The Quality Theory of Money

De-Xing Guan*

April 28, 2022

Abstract

The debate between the quantity theory of money and the real bills doctrine has been the focus of the monetary theory ever since the time of David Hume and Adam Smith. Smith's real bills doctrine is, in effect, a quality theory of money. Misled by Lloyd Mints, Milton Friedman considered real bills doctrine as fallacious. This is why most modern monetary economists have been adherents to the quantity theory of money. In this paper we establish a simple model to elaborate the quality theory of money. We compare our model with various monetary theories and with facts in the real world. In all cases we found that transaction costs are the key to understanding the quality of money. It is the quality, rather than the quantity, of money that matters.

Keywords: Quality of Money, Transaction Cost, Real Bills Doctrine, Quantity Theory
JEL classification: D2, E4, E5

* Department of Economics, National Taipei University.

I. Introduction

Money is a measuring rod. It measures the exchangeable value of almost all goods and services transacted in the market. Money is also a human device or an institution created by human beings to facilitate transactions. Without it the cost of transactions would be much higher and the operation of the market would be very costly. But the value of the measuring rod itself is not always stable. Its value naturally depends on the institutional structure of a society which created it. By institutional structure we mean the economic, political, and legal framework or infrastructure on which both money supply and money demand are based.

Monetary theories have been primarily concerned with the quantity of money since Milton Friedman's restatement of the quantity theory of money.[1] The quantity theory was proposed by David Hume in a series of essays.[2] Though interesting and influential, it was not agreed by one of the author's best friends, Adam Smith. Through some examples against the quantity theory, Smith proposed the so-called real bills doctrine, or a quality theory of money, in the *Wealth of Nations*.[3] From then on these theories became the two major monetary theories, and the debate between them has lasted for more than two centuries and has never truly ended.[4]

Heavily influenced by Lloyd Mints, Friedman considered real bills doctrine as "basically fallacious".[5] This is unfortunate because Mints confused real bills with the paper money issued by John Law in the famous Mississippi scheme.[6] Smith, in effect, viewed such unbacked paper money as fictitious, in the sense as he distinguished real bills from fictitious ones. With Mints in his mind, Friedman naturally inclined to take quantity theory as the right one to explain monetary phenomena in the real world.

Real bills doctrine is, of course, not the only theory about the quality of money. We can find many others in the works of Smith and of those who disagree with the quantity theory. For example, John Maynard Keynes opposed quantity theory since people would have propensity to hoard cash. This liquidity preference, as he coined the term, made the equation of exchange meaningful only if all transactions are

---

[1] Friedman (1969, Ch. 2).
[2] Hume (1752).
[3] Smith (1789, Bk. II, Ch. II). That the real bills doctrine is actually a quality theory of money was recognized by economists such as Friedman. For example, Friedman and Schwartz (1963, p. 266) used "that qualitative, "real bills" point of view" to describe the real bills doctrine.
[4] Sargent (2011) provided a very good review of the debate between these two theories.
[5] Friedman (1969, p. 76).
[6] Mints (1945, Ch. 3).

included, not just those leading to final goods and services.[7] In this sense, Keynes was not only concerned with the quantity of money, but the quality of it. And he used the term liquidity premium to represent this quality.[8]

Another example of the quality theory of money came from the problem of how to allocate the cost of bringing lender and borrower together. In financial markets there are at least three types of transaction costs: searching for information, bargaining and negotiation, and enforcing contracts.[9] During the 1929-1933 Great Depression and the 2008-2009 Great Recession these transaction costs were obviously huge.[10] Either lenders or borrowers could have two strategies to deal with these transaction costs: to incur or to transfer them. By transfer we mean passing on costs to others who do not voluntarily accept them. For example, in the Great Recession investment banks in Wall Street used financial instruments such as collateralized debt obligation (CDO) and credit default swap (CDS) to transfer costs to security buyers and taxpayers, especially when these banks were too big to fail.[11]

Japan's lost decades after the burst of the Heisei bubble in early 1990s gave us the final example concerning the quality of money. The first quantitative easing (QE) in the world was initiated in March 2001 by the Bank of Japan, which was then followed by another round of QE called quantitative and qualitative easing (QQE) in April 2013. That the quality of money was emphasized in QQE implied that the Bank of Japan had noticed that it is not only the quantity but the quality of money that matters.

In Section II we discuss the above four examples of the quality theory of money. Section III proposes a simple model of the quality theory of money as the workhorse to explain phenomena in these examples. Section IV applies our theory to the study of some cases in the real world, such as Japan's lost decades, subprime mortgage crisis and the Great Recession, QE, the optimum quantity of money, and the rise and popularity of mobile payments and digital currencies in transactions. Section V concludes.

[7] Keynes (1936, Ch. 15).
[8] Keynes (1936, Ch. 17).
[9] Coase (1988, Ch. 2).
[10] For a survey and comparison of these two episodes, please see Eichengreen (2015).
[11] See Lewis (2011) and Eichengreen (2015).

II. Examples Concerning the Quality of Money

The quality of money is, in effect, concerned with money's real value in exchange. And if we adopt a broad view of money, it would include both outside money and inside money, where the former is currency plus required reserves, and the latter consists of various deposits. Both currency and deposit have quantitative as well as qualitative parts. The quantity theory of money is mainly concerned with the quantitative part. Real bills doctrine has primarily focused on the qualitative part.

1. Real Bills Doctrine versus the Quantity Theory of Money

Both Hume and Smith were excellent monetary economists. Hume (1752) originated the idea of the quantity theory of money in two essays: "Of Money" and "Of Interest." Smith (1789, Bk. II, Ch. II) initiated the idea of real bills in the following passage:

> When a bank discounts to a merchant a real bill of exchange drawn by a real creditor upon a real debtor, and which, as soon as it becomes due, is really paid by that debtor; it only advances to him a part of the value which it would otherwise be obliged to keep by him unemployed and in ready money for answering occasional demands.

A real bill is contrary to a fictitious one in which the creditor, the debtor, or both of them would be fictitious. Fictitious bills would have harmful effects, as Smith had said in the same chapter:

> [W]hen the same two persons do not constantly draw and redraw upon one another, but occasionally run the round of a great circle of projectors, who find it for their interest to assist one another in this method of raising money, and to render it, upon that account, as difficult as possible to distinguish between a real and fictitious bill of exchange; between a bill drawn by a real creditor upon a real debtor, and a bill for which there was properly no real creditor but the bank which discounted it, nor any real debtor but the projector who made use of the money. When a banker had even made this discovery, he might sometimes make it too late, and might find that he had already discounted the bills of those projectors to so great an extent that, by refusing to discount any more, he would necessarily make them all bankrupts, and thus, by ruining them, might perhaps ruin himself.

Obviously Smith was concerned with the quality of bills of exchange (inside money).

In the days of Smith, there was no such thing as central banking.[12] But even there was a central bank which issued outside money, the focus of Smith would still be on the quality rather than the quantity of both outside and inside money.

The monetary theories of Smith and Hume were not contradictory to each other, but their emphases were different. Hume focused on money supply and implicitly assumed the long-run neutrality of money. But Smith had a balanced view between money demand and money supply. His focus was on the difference between real bills and fictitious ones, and economic fluctuations are mainly due to the over issuance of the latter. In other words, inside money was the focal point of Smith, but it is outside money to which Hume had paid more attention. The implication for an efficient monetary system is obviously different under these two theories. The problem is how to draw the line between outside money (gold or paper money in Hume's case) and inside money (bank money or bills of exchange in Smith's case).[13]

The term real bills doctrine was coined by Mints (1945) but, unfortunately, he thought it came from the idea of John Law, the originator of the infamous Mississippi bubble. Because Friedman had learned monetary theory from Mints, it was no wonder why he interpreted real bills doctrine in the wrong way. Barry Eichengreen was the other scholar misled by Mints. For example, he had said:

> That doctrine, developed in the early eighteenth century by, among others, the Scottish monetary theorist John Law, was intended as a guide to credit creation by the Bank of England…and play a role in the Mississippi Bubble…His real bills doctrine informed the conduct of central bank policy for two centuries and more thereafter.[14]

Law's proposal was, in effect, creating a fictitious bills doctrine. The root of the bubble was that stock dividends of the Mississippi Company were paid by issuing unbacked paper money. This means that there were not real debtors who would have "ready money for answering occasional demands." It is therefore unbelievable that this infamous doctrine had "informed the conduct of central bank policy" for more than two centuries. Misled by Mints, it is no wonder why quantity theorists have been hostile to Smith's real bills doctrine. It is Law's doctrine which should be blamed.

---

[12] Though the Bank of England was established in 1694, it was not on behalf of a central bank until the Peel's Act became legitimate in 1844.

[13] Tobin (1961) was one of a few modern economists who shared similar interests in real bills doctrine. Sargent (2011) had an excellent discussion of the debate between real bills and the quantity theory.

