Should We Subsidize the Use of Currency?

Jeffrey M. Lacker

Three types of money are provided monopolistically by the government in the United States: coin, currency, and reserve balances with Federal Reserve Banks. Together, they make up the monetary base. Although it is common to think of government money as virtually costless to produce, the real resource costs are substantial, as shown in Table 1. In 1991 the cost of providing currency was $393.2 million, most of which was incurred in verifying and sorting deposits of used currency and replacing unfit notes with new currency.

Under current arrangements, banks can deposit used currency and withdraw fit currency at par. Thus currency costs are not borne directly by banks, but instead are funded out of general government revenues.[1] True, restrictions on banks' deposit and withdrawal of currency help limit Federal Reserve costs, but these costs are not borne by users. In contrast, all of the costs associated with the provision of reserve balances are recovered through "service fees," as mandated by the Monetary Control Act of 1980.[2]

The author is associate research officer of this Bank. For helpful comments on an earlier version, the author would like to thank Mike Dotsey, Marvin Goodfriend, Bob Hetzel, Tom Humphrey, Pete Ireland, Jim Reese, and John Weinberg. Thanks to Bruce Champ and Warren Weber for helpful conversations. The views expressed do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Costs associated with currency and coin operations directly reduce the Federal Reserve's payment to the U.S. Treasury. The Fed buys newly printed currency at cost from the Bureau of Engraving and Printing and newly minted coin at face value from the United States Mint.

[2] The Monetary Control Act of 1980 mandated fees for Federal Reserve services, including "currency and coin services" (12 USC §248a), but this phrase is interpreted to mean only coin wrapping and currency and coin transportation services, which are omitted from the tables. See Board of Governors (1980).

Table 1 Costs of Components of the Monetary Base, 1991
Millions of dollars

	Currency	Coin	Reserve Balances
Operating expenses	139.9	24.5	140.8
Replacement costs	253.3	44.0	0
Total	393.2	68.5	140.8
Recovered from fees	0	0	140.8
Unrecovered	393.2	68.5	0

Sources: See Appendix B.

This article examines whether the Fed should continue to subsidize the use of currency. In particular, I argue that the Fed should charge a currency deposit fee, effectively paying less than par when converting currency into reserve balances, and should remove the rationing constraints on currency deposits and withdrawals. This recommendation should not be surprising, since it follows from standard economic reasoning. The costs of a government-provided service should, in general, be paid by the users to ensure that use is efficient.

However, I argue that a partial subsidy is desirable, in the sense that the deposit fee should be less than marginal cost. The reasoning is again standard. There is likely to be a "market failure" that makes the private willingness to pay for fit currency less than the social benefits of fit currency. The market failure arises because currency generally trades at par, regardless of the quality of the note, due to the inconvenience of quality-adjusted currency prices. As a result, willingness to pay is less than the social benefits and the Fed should subsidize the provision of currency quality by charging less than marginal cost. In no case, however, should the deposit fee be zero; efficiency of the division of currency processing between banks and the Fed requires a strictly positive fee.

In what follows I focus solely on currency policy, even though coin use is similarly subsidized and the arguments against its free provision apply with equal force.

1. SOME BACKGROUND ON THE MANAGEMENT OF GOVERNMENT MONEY[3]

Depository institutions can hold reserve account balances at a Federal Reserve Bank. These reserves are book-entry demand deposits that can be transferred

[3] See Booth (1989) for a historical review of Federal Reserve currency and coin operations.

to other banks. Reserve account balances are also used for automated clearinghouse transactions in which recurring payments are made via prearranged wire transfers. Federal Reserve Banks charge fees for transfers of reserve balances and generally recover all of the associated costs.

A branch or office of a bank may withdraw currency and coin from a Federal Reserve Bank, deducting the par value of the withdrawal from its reserve balance. Deposits of currency and coin are credited at par. In both cases the bank pays for transportation, generally via an armored carrier service.

Incoming deposits of currency at Federal Reserve Banks are processed on high-speed equipment that removes wrong denomination and counterfeit notes and verifies the number of notes in the deposit. In addition, the equipment removes and destroys "unfit" notes that have become soiled in circulation.[4] The remaining "fit" notes are repackaged and stored. Withdrawals of currency are met with fit used notes and newly printed notes from the Bureau of Engraving and Printing. Banks are not allowed to request new notes, but must accept the mix that is sent to them. Coins are deposited and withdrawn in bags of standard size and are verified by weighing.

