consumed, $h + \pi Z - c_t$, minus the loss of real balances due to inflation, πz_t. Similar to the previous section, the value function of holding money solves the following HJB equation

$$rW(z) = \max_{c \geq 0, y \in [0,z]} \left\{ c + W'(z)(h + \pi Z - c - \pi z) + \alpha \left[L(y) + W(z - y) - W(z) \right] \right\}. \tag{11}$$

It is straightforward to verify that the value function admits the following closed-form solution:

$$W(z) = -\frac{w_2}{2} z^2 + w_1 z + w_0, \text{ for } z \leq z^* \tag{12}$$

where w_i, $i = 0, 1, 2$, are constant, given by

$$\begin{aligned} w_2 &= \frac{\alpha A}{r + \alpha + 2\pi}, \\ w_1 &= \frac{\alpha A \bar{y} - w_2 (h + \pi Z)}{r + \alpha + \pi}, \\ w_0 &= \frac{w_1 (h + \pi Z)}{r} - \frac{\alpha A}{2r} \bar{y}^2, \\ z^* &= (w_1 - 1)/w_2. \end{aligned}$$

The optimal choices for consumption and early consumption are functions of real balances given by

$$c(z) = \begin{cases} 0, \text{ for } z < z^*, \\ h + \pi Z - \pi z^*, \text{ for } z = z^*. \end{cases} \tag{13}$$

$$y(z) = z, \text{ for } z \leq z^*. \tag{14}$$

The rate of accumulating real balances, $\dot{z} = s_z(z)$, is given by

$$s_z(z) = h + \pi Z - c(z) - \pi z. \tag{15}$$

The distribution is given by the following KFE

$$s_z(z) F_z'(z) = \alpha [1 - F_z(z)], \text{ for all interior } z. \tag{16}$$

The closed-form solution to the distribution function of real balances is given by

$$F_z(z) = \left\{ \begin{array}{l} 1 - \left(1 - \frac{\pi}{h+\pi Z} z\right)^{\alpha/\pi}, \text{ for } z < z^*, \\ 1, \text{ for } z \geq z^*. \end{array} \right\}, \text{ if } \pi > 0,$$

or

$$F_z(z) = \left\{ \begin{array}{l} 1 - \exp\left(-\frac{\alpha}{h} z\right), \text{ for } z < z^*, \\ 1, \text{ for } z \geq z^*. \end{array} \right\}, \text{ if } \pi = 0,$$

where Z is the fixed point solving

$$Z = \mathbb{E}(z) = \frac{h + \pi Z}{\pi + \alpha} \left[1 - \left(1 - \frac{\delta}{h + \pi Z} z^*\right)^{\alpha/\pi + 1}\right]. \tag{17}$$

The following lemma shows that Z is well-defined.

Lemma 1 *There exists a unique solution Z to (17).*

Proof. The fixed point Z exists because the left side of (17) is smaller than the right side for $Z = 0$ but becomes larger than the right side for $Z = \infty$, so there must exist some Z where the left side is equal to the right side. The fixed point is also unique because the right side, as a function of Z, has a slope less than unity. ■

Equation (17) captures the market-clearing condition for money: the left side captures the money supply $Z = \phi_t M_t$ in real terms, and the right side captures the aggregate money demand by agents, $\mathbb{E}(z) \equiv \int z dF_z(z)$. Similar to Proposition 1, the following proposition establishes that the liquidity shock increases the money demand.

Proposition 3 *The aggregate real balances of money, $\mathbb{E}(z)$, is increasing in the magnitude of the liquidity shock, $\bar{y}$.*

Proof. The left side of (17) is a function of Z with a unit slope. The right side has a slope less than unity. An increase in $\bar{y}$ shifts up the right side via z^*, so the fixed point Z increases. ■

Like storage, agents in this economy hold money in order to self-insure against the liquidity shocks. Therefore, the more severe the magnitude of the liquidity shock, the more money agents hold. The value of money is measured by the equilibrium price, which is given by

$$\phi_t = \frac{Z}{M_t} = \frac{Z}{M_0} \exp(-\pi t).$$

Indeed, the fact that $\phi_t > 0$ verifies that money circulates in this economy. Finally, think of the inflation rate, $-\dot{\phi}_t/\phi_t$, as the depreciation rate of a new storage technology, which is preferred by agents to the original storage technology, since the depreciation rate with money is lower, i.e. $-\dot{\phi}_t/\phi_t = \pi \leq \delta$. It verifies the premise that the use of money will "crowd out" individual storage.