[14] Eichengreen (2015, p. 24). Actually the bank to which Law proposed his idea was the Bank of Scotland. There was not the Great Britain in 1705 when Law proposed. Scotland, Law's nation, and England were not united until 1707.

2. Liquidity Premium versus the Optimum Quantity of Money

As is well known, in a classical long-run equilibrium the aggregate supply curve is vertical, and therefore money in neutral. Keynes reminded us of the limitation of this result that "The Quantity Theory is often stated in this, or a similar, form. Now "in the long run" this is probably true…*In the long run* we are all dead."[15] Keynes did not stop there. He provided us with the *General theory* in which money is generally not neutral just because people have the propensity to hoard cash. The reason to keep cash is that it gives us liquidity premium, an extra return on what Keynes called liquidity: at short note on occasional demand.

When money is not neutral, at least in the short run, the supply of money would affect real output, real interest rates, and other real variables. Adapted from the setup of Keynes (1936, Ch. 17), Friedman (1969, Ch. 1) derived an optimum quantity of money where the nominal interest rate should be pushed to zero by the monetary authority such that the real interest rate would be equal to the deflation rate. But unless there is no liquidity premium, the nominal interest rate would not be zero. This was a result emphasized by Kenneth Rogoff, where he claimed that because people hold cash, especially that with large face value, central banks could not decrease nominal interest rates below the so-called zero lower bound (ZLB).[16] This in turn limits the ability for monetary policy, such as QE, to stimulate the economy. The optimum quantity of money in Rogoff's world is obviously not the one with zero nominal interest rate.

An equally important problem is about the stability of money supply and demand. As mentioned above, adherents of the quantity theory would usually assume that the money demand function is stable.[17] Though many empirical works had confirmed this, almost all of them ignored transaction costs.[18] It seems necessary to test the stability of money demand along the line of Friedman (1969) and Lucas and Nicolini (2015). But it should better include positive transaction costs. Similarly, to test the economic implication of the real bills doctrine it is necessary to have a stable money supply function. But unlike the case for the money demand, there are hardly any empirical evidences for the case of money supply.

---

[15] Keynes (1923, p. 80). Italics in this passage were original.
[16] Rogoff (2016).
[17] Many empirical studies about money demand can be found in Friedman (1969). A much more recent study on the stability of money demand was provided by Lucas and Nicolini (2015).
[18] Tobin (1961) was an exception, who had discussed the role of transaction costs. Wallace (1979) was another one, who had also dealt with transaction costs in the last section of his paper, but for unknown reasons this section was omitted in the final version for publication in the *American Economic Review*.

Another difference between Keynes and Friedman is that the optimum quantity of money was derived on the assumption that there are not transaction costs, but Keynes realized that there are costs of bringing lender and borrower together.[19] Friedman had denied this possibility, as he said in his article on the optimum quantity of money:[20]

> Treating the return to the lender and the cost to the borrower as equal assumes that both have the same anticipations and also that transactions costs of borrowing and lending can be neglected.

This neglect of transaction costs by Friedman made his optimum quantity of money impractical because the nominal rate of interest is usually larger than zero. And even if it is zero, it would probably not be an optimum state because the economy might have fallen into the liquidity trap. Maurice Allais had suggested a similar rule to Friedman's that the optimum real interest rate is zero.[21] In this paper we show that both Friedman and Allais would not be adequate when there are transaction costs of bringing lender and borrower together.

3. Great Depression versus Great Recession

When the Federal Reserve Act was enacted on December 23, 1913, real bills doctrine was written into the law. In mid-1920s Benjamin Strong, then governor of the Federal Reserve Bank of New York, decided to support Montagu Norman, then governor of the Bank of England, to restore gold standard in Britain. Through open market operation the interest rates in the United States were lowered to induce the outflow of gold to Britain. Low interest rates and huge money supply laid the foundation of America's stock and housing bubble in late 1920s. And finally, together with other factors, there was the Great Depression. In this case the quantity of money was increased but the quality of it was decreased. This is because in mid-1920s America's economy was good such that there was no reason to lower interest rates just to help other country to restore gold standard.

Adolph Miller, then at the Federal Reserve Board, had opposed Strong's policy. He was also described by Friedman and Schwartz (1963, p. 266) as an opponent of the real bills doctrine in the following passage:

---

[19] Keynes (1936, p. 208).
[20] Friedman (1969, p. 24).
[21] Friedman (1969, p. 22).

The Board's emphasis on direct pressure was related to a general view of monetary policy which distinguished sharply between discounts, bills (bankers' acceptances), and government securities as source of credit expansion. From that qualitative, "real bills" point of view, what mattered was the end use of the credit…Adolph Miller was one of the most consistent and tenacious supporters of the policy of direct pressure, yet apparently he did not accept the real bills view.[22]

The different views between Strong and Miller reflected that the monetary policy of the Federal Reserve Bank had shifted gradually from the original real bills view to the more popular one of the quantity theory of money. The task of central banks becomes to pay more attention to the management of the quantity than the quality of money. A good example of this change of views came from the policy of quantitative easing in the Great Recession. As central banks around the world purchased huge amount of government bonds and mortgage-based securities to provide liquidity service to the market, the quantity of money was the only issue to be concerned with. The quality of money, or equivalently whether those bonds and securities are real or legitimate, had never been an important issue in the policy-making process.

4. Japan's Lost Decades

Since the early 1990s Japan has experienced decades of recession. A very short recovery was found in 2006-2007, but soon the economy fell into recession. In the fifteen years 1990-2005, many firms in Japan had suffered from heavy debts and had tried their best to pay them back to the bank when they had earned any profits. To recover their balance sheets was the first priority for most firms in Japan during these fifteen years. This made the recovery much slower than anticipated. This phenomenon has been called the balance sheet recession by Richard Koo. What Koo (2009) has emphasized was that the first priority of those firms with debts was not to maximize profits but to minimize debts. This made many traditional macroeconomic theories inadequate in dealing with the problem of Japan's lost decades.

[22] Ironically, Miller was considered as a supporter of the real bills doctrine by Eichengreen (2015).

III. A Simple Model of the Quality Theory of Money

1. Risk, Cost, and the loanable funds market

Smith had once said that "the exchangeable value of every commodity is more frequently estimated by the quantity of money, than by the quantity either of labour or of any other commodity which can be had in exchange for it."[23] This indicates that money is usually the nominal measure of the exchangeable value (or price) of every commodity. But the function of money is not restricted to the medium of exchange. It can also act as the store of value. The use of money (or interest) played an important role in Smith's theory of money, which can be shown in the following passage:[24]

> Whoever derives his revenue from a fund which is his own, must draw it either from his labour, from his stock, or from his land. The revenue derived from labour is called wages. That derived from stock, by the person who manages or employs it, is called profit. That derived from it by the person who does not employ it himself, but lends it to another, is called the interest or the use of money. It is the compensation which the borrower pays to the lender, for the profit which he has an opportunity of making by the use of the money. Part of that profit naturally belongs to the borrower, who runs the risk and takes the trouble of employing it; and part to the lender, who affords him the opportunity of making this profit.

When money is not ready for immediate consumption but for the use of other people, the lender of the money would charge the borrower an interest for the use of money. The lender is therefore the supplier, and the borrower the demander, of these loanable funds. Interest is indeed part of the profit, and the profit should be shared by the lender and the borrower. What was wrong with the classical loanable funds theory of interest is at least twofold.[25] First, as argued by Keynes, loanable funds market cannot simultaneously determine the interest rate and aggregate income unless it is at the full employment equilibrium. Second, lenders and borrowers meet directly to determine equilibrium price without incurring any costs. In other words, there are not costs of bringing lender and borrower together. If these costs are important, then the problem would be that what are these costs, and how to measure them? Keynes had given us some suggestions in dealing with this problem. He thought that risk was

---

[23] Smith (1789, Bk. I, Ch. V).
[24] Smith (1789, Bk. I, Ch. VI).
[25] The term classical theory of interest was coined by Keynes (1936, Ch. 14). What he meant by it was the theory of Alfred Marshall, David Ricardo, and the Austrian School (mainly Ludwig Mises and Friedrich Hayek), not that of Smith. Loanable funds theory was also associated with Dennis Robertson.

ignored by loanable funds theory. As he said in the *General Theory*:[26]

There is, finally, the difficulty…of bringing the effective rate of interest below a certain figure, which may prove important in an era of low interest rates; namely the intermediate costs of bringing the borrower and the ultimate lender together, and the allowance for risk, especially for moral risk, which the lender requires over and above the pure rate of interest. As the pure rate of interest declines it does not follow that the allowances for expense and risk decline *pari passu*. (italics original)

Contrary to most economists, who usually used either utility function (risk aversion) or random variable (variance) to represent risks,[27] Keynes considered risk as a kind of cost. In the preface to an earlier book on money, he said:[28]

It is often supposed that the costs of production are threefold, corresponding to the rewards of labour, enterprise, and accumulation. But there is a fourth cost, namely risk; and the reward of risk-bearing is one of the heaviest, and perhaps the most avoidable, burden on production. This element of risk is greatly aggravated by the instability of the standard of value.