Banks can process, sort, and reuse currency themselves. Tellers count and sort currency, holding some for future use, and send unsatisfactory notes, suspected counterfeits or "excess" notes to the Fed. Banks with more than one branch sometimes process currency centrally at a "cash room," collecting currency from branches with net inflows, and disbursing it to branches with net outflows. Some banks sell currency directly to other neighboring institutions. Some private transfers of currency are intermediated by armored carrier companies. The same high-speed currency-processing equipment used by the Federal Reserve is available to private institutions, some of which, particularly larger institutions, run currency-processing operations similar to the Fed's (although they do not destroy worn currency).

[4] The U.S. Treasury defines unfit currency as "currency which is unfit for further circulation because of its physical condition such as torn, dirty, limp, worn or mutilated." 31 CFR 100.5. Federal Reserve Banks have adopted more detailed definitions.

> "Paper currency tendered for redemption in order to be classed as fit for further circulation must be fairly clean so that its class, denomination, and genuineness can be determined without difficulty and must contain a sufficient amount of life or sizing to permit its being handled with facility. It should not contain heavy creases which break the fiber of the paper and indicate that disintegration has begun. A fit note when held by one end in one hand and pressed into a slightly concave shape lengthwise should sustain itself substantially on a line with the hand. It should not present a limp or rag-like appearance. If a note has retained a fair amount of the original strength or sizing, it is fit unless it is so badly soiled as to be offensive, or it is torn, perforated or otherwise mutilated. Mere creasing or wrinkling that has not broken nor seriously weakened the note does not make it unfit. So-called 'dog ears' or bent corners do not render notes unfit." (Federal Reserve Bank of Richmond, *Operating Circular No. 14*, March 30, 1990, p. 3)

Federal Reserve Banks place restrictions on the deposit and withdrawal of coin and currency by banks. The restrictions vary somewhat across Federal Reserve offices, but all follow a set of guidelines adopted by the Board of Governors (Board of Governors 1984). For example, some Federal Reserve Banks offer access to only one bank office per municipality. In large metropolitan areas this constraint can force a bank to consolidate cash from its branch network to its own cash room operation from which all deposits and withdrawals are made. In contrast, a bank operating in a suburban or rural environment can receive service at many of its branches. These constraints, called "access controls," effectively mandate significant private processing for banks in certain geographical areas.

Another key restriction is the prohibition of "cross-shipping"—the "deposit of excess fit currency and re-order of the same denomination within 5 business days." This restriction prevents a branch from depositing its incoming currency without sorting and then obtaining fit currency from the Fed to meet its withdrawal demand. Both access and cross-shipping constraints are explicitly aimed at limiting the volume of Fed currency processing. There is no empirical evidence on the extent to which these constraints reduce such processing volume, but anecdotal evidence suggests that it is quite substantial in some areas.

2. THE POLICY QUESTION

Currency, coin, and reserve account balances are the three components of the monetary base. The essential policy question concerns the terms on which the components of the monetary base can be exchanged for each other. Under current policy the Fed stands ready to exchange any component of the monetary base for any other component at a fixed relative price of one (par) plus transportation costs. The division of the base among currency, coin, and reserve account balances is thus demand determined, while the total monetary base is determined by monetary policy via open market operations.

As a consequence of the current policy and the fragmented structure of the U.S. depository industry, the Fed acts as an intermediary in the flow of currency. Some bank branches experience net inflows of currency deposits, accumulating currency which they wish to exchange for reserve account balances, either because their vault space is limited or because they want to convert excess reserves to loanable funds. Some bank branches experience net outflows of currency which they need to replace from reserve account balances. The Federal Reserve has a substantial inventory of currency, amounting to 21 percent of outstanding currency at the end of 1991, for example.[5]

[5] Out of $366 billion Federal Reserve notes outstanding on December 31, 1991, Federal Reserve Banks held $78 billion, leaving $288 billion in the hands of banks and the public (Board of Governors 1992, p. 246).

Table 2 Currency Costs, 1991
Millions of dollars, except where otherwise noted

Paying and receiving	41.2
High-speed processing	50.9
Other processing	4.4
Overhead	43.3
Total processing costs	139.9
Printing new currency	253.3
Total costs	393.2
Number of notes received from circulation, *millions*	19,613
Cost per note, *dollars.*	.0200

Sources: See Appendix B.