Having shown that money circulates despite the fact that money is intrinsically useless and there is always a coincidence of wants, now I want to check whether or not the economy is better off after the introduction of money. To do so, I need to compare the welfare to the self-insurance economy. The welfare in the monetary economy is given by

$$\mathcal{W}_z = \int_0^\infty \left[c\left(z\right) + \alpha L\left[y\left(z\right)\right]\right] dF_z\left(z\right).$$

The first term is given by

$$\int_0^\infty c\left(z\right) dF_z\left(z\right) = h - \alpha \mathbb{E}\left(z\right),$$

where I have made use of the fact that $Z = \mathbb{E}\left(z\right)$ in the equilibrium. Since the economy features full-depletion, the second term is given by

$$\int_0^\infty L\left[y\left(z\right)\right] dF_z\left(az\right) = -\frac{A}{2}\left[\left[\bar{y} - \mathbb{E}\left(z\right)\right]^2 + VAR\left(z\right)\right].$$

Collecting these terms, the welfare can be written as the similar mean-variance formula:

$$\mathcal{W}_z = c^* - \frac{\alpha}{2A} - \frac{\alpha A}{2}\left[y^* - \mathbb{E}\left(z\right)\right]^2 - \frac{\alpha A}{2} VAR\left(z\right). \tag{18}$$

Proposition 4 *The welfare of the monetary economy is*

(a) negatively related to the dispersion of real balances, $VAR\left(z\right)$*, and*

(b) positively related to the aggregate real balances, $\mathbb{E}\left(z\right)$*, if and only if* $\mathbb{E}\left(z\right) \leq y^*$*.*

Similar to the self-insurance economy, inequality in money holding reduces the welfare because it means there are a lot of money-poor agents exposed to the liquidity shocks. The aggregate real balances of money raise the welfare if and only if the aggregate real balances are less than the first-best level of early consumption, $\mathbb{E}\left(z\right) \leq y^*$, i.e., when agents do not hold more money on average than they should in the first best.

Now I am ready to conclude the social role of money with the following proposition.

Proposition 5 *Welfare is higher under the monetary economy than the self-insurance economy, i.e.,* $\mathcal{W}_z > \mathcal{W}_a$.

Proof. Notice that welfare is also equal to $\mathcal{W}_a = r \int V(a)\, dF_a(a)$ and $\mathcal{W}_z = r \int W(z)\, dF_z(z)$, i.e., the average value in the economy. It is straightforward, although tedious, to check the closed-form solutions that $V(x) < W(x)$ and $z^* < a^*$. Intuitively, the fact that money depreciates at a lower rate and agents receive an additional transfer πZ means that the agent with wealth x has higher value in the monetary economy, for any x. Finally, from the KFE the distribution functions are given by

$$\begin{aligned} \log[1 - F_a(a)] &= \int_0^a \frac{-\alpha}{s_a(x)} dx, \\ \log[1 - F_z(z)] &= \int_0^z \frac{-\alpha}{s_z(x)} dx. \end{aligned}$$

Given the fact that $s_z(x) > s_a(x)$, I have $1 - F_z(x) > 1 - F_a(x)$, i.e. there are more wealthy agents in the monetary economy. Combining the facts that $1 - F_z(x) > 1 - F_a(x)$ and $V(x) < W(x)$, I prove that $\mathcal{W}_a = r \int V(a)\, dF_a(a) < r \int W(z)\, dF_z(z) = \mathcal{W}_z$. ■

If the inflation rate is not too high that $\pi \leq \delta$, then the monetary economy is always more socially desirable than the self-insurance economy with private storage only. But what is the optimal inflation rate? In general, it is not zero, and finding the optimal rate is a quantitative exercise involving the following trade-off. A higher inflation hurts the return of money (inflation as a tax to money holding) and hence discourages its use as a precautionary saving. A higher inflation, however, enables generous monetary transfers and promotes social sharing of liquidity risks. The second effect dominates when h is low (self-insurance is low because income is low) or α is high (liquidity shock comes too frequent to maintain a sufficient level of self-insurance).