Keynes's idea of risk as cost is heuristic. The traditional representation of risk as utility or variance has the problem that utility is not measurable and variance is often exogenous. Cost has no such problems because it is often measurable and can be endogenously determined. It provides us with an alternative interpretation of risk.

When we consider risk as cost and apply the idea of Keynes to the study of a modified loanable funds theory, we would like to adopt the idea of Ronald Coase. This great idea is the transaction cost discussed thoroughly in Coase (1988). In our model the loanable funds, or simply loans, are an intermediate input to the production of final outputs, or investment goods. Loans themselves are produced by labor and capital. Similar to Goodfriend and McCallum (2007), the labor used to produce loans aims to monitor the quality of these loans: to make sure that there are real creditors and real debtors, as required by the real bills doctrine. The capital is used as the collateral of making loans. But there are differences between Goodfriend-McCallum model and ours. In their model all markets are competitive. In our model the labor and capital markets for producing final goods are competitive but the loanable funds market is not. When the cost of bringing lender and borrower together is larger than

[26] Keynes (1936, p. 208).
[27] As for the standard representation of risks in economics, see Samphantharak and Townsend (2018) and Schulhofer-Wohl (2011), among others.
[28] Keynes (1923, pp. ix-x).

the profit of making the loan, either the bank would not lend the money, or the firm which borrows from the bank would not pay it back. Only if the profit is large enough to cover the cost of bringing lender and borrower together, there would be loans in the equilibrium. In our model the cost of bringing lender and borrower together, or transaction cost in the loanable funds market, is larger than the monitoring effort and collateral cost in the model of Goodfriend and McCallum (2007). In their model the profit is no less than the monitoring and collateral cost such that there is always a competitive equilibrium.

2. A Theoretical Model of the Quality of Money

2.1. An Overview of the Model

Assume there are four groups of agents in the economy: households, banks, firms, and the central bank. Central bank issues fiat money for households to buy goods and services produced by the firm. For simplicity, suppose that firms earn zero profits because the market for final goods and services are perfectly competitive. The bank gets fiat money by taking deposits from households, and pays them interest at the real rate of r'. In a fractional reserve system the bank could lend part of the deposits to the firm as a loan and earn an interest at the real rate of r.

In reality the operation of both the bank and the central bank would have costs. The fiat money is actually a social contract between the central bank and the people who use it. For example, the Federal Reserve Bank is the legal tender of the Federal notes or U.S. dollars. In the era of silver standard or gold standard the bank, which issued what Smith (1789) called the bank money, had the responsibility to pay equal value of silver or gold to the depositors on their demand. There are costs in enforcing this social contract. In the commodity standard era the cost might be in digging, producing, and transporting the species from the mines to the market. But even in an era of fiat money there are still costs in issuing the irredeemable paper money. After all, the main difficulty of maintaining a fiat money system is how to restraint the inappropriate over-supply of the paper money by the central bank, or how to make the long-run price level more predictable. Though transaction costs had been ignored by Friedman, he finally admit that they are important for the conduct of the monetary policy. As he said in the following passage:[29]

[29] Friedman (1986, p. 643).

I took it for granted that the real resource cost of producing irredeemable paper money was negligible, consisting only of the cost of paper and printing…such an assumption, while it may be correct with respect to the direct cost to the government of issuing fiat outside money, is false for society as a whole and is likely to remain so unless and until a monetary structure emerges under an irredeemable paper standard that provides a high degree of long-run price level predictability.

This means that the resource cost, or prime cost, of producing fiat money is near zero, but what is important is the cost of maintaining the credibility of this paper money and of keeping the price to be predictable and the inflation or deflation to be under control. These costs might be considered as the transaction cost of the money supply, which was unfortunately ignored by most monetary economists, including Friedman himself in his 1969 monograph on the optimum quantity of money.[30]

On the contrary, transaction costs were not ignored by authors such as Baumol (1952), Tobin (1956), Goodfriend and McCallum (2007), Lucas (2013), and Lucas and Nicolini (2015). They had adopted similar settings in which there are costs in replenishing households' cash balances (Baumol and Tobin), managing cash (Lucas and Nicolini), or monitoring loans (Goodfriend and McCallum). We follow their settings in assuming that there are transaction costs when using cash as a medium of exchange without specifying the particular structure of the demand for money, so one can interpret our model in the sense of either one of these authors. Nevertheless, our model is most related to Goodfriend and McCallum (2007) because there is also a loan production function as in theirs.

2.2. Banks, Firms, and the Loanable Funds Market

As emphasized by Modigliani and Miller (1958) and Koo (2009), assume that the representative firm wishes to minimize its cost of (per capita) capital. The factor of production consists of labor, (physical) capital, and loans (or financial capital). The final good is produced by labor and capital with, or without, loans.[31] Loans are considered as an intermediate input, with which firms can produce more efficiently because financial markets are supposedly to provide firms with more and better

[30] Keynes (1923) might be the first economist who noticed the importance of the transaction cost in issuing the paper money. This is probably one of the reasons he would have in mind when he was proposing to the British monetary authority (mainly Bank of England) with *A Tract on Monetary Reform* about the arrangement of the international monetary system after the 1922 Genoa Conference.
[31] Firms without having loans can get funds through homemade leverage.

loanable funds. High quality loanable funds require more monitoring efforts and better collateral physical capital. If so, they are what Smith called real bills. If not, they are fictitious ones.

The cost minimization problem of firms can be described as a two-stage problem. At the first stage banks use consumers' deposits and their efforts of monitoring and collateralized capital to make loans. Then at the second stage firms borrow these loanable funds from banks to finance their investments and to produce final goods.

To introduce Coasian transaction cost into our model let us assume that some efforts *X* are necessary in using loanable funds markets to produce goods, such as monitoring labor and collateralized capital, as required by Goodfriend and McCallum (2007). In general, these efforts include other costs of searching for information, bargaining and negotiating, and enforcing the contracts. Without loss of generality, assume that the efforts of using loanable funds markets are linearly related to loans actually made by professional loan makers, or assume that $A_E = \mu X$, where A_E are loanable funds which embody the idea or expertise a typical professional loan maker would have in making loans, and $\mu > 0$ is a variable representing the efficiency of using efforts to make loans. A larger value of μ implies that professional loan makers have better expertise such as more information, better knowledge and know-how, better skills in making loans, and so on.

Let the price of efforts be P_X, that is, the cost of a unit of efforts in terms of the final good. Note that $1/\mu$ is the cost of producing a unit of expertise in terms of efforts, so P_X / μ is the cost of producing a unit of market-made goods in terms of the final good, which we define as marginal transaction cost (C^T) of producing market-made goods, or $C^T = P_X / \mu$. The efforts of using markets are factors of production and therefore intermediate goods of producing final goods. They are produced by other factors of production such as labor and capital. Assume that this production function is Cobb-Douglas: $X = K_E^{\beta} L_E^{1-\beta}$, such that we have $A_E = \mu X = \mu K_E^{\beta} L_E^{1-\beta}$, where L_E, K_E are labor and capital devoted to the accumulation of expertise, respectively, $0 < \beta < 1$.

According to Coase (1988), transaction costs are the costs involved in using institutions such as markets, firms, and the law. When there are no transaction costs,

the equilibrium condition would require that the price of loans be equal to the discounted sum of profits or net cash flow the loan will generate.[32] But when there are transaction costs the equilibrium arbitrage condition would require that

$$(1)\quad C^T + P_E = \frac{\pi_1}{1+r} + \frac{\pi_2}{(1+r)^2} + ... + \frac{\pi_n}{(1+r)^n}$$

where P_E is the price of A_E, π_i is the flow of profits generated by the banker's expertise of making loans in the *i*th period, and *n* is the duration of loans, $i = 1,2,...,n$. Equation (1) indicates that the sum of the discounted profits or net cash flow of acquiring new loans is equal to the full cost of doing so. And the full cost includes not only the cost of acquiring the loan itself, but also the transaction cost of protecting and enforcing the property rights of it.

After acquiring the expertise people have to provide some efforts for protecting and enforcing their property rights. The price of doing this is P_X, as discussed above, and the full cost would be $P_E A_E + P_X X = FA_E$, where F is the unit full cost of the expertise. When full cost is greater than net cash flow, banks would have less incentives to learn new skill in making loans; otherwise they would like to learn more. In equilibrium the full cost must be equal to the net cash flow of making loans. Note that $P_X = C^T \mu$ and $A_E = \mu X$, so $P_X X = C^T A_E$. This implies $FA_E - C^T A_E = P_E A_E$, or simply $F = C^T + P_E$. In equilibrium the full cost *F* is obviously the full price of the loan.