When a Federal Reserve Bank accepts a deposit of currency from a bank, in effect it *buys* notes in "straps" of 100 notes each. When a Federal Reserve Bank pays out currency to a bank, in effect it *sells* straps of currency. The objects sold by the Fed—verified straps of fit, genuine notes ("sorted straps")—are different from the objects bought by the Federal Reserve—unverified straps of fit and unfit notes, including some wrong denomination and counterfeit notes ("unsorted straps"). The two are distinctly different economic commodities, just as aluminum is different from bauxite and flour is different from wheat. Real resource costs must be incurred in receiving and storing unsorted straps, in processing currency, in printing new replacement notes, and in paying out currency withdrawals. From this point of view, Federal Reserve currency operations are a productive economic activity, consisting of the transformation of unsorted straps of mixed-quality currency into sorted straps of fit currency. Table 2 provides a breakdown of Federal Reserve Bank currency-processing costs.

The "price" of Federal Reserve processing is the difference between what the Fed charges for withdrawals of sorted straps of currency and what it pays for deposits of unsorted straps. Access and cross-shipping controls act as rationing devices to limit the demand for Federal Reserve processing. The policy issue, then, is a classic one. On what terms should a publicly provided service be offered? Should it be provided free of charge with quantitative restrictions to ration demand? Or should it be offered at a price without quantitative restrictions?

3. WHAT IS WRONG WITH THE CURRENT POLICY?

Under the current policy of rationed free provision—buying and selling currency at par with quantitative restrictions on deposits and withdrawals—banks' currency-handling decisions are likely to be socially inefficient. For example,

a bank with a large branch network could operate a cash room, incurring the associated costs of processing. Alternatively, the bank could have branches deposit and withdraw directly to and from the Fed, avoiding the cost of processing altogether. By buying and selling currency at par, the Fed subsidizes central bank processing relative to private currency processing. Some currency processing now performed by the Fed is socially wasteful.

The quantitative restrictions embodied in access and cross-shipping constraints attempt to prevent overuse of Federal Reserve processing. For example, restrictions on the number of bank branches served in a given geographic locality prevent duplicative shipping to and from the Fed when direct currency transfers between branches would be less costly. Similarly, cross-shipping restrictions prevent some branches from having the Fed process currency that they could process at lower cost themselves.

However, quantitative restrictions are generally less efficient as a rationing device. If the Federal Reserve knew the costs of various alternative private currency-processing arrangements, it could design quantitative restrictions which would exactly replicate the effect of efficient prices. Unfortunately, the necessary information is costly to acquire, and implementable rules cannot be too complex or subjective. As a result, the simple rules in place are often unrelated to the real resource costs of currency processing.

For example, two banks might have neighboring branches that qualify for access, but if the banks merge only one branch would qualify. The real relative costs of currency processing would not have changed, but access controls would treat the two situations differently. Similarly, cross-shipping constraints affect banks unequally, depending on their typical net deposit flows. A branch with equal currency deposits and withdrawals is forced to process currency itself. By contrast, two neighboring branches with complementary currency flows—one receiving net inflows, the other meeting net outflows—have no incentive to sort and transfer currency together, since they can have the Fed process it free. Cross-shipping thus encourages private processing at branches with balanced currency flows but not at branches with net inflows or outflows.

4. PRICING WOULD BE BETTER THAN QUANTITATIVE RESTRICTIONS

An attractive alternative to rationing is to provide currency processing at a positive price without quantitative restrictions. One way to do this is to impose a fee for depositing currency.[6] Economic theory tells us that the price of a publicly

[6] An alternative method would be to charge a fee for withdrawals of sorted fit currency. Any combination of deposit and withdrawal fees are possible as long as the price at which the Fed sells currency is greater than the price at which the Fed buys it. I will restrict attention to a deposit fee.

provided good should be set equal to the marginal social cost of production.[7] This suggests that a currency deposit fee should be set equal to the marginal cost of Federal Reserve currency processing. As a rough guide to the likely magnitude of such a price, Table 2 shows that the average total cost of Federal Reserve currency operations is 2 cents per note. Under constant returns to scale, marginal cost would equal average cost implying a deposit fee of 2 cents per note.[8]

A deposit fee equal to the marginal cost of Fed processing would have the desirable property that banks' decisions about currency handling would no longer be biased towards Fed processing. If operating its own cash room is less costly to a bank than sending currency to the Fed, that is because private processing is socially less costly than Fed processing. Conversely, if Fed processing is less costly to the bank than processing currency itself, it is because Fed processing is socially less costly than private processing. The cost of acquiring sorted currency from the Fed, relative to the cost of obtaining privately sorted currency, would reflect actual relative resource costs. Appendix A describes a simple model that illustrates this point. The model shows that in the absence of any "market failure," setting the deposit fee equal to the Fed's marginal cost results in the socially optimal allocation of currency processing.