4. CONCLUSION

As rightly pointed out by Winston Churchill in my opening quotation, capitalism that builds on money and market inevitably results in unequal sharing of blessing and misery. A lesson from this review is, however, that money can well be our best response to mitigate unequal sharing of blessing and misery, so economy without money is even worse. How can money improve social welfare? My model illustrates three channels. Firstly, the usage of money mimics, though not perfectly, the first-best allocation. When agents hold and accumulate money, they sell their endowments to others who want to sell money. Let's call the former the seller (of endowment) and the latter the buyer

(of endowment). In the monetary economy, these buyers are exactly the ones who are hit by the liquidity shocks. When the buyers buy endowments and sell money to the sellers in the market, resources are efficiently transferred from nonshocked agents to the shocked agents—the monetary mechanism of risk sharing. Why do the sellers want to buy money? Because when they sell their endowments, the sellers actually buy an insurance against the future liquidity shocks, and the insurance premium is essentially the consumption forgone for acquiring money in the market. It is exactly what should happen in the first best, illustrated in Section 2. Secondly, money allows agents to save with other agents, instead of decentralized storage. And saving in money gives higher returns to agents, comparing the loss due to inflation with the loss due to depreciation. It avoids wasting resources to depreciation and allows more resources for consumption. Finally, agents who are poor in money are the unlucky ones frequently hit by the liquidity shocks. When the central bank keeps injecting new money into the economy, it helps poor agents build up their purchasing power. As a result, agents in the monetary economy are better prepared for the liquidity shocks.

REFERENCES

Aiyagari, S. Rao. 1994. "Uninsured Idiosyncratic Risk and Aggregate Saving." *Quarterly Journal of Economics* 109 (August): 659–84.

Bewley, Truman. 1980. "The Optimum Quantity of Money." In *Models of Monetary Economies*, edited by John H. Kareken and Neil Wallace. Minneapolis: Federal Reserve Bank of Minneapolis. 169–210.

Bewley, Truman. 1983. "A Difficulty with the Optimum Quantity of Money." *Econometrica* 51 (September): 1485–1504.

Davila, Julio, Jay H. Hong, Per Krusell, and José-Victor Ríos-Rull. 2012. "Constrained Efficiency in the Neoclassical Growth Model with Uninsurable Idiosyncratic Shocks." *Econometrica* 80 (November): 2431–67.

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

Hellwig, M. 1982. "Precautionary Money Holding and the Payment of Interest on Money." CORE Discussion Paper 8236 (September).

Lagos, Ricardo, Guillaume Rocheteau, and Randall Wright. 2017. "Liquidity: A New Monetarist Perspective." *Journal of Economic Literature* 55 (June): 371–440.

Lagos, Ricardo, and Randall Wright. 2005. "A Unified Framework for Monetary Theory and Policy Analysis." Journal of Political Economy 113 (June): 463–84.

Lippi, Francesco, Stefania Ragni, and Nicholas Trachter. 2015. "Optimal Monetary Policy with Heterogeneous Money Holdings." *Journal of Economic Theory* 159 (September): 339–68.

Rocheteau, Guillaume, and Ed Nosal. 2017. *Money, Payments, and Liquidity,* 2nd Ed. Cambridge, Mass.: MIT Press.

Rocheteau, Guillaume, Pierre-Olivier Weill, and Tza-Nga Wong. 2015. "Working through the Distribution: Money in the Short and Long Run." Working Paper 21779. Cambridge, Mass.: National Bureau of Economic Research. (December).

Rocheteau, Guillaume, Pierre-Olivier Weill, and Tza-Nga Wong. Forthcoming. "A Tractable Model of Monetary Exchange with Ex-Post Heterogeneity." *Theoretical Economics.*

Scheinkman, Jose A., and Laurence Weiss. 1986. "Borrowing Constraints and Aggregate Economic Activity." *Econometrica* 54 (January): 23–45.

Wallace, Neil. 2014. "Optimal Money Creation in 'Pure Currency' Economies: a Conjecture." *Quarterly Journal of Economics* 129 (February): 259–74.

Wong, Tza-Nga. 2016. "A Tractable Monetary Model under General Preferences." *Review of Economic Studies* 83 (January): 402–20.