Now we consider the representative firm's problem. Assume that firms face a Smithian make-or-buy decision:[33] to make the loan by themselves or to buy (borrow) it in the loanable funds market.[34] The purpose of firms is assumed to get loanable funds they want in the least costly way. According to the principle of comparative advantage, sellers in the market are usually better at producing goods than buyers. Because using markets is costly, buyers should pay transaction costs such that sellers are willing to bring goods to the market. The cost minimization problem of firms can

[32] Paul Romer (1990, p. S87) had called this condition the intertemporal zero profit constraint, where the economic profit is zero at the equilibrium. But he did not consider transaction costs. The accounting profit is equal to the fixed cost at equilibrium and, therefore, would be greater than zero.
[33] Smith (1789, Bk. IV, Ch. II).
[34] Making loans on firm's own credit, not on that of banks, was usually called the homemade leverage. See, for example, Modigliani and Miller (1958, p. 269).

be described as follows:

(2) $C = Y \min\{\min(F, \gamma w^{1-\alpha} r^{\alpha}), (1-\gamma) w^{1-\alpha} r^{\alpha}\}$

where C is total cost of producing the final good Y, $w^{1-\alpha} r^{\alpha}$ is the unit cost of labor and capital, where $0 < \alpha < 1$, γ is the fraction of labor and capital devoted to the production of the final good with loanable funds borrowed from the bank, and $1-\gamma$ is the other fraction devoted to the production of final goods with funds financed by firms themselves (homemade leverage), $0 < \gamma < 1/2$.[35] Equation (2) indicates that all three factors of production: expertise, labor, and capital are necessary to produce final goods with loanable funds borrowed from banks, but only labor and capital are required for firms with homemade leverage.[36] Firms can either make loans by themselves, or borrow them from banks in the loanable funds market. They just choose the least costly way to finance their investments.

In equilibrium the total cost of borrowing funds from banks and homemade leverage would be the same.[37] Because there are three inputs: professional loan maker's expertise, labor, and capital, the total cost function can be written as $C = FA_E + wL_N + rK_N$, where w is wage rate, r is real loan rate, and L_N, K_N are labor and capital in producing the final output.[38] We assume that both labor and capital markets are competitive but the market for loans is not. The first minimization problem inside the curly bracket of equation (2) requires that loan, labor, and capital are all necessary in producing final goods with loanable funds borrowed from banks, that is,

(3) $C = FY = \gamma w^{1-\alpha} r^{\alpha} Y$

---

[35] The expertise of professional bankers could be their knowledge concerning loans, their skills in making loans, or any other know-how which ordinary people could not easily obtain. Because banks need both experts and ordinary workers, the fraction of labor/capital making loans by firms themselves should not be less than one half. Otherwise no firms will go to banks because loans are too expensive to make there.

[36] Of course, homemade leverage also needs monitoring efforts and collateralized capital. We assume that there are not these costs just for simplicity.

[37] The total cost should include both prime cost and transaction cost.

[38] As will be shown later, the expertise is in turn produced by both labor and capital, and the aggregate production function will be a weighted average of the outputs produced by firms with loanable funds or through homemade leverage, with the weights being the fractions of labor and capital allocated to these two kinds of production.

This means that $FY = C = FA_E + wL_N + rK_N$, or $Y = A_E + (wL_N + rK_N)/F$. Since the unit cost function is assumed to be Cobb-Douglas, an immediate implication of this result is that $(1-\alpha)(wL_N + rK_N) = wL_N$. Combining this with the above equations we have

(4) $Y = A_E + (wL_N)/[\gamma(1-\alpha)(w^{1-\alpha}r^{\alpha})]$

By Shephard's Lemma, $L_N = \partial C/\partial w = \gamma(1-\alpha)w^{-\alpha}r^{\alpha}Y$,

$K_N = \partial C/\partial r = \gamma\alpha w^{1-\alpha}r^{\alpha-1}Y$, so $K_N/L_N = \alpha w/[(1-\alpha)r]$. Inserting this into (3) and rearranging terms would have

(5) $Y = A_E + A_N' K_N{}^{\alpha} L_N{}^{1-\alpha}$

where $A_N' = [(1-\alpha)/\alpha]^{\alpha}/[\gamma(1-\alpha)]$.

The solution for the second cost minimization problem outside the curly bracket of equation (2) requires that the total cost of producing the final good would be the same in equilibrium, so we have

(6) $(F + \gamma w^{1-\alpha}r^{\alpha})Y = (1-\gamma)w^{1-\alpha}r^{\alpha}Y$

The solution to equation (6) is equivalent to that of the following redefined problem:

(7) $C = Y\min\{F, (1-2\gamma)w^{1-\alpha}r^{\alpha}\}$

A similar aggregate production function to equation (5) could be derived with only a modification of replacing A_N' by A_N, where $A_N = [(1-\alpha)/\alpha]^{\alpha}/[(1-2\gamma)(1-\alpha)]$.

In this paper we assume that $\beta > \alpha$. This means that the marginal productivity of per capita capital at the extensive margin (acquiring expertise) is greater than that at

the intensive one (no-expertise efforts), or that the production function with loans made by banks has larger marginal product than that without them. Otherwise there are no banks which would make loans if the quality and convenience of these financial services are the same. Combining $A_E = \mu X$ with equation (5) gives rise to the following aggregate production function:

(8) $Y = \mu K_E^{\beta} L_E^{1-\beta} + A_N K_N^{\alpha} L_N^{1-\alpha}$

Before the completion of the model, we first explore the relation between the aggregate production function and the market equilibrium of final goods. First, when there are no transaction costs ($C^T = 0$), $P_E = F$, and this is the standard arbitrage equilibrium condition: at the margin, the cost of buying the good is equal to the discounted sum of profits (or monopoly rent) generated by selling this good in the market. But when $C^T \to 0$, $\mu = P_X / C^T \to \infty$, so $X = A_E / \mu \to 0$: no efforts will be devoted to using the market. This contradicts the fact that using markets is costly in the real world. The second aspect is that when firms would like to borrow from banks in the loanable funds market it must pay the costs involved in using the market. If it does pay the full price, that is, prime costs plus transaction costs, then its demand for the good becomes Smith's effectual demand; otherwise, it is an absolute demand.[39] Obviously here the effectual demand is represented by the full price $P_E + C^T$ such that without paying for transaction costs, the firm's demand would become absolute and it will not be realized in the market. The firm must pay not only the prime cost but the transaction cost to bring the good to the market. The firm would borrow nothing if it only pays for the fixed cost. Another implication of equation (6) is that, for any goods to be effectively brought to the market, marginal benefits (rent) must exceed marginal costs (transaction cost) of doing so, or $F > C^T$. If the benefit fails to be larger than the cost, no new loans would be made. In the extreme case that $C^T \to \infty$, it is too costly for the firm to start a new business, such that there are no new loans to be made at all. Mathematically, $A_E = P_X X / C^T \to 0$ as $C^T \to \infty$.

To close this model we need market-clearing conditions for both labor and capital. Assume that there is a θ fraction of people who would like to learn the expertise of

[39] For the distinction between effectual demand and absolute demand, see Smith (1789, Bk. I, Ch. VII).

making loans, where $0 < \theta < 1$, and the remaining $1-\theta$ has two choices: γ fraction of it would choose to work at the extensive margin (in the market), while $1-\gamma$ of it would work at the intensive margin (at home). For simplicity, we also assume that the proportions of capital employed at these two margins are the same as those of labor. Again nothing important would be changed if this assumption were relaxed. The labor and capital markets clear if $L_E + L_N = L$ and $K_E + K_N = K$, where L, K are the aggregate supply of labor and capital, respectively. When all markets clear, equation (5) would become

(9) $Y = \theta\mu K^{\beta} L^{1-\beta} + (1-\theta) A_N K^{\alpha} L^{1-\alpha}$

$\theta\mu K^{\beta} L^{1-\beta}$ is the fraction of skilled labor/capital devoted to the accumulation of the expertise. $(1-\theta) A_N K^{\alpha} L^{1-\alpha}$ can be decomposed into two parts: $\gamma(1-\theta) A_N K^{\alpha} L^{1-\alpha}$ and $(1-\gamma)(1-\theta) A_N K^{\alpha} L^{1-\alpha}$. The first part is the fraction of unskilled labor/capital devoted to producing final goods with loanable funds, and the second part is that devoted to producing goods through homemade leverage. Equation (9) characterizes the aggregate production possibility frontier. It is a weighted average of the production functions at both extensive and intensive margins.