A currency deposit fee would allow the elimination of quantitative restrictions on currency deposits and withdrawals, and would let banks decide on the cost-minimizing pattern of use. Banks would determine whether it is less costly for a particular branch to obtain currency directly from the Fed, from other branches, or from central cash rooms. They would determine the cost-minimizing frequency of shipments and whether cross-shipping is the least costly arrangement. The potential overuse of Federal Reserve processing would be curbed, since banks would efficiently ration such processing themselves. In addition, and perhaps more important, the resulting volume of Fed currency processing would be allocated efficiently across banks. The substitution of a deposit fee for quantitative restrictions also would decentralize cost-minimization decisions and let banks assess whether Fed or private processing is more efficient.

5. WOULD THIS INTERFERE WITH PAR VALUE EXCHANGE OF CURRENCY?

One advantage of paper currency in retail transactions is that it trades at par value, which can be ascertained at a glance.[9] In contrast, in many commodity

[7] See Bös (1985), for a survey.

[8] Using three years of data by Federal Reserve office, Michael Dotsey (1991) estimates long-run cost functions for high-speed sorting and for currency paying and receiving. He finds economies of scale in both activities, with scale economy coefficients (elasticity of total cost with respect to output) of .92 and .84, respectively. This finding implies that marginal cost is about 1.7 to 1.8 cents per note.

[9] See Fama (1983) for a discussion.

money schemes it was costly to establish the value of money being tendered in exchange. For example, gold coins often would be weighed and assayed, with the coins valued on the basis of the quantity of the gold they contained. One would expect to see merchants offer discounts for less costly forms of payment. Gasoline retailers sometimes offer discounts for purchases by cash, check, or debit card, reflecting the costs of the float associated with credit card purchases.[10] If the Federal Reserve paid less than par for deposits of currency, then a merchant's currency deposit would be more costly for a bank to accept. Would banks impose new fees for currency deposits, passing along the cost of the deposit fee charged by the Fed? Would merchants then charge a premium for accepting payment in currency or offer a discount for non-cash payments? Perhaps, but it provides no reason not to charge for currency deposits at the Fed. The Fed charges for wire transfer payments and allows banks to decide whether to pass the charges on to their customers. Similarly, efficiency would be maximized by allowing banks to decide whether to pass on currency deposit fees.

Merchants and banks already face many costs of handling currency, but choose not to pass them on to consumers. For example, both currency and checks are accepted at par in retail transactions, despite the fact that their costs are very different. As it is, banks pay the cost of transporting currency to and from the Federal Reserve, a fee than can be as high as 4 cents per note—twice as high as the 2 cents per note deposit fee proposed above. Banks do not charge consumers directly for cash transactions, but often charge retail merchants for currency transactions. Retail businesses do not directly pass these costs along to customers offering cash; instead they incorporate them into their overall cost of doing business and assess customers only indirectly. In essence, we already have nonpar currency at the wholesale level, coexisting with par currency at the retail level. Such evidence suggests that currency would still generally exchange at par under a deposit-fee scheme.[11]

A distinct but related question concerns the variation in quality across notes. Par value exchange implies that old worn notes exchange one-for-one with crisp new notes of the same denomination, despite the inferiority of worn notes. If the Federal Reserve paid less than par for used currency deposits, then accepting unfit currency would be more costly for a bank than accepting fit. Would banks impose fees for deposits of unfit currency, passing along the cost of the deposit fee charged by the Fed? Would old, worn-out currency be discounted in retail transactions, trading at quality-dependent prices like used cars? Would the benefits of improved efficiency in currency processing be outweighed by the higher transactions costs of quality-dependent prices?

[10] See Barron, Staten and Umbeck (1992).

[11] Curtin (1983) reports evidence on the costs of different payment media to retail establishments and why so few offer cash discounts.

It seems unlikely that our currency would trade at quality-dependent prices under a deposit-fee scheme. First, it obviously would be cumbersome and inconvenient for the price of a note to vary with the note's physical condition. The quality of each note tendered in a transaction would have to be assessed individually and a value mutually agreed upon by both parties. Since quality is difficult to define objectively, this could be a contentious process. The number of computations required for a typical transaction would be larger, since the quality-adjusted price of notes would have to be calculated both for the notes tendered and for the notes given in change. Furthermore, merchants would have to keep track of each note's quality, either segregating notes or labeling them. For retail transactions it seems far more convenient for currency to trade at face value.

If everyone else is accepting currency at prices that do not depend on note quality, then your incentive to pay quality-dependent prices is minimal. You might be willing to pay for the lower inconvenience provided by a high-quality note while it is in your possession, but you will be unwilling to pay for the added convenience to the person that accepts it from you at par. Quality differentials would have to become relatively large to induce quality-dependent currency prices. Therefore, one reason we do not see quality-dependent currency prices is that the benefits—improved resource allocation resulting from more accurate relative prices—are less than the associated transactions costs.