All of these results can be illustrated by Figure 3. In a world without transaction costs, no market-made goods would be produced because using the market is not costless. This implies that $A_E = 0$, and the point *B* in Figure 3 will shrink to the origin immediately. In a world with positive transaction costs there are two situations. First, if transaction costs are no less than the rent the firm might earn from its production of the new good, that is, if $F \le C^T$, then obviously no goods will be produced. The point *B* in Figure 3 will again shrink to the origin. Second, if $F > C^T$, then the new good will be produced, and in equilibrium, $F - C^T = P_E > 0$, a positive price which is necessary for A_E to exist.

Transaction costs, therefore, act as thresholds to the introduction of new ideas or new goods into the economy. When transaction costs are lower because of better legal system, more information, less unnecessary lawsuits, less political conflicts, among

others, point *B* in Figure 3 will move rightward to point *B*', and the intersection point of the two production functions (point *A*) will move upward along the production curve at the extensive margin to another *newer* extensive margin (to point *A*' in equilibrium). This is because now the firm would have better expertise due to the reduction of transaction costs. This process will go on and on if more transaction costs are reduced and therefore better institutions are established. The long-run aggregate production possibility frontier will be the upper envelope of the production functions at various margins. There is always another better extensive margin out there for people to pursue if they can find a better way to get to it.

Before describing the behavior of consumers and the central bank the reader should note that there is a similar arbitrage condition such as equation (1) for firms with homemade leverage. Assume that there are no transaction costs when using homemade leverage. The intertemporal zero profit constraint for such firms would be

$$(10)\quad P_E = \frac{\pi_1}{1+MEC} + \frac{\pi_2}{(1+MEC)^2} + \ldots + \frac{\pi_n}{(1+MEC)^n}$$

where MEC stands for the marginal efficiency of capital.[40] If profits in each period were unchanged for firms with and without using loanable funds, then MEC must be greater than *r*. When there are no transaction costs, MEC is equal to *r*. $MEC - r$ is measuring the excess rate of return of the entrepreneur over the banker. Otherwise, there would be no business firms because bankers could have done better jobs in investment than entrepreneurs do.

Equation (10) can also be interpreted as follows. If firms can have loans provided by banks without incurring any transaction costs, just as if they could have funds with the same quality but through homemade leverage, then the internal rate of return (that is, MEC) would be higher than that of the firm which have to incur transaction costs. If both types of firms co-exist in the world, then on average MEC must be greater than the real interest rate for the loanable funds market (that is, *r*).

2.3. Banks, Consumers, and the Central Bank

The ultimate sources of loanable funds come from consumers. They deposit their

[40] MEC was adapted from Irving Fisher's "rate of return over cost" by Keynes (1936, p. 140), which is the internal rate of return of the replacement cost, or of the supply price of capital owned by the firm itself.

savings in banks in return of a real deposit rate. These savings become funds to be borrowed by business firms. Consumers can use, save, or hoard the money they have. How much of the money they would hold as cash was discussed by Baumol (1952) and Tobin (1956), to cite the two most famous works on this topic. Much later, Lucas (2013) and Lucas and Nicolini (2015) extended the Baumol-Tobin framework to the case where there are cash management costs. The role played by the central bank in determining how much cash and credit a society should hold was the main theme of Woodford (2010), Sargent (2011), and Stiglitz (2016). The model presented in the current paper might shed some light on this important issue.

Assume that there are no taxes and public debts in our economy. All government spending is financed by issuing fiat money. This outside money plus the inside money created by the banking system as a whole consist of the total amount of money circulated in the economy. Call this sum total M1, which is equal to currency plus deposits, or $M1 = CC + DD$, where CC is the currency held by the public and DD is the deposit. The outside (or high-powered) money created by the central bank is denoted by M0, which is the sum of currency and the required reserves of the bank, or $M0 = CC + rrDD$, where rr is the required reserve ratio such that $rrDD$ is the required reserve. $P_E A_E = (1 - rr)DD$ is therefore the nominal value of loanable funds. Let $cr = CC / DD$ be the currency ratio. In a fractional reserve system the money multiplier is $m_G = \Delta M1 / \Delta M0 = (cr + 1)/(cr + rr)$.[41]

The reason that consumers would deposit their savings in the bank must be that the money they had could not give them more returns in alternative uses. To put it another way, this means that there are high enough transaction costs when consumers use their money in other investment opportunities. Again, there is another zero profit condition for consumers:

$$(11)\quad C^{T'} + C^{T} + P_E = \frac{\pi_1}{1+r'} + \frac{\pi_2}{(1+r')^2} + \ldots + \frac{\pi_n}{(1+r')^n}$$

where $C^{T'}$ is transaction costs consumers have to incur. They are the cost of bringing lender (consumer) and borrower (bank) together in the deposit market, just as C^{T} are the cost of bringing lender (bank) and borrower (firm) together in the loan market. As in the above discussion, if profits in each period were unchanged for consumers with and without depositing money in the bank, then r must be greater than r'. When

[41] For definitions of these concepts see, for example, Friedman and Schwartz (1963, Appendix B).

there are no transaction costs, these two rates are equal to each other. Thus $r - r'$ measures the excess rate of return, or the spread, of the bank over the consumer.[42] Otherwise, there would be no banks because consumers could have done better jobs in lending money to entrepreneurs than bankers do.

Though the above model is simple enough, the novelty here is the transaction cost in the operation of the bank and the money market. When this cost is higher, the bank's profit becomes lower, and the loanable funds created by the banking system would decrease. This is the typical phenomenon of a recession. On the other hand, when transaction cost is lower, the credit or inside money created by the bank would increase, and this is probably a situation of prosperity. But an over-heated economy needs more fuel from outside the banking system, and a great contraction is usually the result of a liquidity crisis. Thus, we have to consider the behavior of the government or the central bank because they are responsible for the creation of public debts and the outside money, which are probably roots of financial booms and busts.[43]

To sum up, when transaction costs are greater than zero as in the real world, there are costs of bringing lender and borrower together, as claimed by Keynes (1936). And in arbitrage equilibrium we would have the following inequality: $MEC > r > r' > 0$. The equilibrium real interest rates should be greater than zero. Otherwise there would be no loanable funds market in the first place. The real zero lower bound (RZLB) for the loanable funds market in particular, and for the financial markets in general, will be further discussed in Section IV below.[44]

Now we turn to the problem of the central bank. Assume that there are costs in implementing the monetary policy for the central bank. If the central bank's goal is to pursue full employment and price stability, then it needs to stay focused with the targeted interest rate and unemployment rate. The announcement of the Fed in 2013 to target the expected inflation rate at 2% and unemployment rate at 6.5% was an example. The efforts the Fed pays for implementing these monetary policies could be considered as costs of enforcing the mandates authorized by the Federal Reserve Act

[42] A positive spread means that there are at least two interest rates in equilibrium, that is, deposit and loan rates. Unfortunately, mainstream macroeconomic models usually have only one equilibrium interest rate. This embarrassing situation has been changing. Woodford (2010) was one such example.
[43] The theory of monetary business cycles can be found in the works of Hayek, Friedman, and even Keynes, but their emphases were not the same. The debate between Keynes and Hayek was the most profound one. A recent example about this debate was Rognlie, Shleifer, and Simsek (2018).
[44] Zero lower bound (ZLB) for the nominal interest rate has been a popular idea since the Bank of Japan initiated the world's first QE in 2001. The ZLB came from the idea of liquidity trap suggested by Keynes (1936, p. 207). During and after the Great Recession, ZLB has become one of the main debates in both macroeconomic theory and practice. See, for examples, Koo (2009), Woodford (2010), Rogoff (2016), and Rognlie, Shleifer, and Simsek (2018). In effect, what really matters is not ZLB, but RZLB.

of 1913. These costs of enforcing legal contracts are what Coase (1988) called transaction costs which was finally accepted as the resource cost of irredeemable paper money by Friedman (1986).

Another cost of monetary policy stems from the plausible time inconsistent behavior of the central bank. For example, in December 2015 Fed raised the federal funds rate target for the first time since the Great Recession. The rate was increased again in December 2016, and three times in 2017. In early 2017 Fed officials, such as the then vice chairman Stanley Fischer, claimed that Fed should raise rates at least four times that year. But it turned out that there were only three times. If there was no transaction cost, then according to the *Coase Conjecture* (Coase (1972)) Fed could choose any level of interest rates as targets without generating unnecessary financial market fluctuations. But in the real world there are so many transaction costs such that any unanticipated monetary policies could have harmful effects for market participants.

The purpose of the government or the central banker is assumed to maximize the welfare of the household, which is represented by the rent households obtain by using the fiat money. Again the household has two options in using money. First, she could use it to buy goods. In this case the fiat money provides her with liquidity services. Otherwise she could lend the money to someone else who needs it to consume goods or invest in bonds and stocks. In this second case she abandons the right to use the money in exchange for some interest returns. With competitive uses from numerous households the rent would be equal to the interest income in equilibrium. The central bank therefore wishes to maximize the rent, or equivalently the interest income, net of the transaction cost of implementing the mandates: $(R'-C^g)M0$, where R' is the nominal deposit rate and $C^g > 0$ is the relevant transaction cost. We assume that the central bank would follow the advice of Friedman (1969) to choose the quantity of money (M0 here) instead of interest rates as the instrument of the monetary policy.

The solution to the central bank's maximization problem is simply $R' = C^g$, that is in equilibrium the transaction cost is just equal to the opportunity cost of forgoing the right to use the money to other people. The novelty here is that this result could help us explain the optimum quantity of money proposed by Friedman. In his famous statement Friedman (1969, p. 34) had claimed that "[*T*]*he optimum quantity of money is that it will be attained by a rate of price deflation that makes the nominal rate of interest equal to zero.*" (italics original) Although the process of deriving the optimum

quantity of money is without error, the assumption Friedman used is not. Assuming zero nominal interest rate is equivalent to assuming that there is no right to use the money, or that the right to use money cannot be transferred to someone else. This amounts to saying that only the person who originally owns the money can use it. Nobody can use the medium of exchange except the owner of the money. This is a world with zero velocity of money. And in the end the money is useless because it has lost its role as medium of exchange and as store of value.

To have a meaningful monetary equilibrium, be it optimum or not, a positive nominal interest rate would be necessary. If there were no transaction cost of supplying money, there would not be nominal interest rate, at least in the sense of Friedman's optimum quantity of money. The equilibrium nominal interest rate in our model is $R' = C^g > 0$,[45] which is consistent with the later work of Friedman (1986). He finally admitted the importance of the resource cost of issuing irredeemable paper money.

IV. Applications of the Quality Theory

1. Subprime Mortgage Crisis

The quantity theorist has focused on the stability of money demand, as described by Friedman (1969, pp. 62-63) in his restatement of the quantity theory:

The quantity theorist accepts the empirical hypothesis that the demand for money is highly stable…A stable demand function is useful precisely in order to trace out the effects of changes in supply…The classical version of the objection under this head to the quantity theory is the so-called real-bills doctrine: that changes in the demand for money call forth corresponding changes in supply and that supply cannot change otherwise, or at least cannot do so under specified institutional arrangements.

It is therefore no doubt that quantity theorists must verify that the demand function for money is stable. If the quantity theory is based on the assumption of stable money

[45] This is the equilibrium nominal deposit rate. The equilibrium nominal loan rate would in general be higher than the deposit rate, as shown above. In any case the equilibrium nominal interest rates in our model are positive. If there are no transaction costs, then both the nominal deposit rate and the nominal loan rate would be zero in our equilibrium, which is consistent with Friedman's optimum quantity of money. Friedman's theory is therefore a special case of ours.

demand, then the quality theory is concerned with the probable instability of money demand. This instability came from the cost of bringing lender and borrower together and the resulting separation of market-made loans from homemade leverage. We can use the CDO in subprime mortgage crisis as an example to explain how the quality theory works, and why the quantity theory and the corresponding policy such as QE could not solve the problem of boom-bust cycles.

In our theory the loanable funds are an intermediate input to the production of the final good. But in early 2000s mortgage played an important role in loanable funds markets. This is partly because the government of the United States promoted the idea of homeownership, and partly because the financial instrument CDO was invented. A nice story of this process has been told by Michael Lewis, as he said:[46]

> In the process, Goldman Sachs created…the synthetic subprime mortgage bond-backed CDO, or collateralized debt obligation…The rating agencies, who were paid fat fees by Goldman Sachs and other Wall Street firms for each deal they rated, pronounced 80 percent of the new tower of debt triple-A…The goal of the innovation, in short, was to make the financial markets more efficient. Now, somehow, the same innovative spirit was being put to the opposite purpose: to hide the risk by complicating it. The market was paying Goldman Sachs bond traders to make the market less efficient.

Without mortgages or loans there would be no mortgage backed bonds, and without these bonds there would be no CDOs. Loans thus became an intermediate input to the production of CDO instead of real investment. And as claimed by Lewis, the risk of CDO was hidden by Wall Street firms, and the resulting transaction cost was passed on to investors and taxpayers if these security firms were considered by government as too big to fail.

In standard macroeconomics the purpose of financial instruments is to make markets more efficient. The risk would be shared with or transferred to other people through the functioning of financial markets. Stock, bond, insurance, CDO, and many others are all financial instruments to make sure that the risk will be reduced and the pie will thus be bigger. But at least for CDO, it seems that this is not the case. The risk was not shared but hidden, and this made investors much less informative. The rise of the cost of information eroded the quality of CDO and that of loans, its intermediate input. The production function of these investors would therefore shift to the left and, at the same time, the production function of Wall Street firms shifts to the right. Total output in the society shrinks because the number of investors and taxpayers are much

[46] Lewis (2011, pp. 72-74).

larger than that of bond traders in Wall Street firms.

The case of CDO reminds us of the importance of the quality of money. When almost everyone could get mortgage with a very low loan rate, as in the era of the Great Recession, the quality of the loan was declining, even though the quantity of it was increasing. When both the quality and the quantity of loans changed, the demand for loanable funds could not have been stable. Without stable money demand function the quantity theory cannot provide the central bank with a credible guidance to its monetary policy. The central bank should put in mind not only the quantity but the quality of money.

2. QE: Japan versus the United States

After the burst of the Heisei bubble in 1990, Japan had fallen into a long and great recession. The Bank of Japan, under suggestions of Paul Krugman and Ben Bernanke, launched the world's first QE in March 2001.[47] Seven years later, Bernanke's Fed did the same thing. In late 2008, after the burst of subprime mortgage bubble, the Fed initiated the first QE in the United States. The Fed purchased not only government bonds but mortgage based securities (MBS). As mentioned by Lewis (2011, p. 73), the rating agency gave triple-A to more than 80 percent of the debt rated. The quality of many of these debts and the derivatives based on them, such as CDOs, was definitely not good enough to get a triple-A. What the Fed cared about was, of course, the quantity of money. To provide unlimited liquidity service to the market was the first priority in the mind of many Fed officials. This is not all wrong, but it is not all right, either. Without paying more attention to the quality of money, the implementation of QE was at the cost of at least two things. First, as an almost unlimited lender of the last resort, the Fed would have encouraged security firms in Wall Street to take more risk in designing financial assets more toxic than CDO and CDS,[48] and then pass it on to innocent investors and taxpayers to earn a pure rent.[49]

The second social cost of QE was that it might have encouraged the speculation in stock, housing, and high-yield bond markets. When huge amount of money through QE was borrowed from the bank to invest in these asset markets, the Fed, in effect, had provided the projector with the easiest money to make a speculative bubble. If

---

[47] Koo (2009, p. 73).

[48] In addition to CDO, the CDS also played an important role in the subprime mortgage crisis. See, for example, Lewis (2011).

[49] Keynes (1936, p. 144) had recognized the importance of this moral hazard problem when he was discussing the cost of bringing lender and borrower together.

firms in Wall Street are too big to fail, then firms in Main Street would be too many to fail. In either case the central bank should be more concerned with the quality, rather than the quantity, of both outside and inside money to avoid the next financial bubble.

3. Cash: A Curse or a Blessing?

In a world with the internet of everything, the transaction cost will be reduced, but it will never disappear. It just transforms into different types. As first mentioned by his paper on "The Nature of the Firm," Coase told us that there are three types of costs when we prepare to make transactions, as he said in the following passage:[50]

> The main reason why it is profitable to establish a firm would seem to be that there is a cost of using the price mechanism. The most obvious cost of "organizing" production through the price mechanism is that of discovering what the relevant prices are. This cost may be reduced but it will never be eliminated by the emergence of specialists who will sell this information. The costs of negotiating and concluding a separate contract for each exchange transaction which takes place on a market must also be taken into account.

The cost of using cash is not just the cost of replenishing cash as in the Baumol-Tobin model. The implementation of the third-party payments, mobile payments, and digital currencies, among others, needs the establishment of relevant economic and legal frameworks. These are costs of enforcing contracts. The popularity of mobile phones and other mobile devices is a necessary condition for the functioning of a world without paper money. In such a world the inside money, or credit, will become more and more important, and the role of the central bank will be narrowed because the amount of the high-powered money would be reduced. For some economists, such as Rogoff (2016), cash might be a curse because it would facilitate corruption and make monetary policy of the central bank more limited because there will always be a ZLB. If cash were banished, then people could not hoard money, and they would be forced to spend it. This means that the bank could pay a negative nominal interest rate when people would still like to deposit savings in the bank just for, say, security.