Part of the reason quality differentials are not large enough to induce quality-dependent prices is that the Fed essentially supports the price of old currency relative to new currency. The Fed buys old currency at par, withdrawing unfit notes from circulation and replacing them at par with new notes. Worn notes are tendered to the Fed rather than exchanged at less than par. As a result, the range of quality of notes in circulation is not large enough to make quality-dependent pricing worthwhile. In most private durable goods markets, in contrast, producers supply new goods but do not intervene in the market for used goods. For example, automobile manufacturers sell new cars but do not generally intervene in the market for used cars.[12] The prices of used cars adjust to reflect the value of old cars relative to new cars. If automobile manufacturers supported the price of their used cars at close to the price of their new cars, they would be forced to buy many old cars. The quality of the stock of cars in use would be far higher on average and far more uniform, perhaps even trading "at par." Without manufacturer intervention, many consumers would pay large sums to have an old car replaced with a new one. Under current Fed policy, low-quality notes are rarely so bad that a consumer would be willing to pay more than a small fraction of its value to have it replaced with a new one.

The legal tender provision of the Coinage Act is another reason currency is likely to keep trading at par. The law prevents creditors from refusing to take worn currency at par. This probably would prevent merchants from offering discounts for high-quality notes.

[12] Supel and Todd (1984) first suggested this analogy.

6. WOULD THE QUALITY OF CURRENCY IN CIRCULATION DECLINE?

The preceding discussion raises some pertinent questions. How would a deposit-fee scheme affect the average quality of currency in circulation? By discouraging Fed processing, would it reduce the rate of destruction of unfit notes, increasing the average life of a note? More generally, what is the optimal quality of currency in circulation, and how can we attain the optimum?

As a benchmark, suppose that there is no "market failure" for currency quality, in the sense that the willingness of private individuals to pay for a higher-quality note is identical to its total social benefits. We could then let market forces determine the socially optimal level of currency quality, just as they determine the quality of any privately produced durable good. It would be crucial for the Fed to set the relative price of sorted and unsorted currency to reflect the real costs of transforming unsorted straps into sorted straps. In this case, private decisions would result in the socially optimal currency quality.

Figure 1, based on a model of currency processing described in Appendix A, illustrates the interaction between currency quality and the price of processing. The quality of currency in circulation is measured by the variable q, the fraction of notes in circulation that are fit rather than unfit.[13] The vertical axis displays the marginal benefits and marginal costs associated with each level of currency quality. The curve labeled MSC is the marginal social cost of Federal Reserve currency processing and includes the cost of processing notes plus the cost of printing new notes to replace the unfit notes destroyed. The curve labeled MSB is the marginal social benefit of higher-quality currency. The socially optimal currency quality is q^*, where MSC equals MSB. This currency quality maximizes social benefits minus social costs, or net social welfare.

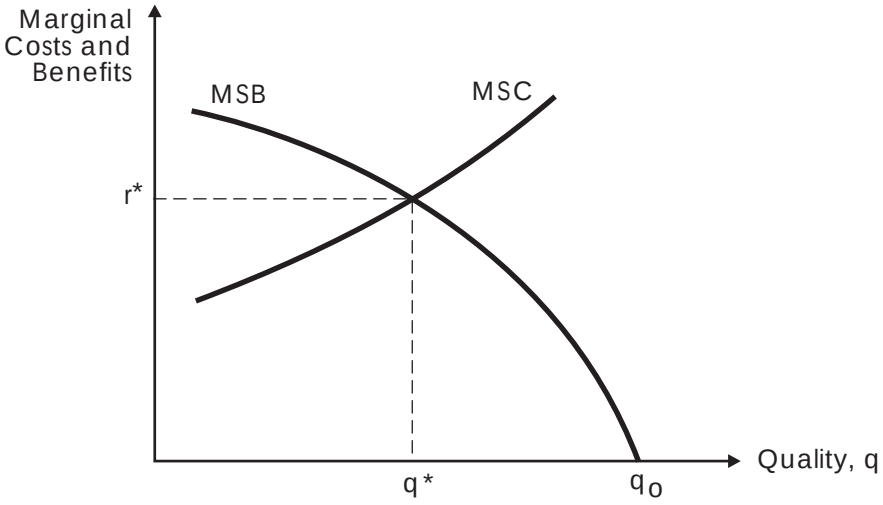
Underlying Figure 1 is the demand and supply for Fed processing. Banks choose how much currency to send to the Fed and to withdraw from the Fed based on the price of Fed processing—the difference between what the Fed charges for withdrawals of fit currency and what the Fed pays for unfit currency. A bank's demand for Fed-processed currency determines the quality of currency circulated by the bank. Fed processing volume determines the fraction of circulating unfit notes that are destroyed each period and replaced with new notes, and thus directly determines the quality of currency in circulation. A price for Fed processing thus is equivalent to a "price" for currency quality, since quality varies one-for-one with Fed processing volume.[14]

In the absence of market failure, private willingness to pay for currency quality corresponds exactly to the marginal social benefits of currency quality.

[13] For simplicity, think of currency as either fit or unfit, and suppose that over time fit notes randomly become unfit.

[14] More precisely, the "price" of currency quality is the price of Fed processing times the amount by which Fed processing volume must rise to increase quality by one unit.

Figure 1 Optimal Currency Quality, "No Market Failure" Case



In this case the optimal price of currency quality, r^*, is the marginal social cost of quality and it directly determines the optimal price of Federal Reserve currency processing. Facing r^*, banks bear the full social cost of higher currency quality, and so choose the socially optimal quality. Note that r^* includes the cost of printing new currency to replace the unfit currency in a deposit.

Would replacing the current system of quantitative restrictions with a deposit fee result in a decline in the quality of currency in circulation? Figure 1 shows that, by itself, imposing an optimal deposit fee reduces quality from q_0 to q^*. But the level of quality under free provision with quantitative rationing could be much less than q_0, or q^* for that matter. Therefore, what happens to currency quality depends upon how the level of quality under the current policy compares with the optimal level of quality. Is currency quality now too high or too low? One recent study of this question suggests that, if anything, currency quality is currently above the social optimum.[15]

[15] High-speed currency verification and sorting machines are capable of distinguishing between 16 different currency soil levels and are set to destroy all notes at or below a given soil level. The Federal Reserve System recently decided to reduce the threshold for destruction by one setting, from 9 to 8 for one-dollar bills and from 10 to 9 for other denominations. This decision was made after a study that documented the likely public reaction to the predicted change in currency quality and the effect on new currency printing cost. Extensive surveys revealed that lowering the threshold "would have a minimal impact on public acceptability of currency." The decision reflects the implicit judgment that the marginal social cost of the higher soil level setting is larger than the marginal social benefits of higher currency quality. See Federal Reserve System (1991).

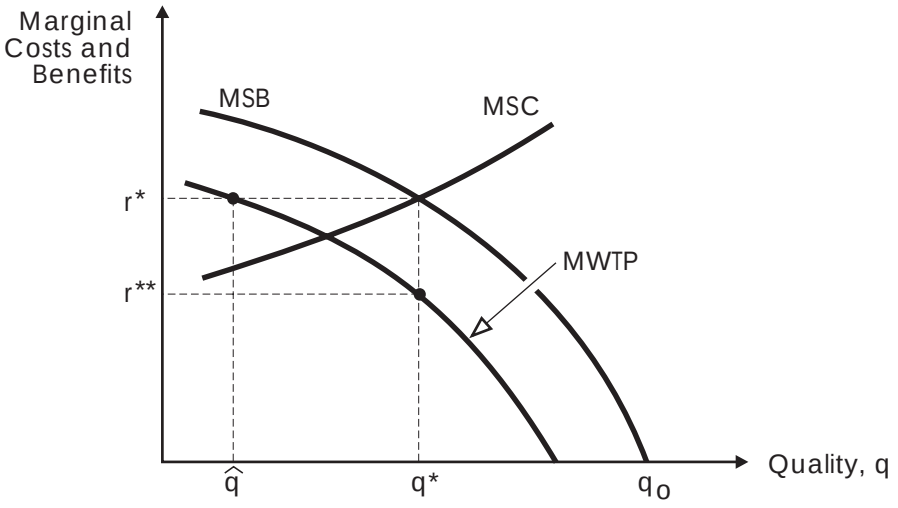
7. WHAT IF THERE IS A MARKET FAILURE FOR CURRENCY QUALITY?

The fact that currency trades at par can lead to a market failure that complicates the attempt to achieve the optimum level of currency quality. Banks choose the quality of currency to supply for withdrawals, but that currency is passed on and deposited at other banks. Because banks are small relative to the currency market, they neglect the effect of their currency outflow on the quality of other banks' currency inflows. Banks ignore the effect of their currency-processing decisions on the quality of currency in circulation, because they are not fully compensated for the social value of the currency quality they supply. At a more fundamental level, the source of the market failure is that the social benefits of quality-dependent note prices are smaller than the costs associated with the inconvenience of such prices. Because prices do not vary with note quality, people do not bear the full social costs and benefits of their decisions.