Rogoff's arguments were based on the assumption that people cannot use digital currency to make corruptions easily, and they also cannot hoard their money without using cash. These arguments might not be all right. With digital currency, such as the bitcoin, some of China's corrupted money has been secretly transferred abroad

[50] Coase (1988, pp. 38-39).

because renminbi is still not an international currency. Bitcoin is more internationally accepted than Chinese yuan. This was one of the reasons why China would banish bitcoin last year. This is not the case for corrupted money in the United States because U.S. dollar is much more internationally accepted than bitcoin. It is more costly to use bitcoin to hide the corrupted money than to use the U.S. dollar. A curse for America might therefore be a blessing for China, if no digital currencies, such as bitcoin, are allowed to be used in China. Once again, the quality of money matters. The popularity of the U.S. dollar has been the result of America's open and free financial markets. In other words, the quality of U.S. dollar is better than that of the Chinese yuan. To improve the quality of yuan, China at least has to open its financial markets.

4. Optimum Quantity of Money: ZLB or RZLB?

Friedman's optimum quantity of money is where the nominal interest rate is reduced to zero by central bank's deflationary policy. Zero nominal interest rate could also exist when the central bank implements an inflationary policy such that real interest rate becomes negative. In this case, hoarding cash is not a good idea, and the optimum quantity of money should be where the demand for money is zero. When optimum quantity is restricted to be positive, as required by Friedman (1969), the monetary policy of the central bank should be deflationary. This conclusion is somehow controversial. The existence of ZLB indicates that the nominal interest rate is near, but seldom equal to, zero. Even if the nominal interest rate is approaching ZLB, it is often a state in deep recession, and it is hard to call it an optimum. A meaningful optimum quantity of money should better be in a state where both the quantity of money and nominal interest rate are greater than zero, and the central bank is pursuing a mildly inflationary monetary policy. As central banks all over the world are targeting the 2% inflation rate, it naturally implies that the optimum quantity of money will be more consistent with inflationary than deflationary monetary policies. And if, according to Friedman, the optimum real interest rate is positive, then with a positive inflation rate, the optimum nominal interest rate should be greater than zero.

As shown in Section III, optimum nominal interest rate is greater than zero only if there are transaction costs when central bank issues the irredeemable paper money. This means that Friedman's conclusion in 1969 should be modified if his conclusion in 1986 was accepted. It took seventeen years for Friedman to tacitly recognize that the optimum quantity of money should be in such a state that both the nominal and real interest rates are positive, and the central bank should be pursuing very moderate

inflation.[51] All of the results can be illustrated in Figure 4.

The quality of money can, as usual, shed some light on the problem of optimum quantity of money. Friedman (1969) thought that the central bank should not pay any interest to cash holders except that there is inflation. When there is deflation, holding cash would have benefits which would be balanced with the loss of time preference due to delayed consumption. This argument is true, but what if there are transaction costs in the process between printing and hoarding, as well as between hoarding and spending money?

As argued later by Friedman (1986), there are resource costs of printing fiat money. In addition, according to Keynes (1936), there is cost of hoarding money such that it should be compensated with a liquidity premium. Finally, when we spend money, the same Keynes told us that there would be the cost of bringing lender and borrower together. All the above costs are transaction costs. The first is associated with money supply, the second with money demand, and the third with both supply and demand. Because interest is the cost of foregoing the right to use money, if these transaction costs are positive, there is no reason why the interest rate, nominal or real, should be reduced to zero. As mentioned above, the optimum rule for Friedman was zero nominal interest rate, and for Allais it was zero real interest rate. Both ignored positive transaction costs associated with the demand and supply of money.

Coase (1972) taught us that the durable goods monopolist has an incentive to charge additional customers a lower price. As the central bank is the monopolist of issuing the durable good called money, it has also the incentive to print more of it. The policy of QE or QQE in Japan and the United States were standard examples for such a privilege the central bank would have. QE has generated a credibility problem for the country which used it. The more money the central bank issues, the less credible the value of the money would be, and the quality of money would be worse.

In the era of QE some central banks have adopted the policy of negative nominal interest rates. This policy should not be an optimum in any useful sense, but it was supported by many central bankers and scholars such as Rogoff (2016). If it were adopted by more and more countries, then the ZLB would never be a problem. But according to the arguments of this paper, we do not think that it will be an ordinary monetary policy for most countries because the true lower bound for the central bank

[51] To my best knowledge, Friedman had mentioned transaction costs only three times, and two of these were in Friedman (1969, 1986), respectively. Another time was in a 1953 *Journal of Political Economy* paper, when he was discussing inequality and income distribution.

is quite likely the real zero lower bound (RZLB), rather than the usual ZLB.

The reason is simple. The price for the loanable funds market is the real interest rate because the ultimate return of loans would come from real investment. In terms of Fisher or Keynes, it is the rate of return over cost, or MEC, that matters. The cost would never be negative such that the return over cost would never be negative, too. If there are transaction costs, as in the real world, then the total cost of using money must be positive. This means that the equilibrium real interest rate in the loanable funds market must be positive, too. Otherwise, there would be no loanable funds market. Both the operation of the market and the determination of the price need cost. A positive price is the necessary condition for the market to exist. In the case of the loanable funds market, the relevant price is the real interest rate.

Unless the central bank is pursuing a deflationary monetary policy such that the inflation rates would generally be negative, a positive real interest rate would, in general, imply a positive nominal interest rate. This is what we observe in most of the time. If a positive real interest rate is a normal situation for the loanable funds market, then the lower bound for the real interest rate should naturally be zero. This is why we have claimed that what really matters is the RZLB, rather than the ZLB emphasized by many economists.

V. Conclusions

Monetary system is perhaps the most important economic system. Business cycles are largely a monetary phenomenon. The debate between the quantity theory of money and the real bills doctrine has been the focus of the monetary theory ever since the time of Hume and Smith. Smith's real bills doctrine is, in effect, a quality theory of money. Misled by Mints, Friedman considered real bills doctrine as fallacious. This is why most modern monetary economists have been adherents to the quantity, rather than the quality, theory of money.

In this paper we establish a simple theoretical model to elaborate the quality theory of money. We see this problem through different angles. We also compare our model with various monetary theories proposed by economists such as Smith, Friedman, and Keynes, and with facts in the real world. In all cases we found that transaction costs are the key to understanding the quality of money. Much more empirical studies are needed to see if transaction costs are as important as they are in our theoretical model.

References

Baumol, William, "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quarterly Journal of Economics*, 1952, 545-556.

Coase, Ronald H., "Durability and Monopoly," *Journal of Law and Economics*, 1972, 143-149.

--- *The Firm, the Market, and the Law*, University of Chicago Press, 1988.

Eichengreen, Barry, *Hall of Mirrors*, Oxford University Press, 2015.

Friedman, Milton, *The Optimum Quantity of Money and Other Essays*, Aldine, 1969.

--- "The Resource Cost of Irredeemable Paper Money," *Journal of Political Economy*, 1986, 642-647.

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

Goodfriend, Marvin, and Bennett T. McCallum, "Banking and Interest Rates in Monetary Policy Analysis: A Quantitative Exploration," *Journal of Monetary Economics*, 2007, 1480-1507.

Hébert, Benjamin, and Michael Woodford, "Neighborhood-Based Information Costs," *American Economic Review*, 2021, 3225-3255.

Hume, David, "Of Money," in *Political Discourses*, Kincaid and Donaldson, 1752.

Keynes, John Maynard, *A Tract on Monetary Reform*, Macmillan, 1923.

--- *The General Theory of Employment, Interest, and Money*, Macmillan, 1936.

Koo, Richard C., *The Holy Grail of Macroeconomics*, revised edition, Wiley, 2009.

Lewis, Michael, *The Big Short*, paperback edition, Norton, 2011.

Lucas, Robert E., Jr., "Glass-Steagall: A Requiem," *American Economic Review*, 2013, 43-47.

--- and Juan P. Nicolini, "On the Stability of Money Demand," *Journal of Monetary Economics*, 2015, 48-65.

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

Modigliani, Franco, and Merton H. Miller, "The Cost of Capital, Corporate Finance and the Theory of Investment," *American Economic Review*, 1958, 261-297.

Rognlie, Matthew, Andrei Shleifer, and Alp Simsek, "Investment Hangover and the Great Recession," *American Economic Journal: Macroeconomics*, 2018, 113-153.

Rogoff, Kenneth S., *The Curse of Cash*, Princeton University Press, 2016.

Romer, Paul M., "Endogenous Technological Change," *Journal of Political Economy*, 1990, S71-S102.

Samphantharak, Krislert, and Robert M. Townsend, "Risk and Return in Village Economies," *American Economic Journal: Microeconomics*, 2018, 1-40.

Sargent, Thomas J., “Where to Draw Lines: Stability versus Efficiency,” *Economica*, 2011, 197-214.

Schulhofer-Wohl, Sam, “Heterogeneity and Tests of Risk Sharing,” *Journal of Political Economy*, 2011, 925-958.