There are some settings in which suppliers of high-quality currency can obtain direct or indirect compensation for some of the social benefits associated with high quality. Some institutional currency users actually pay to obtain high-quality currency. For example, some banks pay to obtain higher-quality currency for loading into automated teller machines. Many banks, hotels and retail merchants prefer to use high-quality or even new currency in transactions with their customers. Since obtaining high-quality currency often involves some effort or expense, we can presume that the institution believes that its customers prefer high-quality currency, and that providing it to them will result in implicit remuneration such as enhanced goodwill or a higher probability of repeat business. Thus the institution may be partially compensated by its customers for the quality of currency disbursed, making them willing to pay a small premium to arrange for better-quality currency. The demand by these institutions for better-quality currency will reflect a portion of the social benefits of better quality, even though currency trades at par, without adjustment for quality.

For most transactions, however, such mechanisms of implicit compensation are generally lacking or incomplete. Therefore, the demand for fit sorted currency from the Fed, relative to the demand for unsorted currency, is likely to understate society's true demand for fit currency. If so, setting the price of currency processing equal to the Fed's full marginal cost will lead to suboptimal currency quality. Figure 2 illustrates the dilemma. It is based on a model (also described in Appendix A) in which private willingness to pay, the curve labeled MWTP, is less than marginal social benefits, the curve again labeled MSB. Charging the full marginal social cost, r^*, results in suboptimal currency quality, $\hat{q}$, because the demand for Fed processing does not reflect true social benefits. The optimal price is the one that attains the optimal level of Fed currency processing and is equal to the marginal willingness to pay at the optimal level of processing, r^{**} in Figure 2. This price is less than full marginal cost and

Figure 2 Optimal Currency Quality, "Market Failure" Case



subsidizes processing in order to counteract the divergence between the private and social marginal benefits of currency quality. The divergence is the direct result of the fact that currency generally trades at par, unadjusted for quality.

It is quite unlikely that the optimal price is zero. For this to occur, marginal private willingness to pay must be zero at the optimal quality. One component of private willingness to pay is the costliness of bank currency processing; banks are willing to pay to substitute Fed processing for their own. As long as bank processing is at all costly there will always be a positive willingness to pay for Fed processing.

8. AN ALTERNATIVE PRICING SCHEME

I have proposed a simple fee on all currency deposits, but other pricing schemes are also worth considering. One attractive idea is a two-tiered scheme, with a higher fee on deposits of fit currency and a lower fee on deposits of unfit currency.[16] This would provide an enhanced incentive for banks to sort currency themselves and deposit only the unfit, keeping the fit in circulation. If banks' currency-processing decisions are multidimensional, then a two-tiered scheme can bring us closer to the social optimum. A two-tiered, quality-dependent

[16] Cash Services Strategic Planning Task Force (1991).

pricing scheme thus could improve upon a simple deposit fee by inducing banks to make more efficient currency-processing choices.

Unfortunately, such a scheme would be quite costly to administer using existing technology. Currently, notes are fed through processing machines in continuous batches made up of many different deposits, and the number of unfit notes is recorded for the entire batch, but not for each bank's deposit. Measuring the fitness of a single deposit requires processing that deposit in isolation as a single batch, at a significantly larger cost. Measuring deposit fitness would be essential to a two-tiered pricing scheme, since banks would have an incentive to misrepresent their deposits as unfit. An additional social cost associated with a two-tiered scheme is that many depository institutions would be induced to acquire the costly devices the Fed uses to measure quality.

If the social benefits of more efficient currency processing outweigh the required administration and enforcement costs, a two-tiered pricing scheme would represent an improvement over a simple deposit fee. A two-tiered scheme probably is not significantly better than a simple deposit fee under current arrangements, however. The next generation of currency-processing equipment, to be deployed in the next five years, will have the capability of measuring individual deposit fitness at little additional cost. A two-tiered scheme could be worthwhile at that time.

9. PRECEDENTS

The position advocated here—paying less than par for worn money—may seem radical, but it is far from new. Historically, mints often bought foreign and worn domestic coin at the market price of the metal and then issued overvalued coins, with most of the difference accounted for by "brassage," the cost of minting.[17] When the Bank of England bought worn gold coins for reminting, "there was a small series of charges or profits made for weighing, melting, assaying, the turn of the scale, the difference of the assay reports, etc.," plus a charge for "demurrage," the interest lost while the gold was being coined.[18] Thus the deterioration of coin has been a perennial problem for monetary authorities, and has evoked changing policies as the technology underlying monetary arrangements has changed.[19]

There is a more immediate precedent, however. From 1863 to 1935 in the United States, national banks could issue their own notes, which circulated as

[17] See de Roover (1948) p. 241.