Smith, Adam, *An Inquiry into the Nature and Causes of the Wealth of Nations*, 5th edition, 1789; Modern Library, 1994.

Stiglitz, Joseph E., “The Theory of Credit and Macroeconomic Stability,” National Bureau of Economic Research Working Paper 22837, 2016.

Tobin, James, “The Interest-Elasticity of Transactions Demand for Cash,” *Review of Economics and Statistics*, 1956, 241-247.

--- “Money, Capital, and Other Stores of Value,” *American Economic Review*, 1961, 26-37.

Wallace, Neil, “A Modigliani-Miller Theorem for Open-Market Operations,” Federal Reserve Bank of Minneapolis Staff Report 44, 1979.

Woodford, Michael, “Financial Intermediation and Macroeconomic Analysis,” *Journal of Economic Perspectives*, 2010, 21-44.

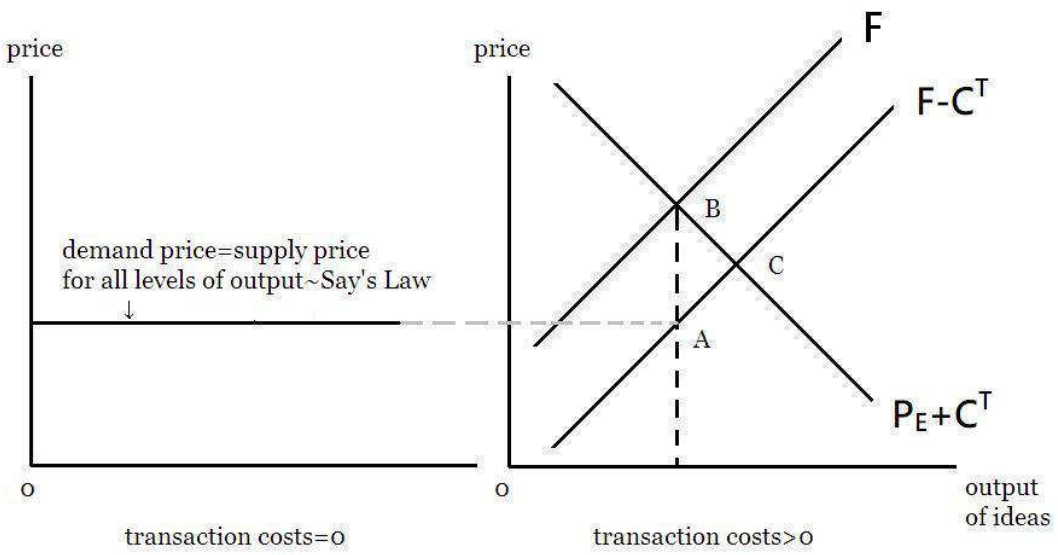

Figure 1: The Market with and without Transaction costs

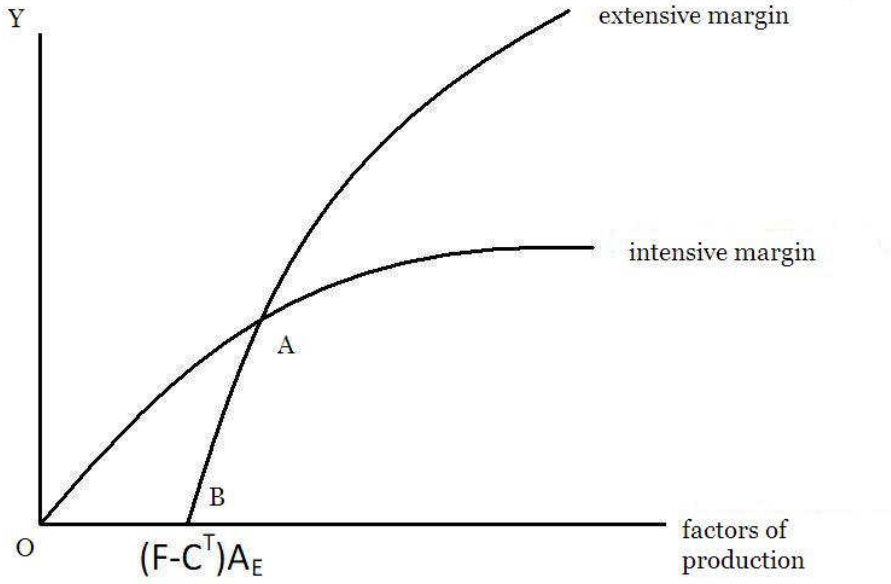

Figure 2: The Extensive and Intensive Margins

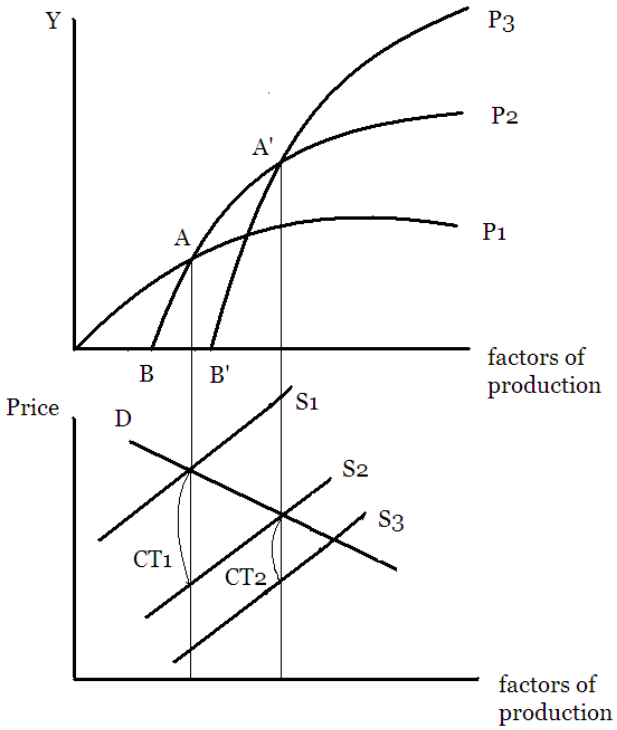

Figure 3: Transaction Cost and the Demand Curve

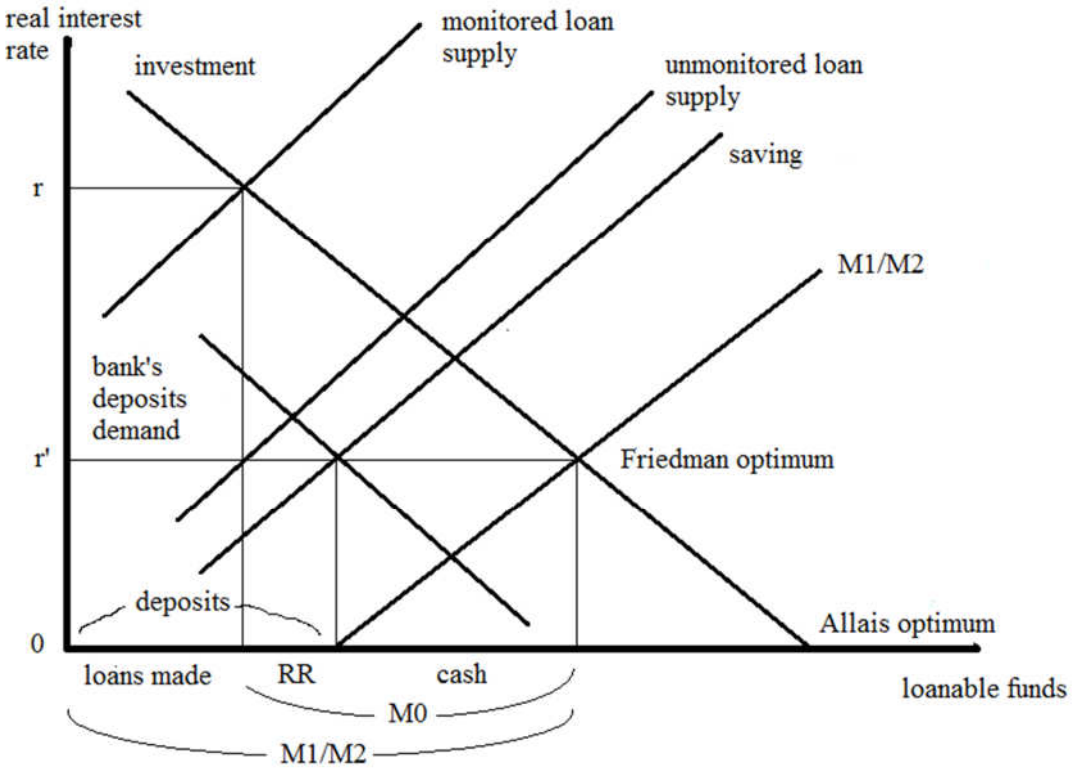

Figure 4: The Quality Theory of Money