[18] Jevons (1875). The 19th century English economist was quite concerned about the quality of English currency and actually performed his own physical experiments to estimate the rate at which English gold sovereigns declined in weight. He also gathered information on the fraction of coins that were below "legal limit," meaning that they no longer qualified as legal tender and could only be sold by weight (at a loss).

[19] See, for example, Redish (1990).

currency. These notes were printed by the U.S. Treasury and shipped to the bank for issue. At first, notes could be redeemed at the issuing bank or at a "redemption agent" designated by the issuing bank, generally in New York. This made redemption costly, since by law the notes had to be redeemed at par. As a result, the condition of national bank notes in circulation deteriorated. In 1874, Congress authorized a "Redemption Agency" within the Treasury that would redeem national bank notes at various locations around the country, destroying worn and mutiliated notes and sending back to the issuing bank the fit notes and, if necessary, newly printed replacement notes.

National banks were charged for all of the expenses associated with note issue and redemption. The banks were charged directly for the cost of the plates and dies used to print their notes. They were also assessed a 1 percent tax on outstanding notes (1/2 percent after 1900), out of which the other costs of producing new notes were paid. The costs of redemption were divided pro rata among banks on the basis of the amount of their notes redeemed. In addition, banks paid all of the costs of shipping their own notes.

By all accounts national bank notes functioned as hand-to-hand currency, and since they were collateralized by government securities, they were, in effect, government currency. The resource costs associated with national bank notes, arguably the model for Federal Reserve notes, were borne by private entities. The policy of charging national banks for the costs of note issue and redemption is thus directly analogous to my proposal to charge a fee for currency deposits.[20]

10. CONCLUDING REMARKS

A central bank fee for deposits of worn currency is an uncomfortable notion for many, but the proposal merely reflects standard economic reasoning. When the government provides a service for which there are privately provided substitutes, economic theory tells us that it should *not*, in general, be provided free. The current system of fees for wire transfers of reserve balances reflects just such reasoning and was designed to improve economic efficiency by providing the proper incentives to private decision makers.[21] The goal of a currency-processing fee is identical—to improve the efficiency of currency handling by providing private decision makers with the proper incentives.

The fact that currency processing plays a role in our monetary system can distract from the essential economics of the issue. Some express the view that

[20] There are some minor differences. The cost of printing notes was charged to the issuing national bank at the date of issue; under my proposal, the cost of printing notes would be charged to depositors of used notes. Under the national bank policy, the deposit of notes with the Treasury resulted in a charge to the issuing bank, not the depositing bank; under my proposal, the depositing bank would be charged.

[21] Board of Governors (1990).

par replacement of worn currency is somehow intrinsic to central banking or the monetary mechanism. But monetary policy concerns the size of the total monetary base, not the division of the base into its components; a deposit fee would merely add to the cost of exchanging currency for reserve account balances. Some express the view that par replacement is somehow required "to furnish an elastic currency," but again, this can be accomplished through variations in the size of the total monetary base; as long as the composition of the base is determined by demand, the currency supply will be sufficiently elastic.[22] Similar arguments were raised in objection to pricing other Federal Reserve services in the early 1980s, but are no longer taken seriously. Indeed, payments system efficiency is now a widely recognized public policy goal. A currency deposit fee would help.

APPENDIX A: A Model of Currency Processing

This appendix describes a simple model designed to capture the interaction between the price of Fed processing and the quality of currency in circulation. The heart of the model is the decision problem of a representative "bank" facing currency deposit and withdrawal flows. The bank chooses the quality of currency to supply, how much currency to process internally and how much currency to deposit and withdraw from the Fed. The Fed processes deposited currency, destroying unfit notes and replacing them with newly printed notes. The Fed sets fees for deposits and withdrawals, and these directly affect banks' processing decisions. Alternative fee policies affect currency quality, and I show how policy can achieve the socially optimal currency quality.

I consider two versions of the model. In the first version, the "no market failure case," currency prices depend on quality and affect consumers' demand for currency quality. In this case the optimal price of Fed currency processing is just the Fed's marginal cost. In the second version, the "market failure case," currency prices do not depend on quality, and the optimal price of Fed currency processing is less than marginal cost, but strictly greater than zero.

The model is an extension of the general equilibrium monetary model of Lucas (1980). It is essentially a model of bank processing embedded within a simple cash-in-advance framework. Articulating such a framework is useful because it allows us to evaluate Fed currency policy using the standard tools of welfare economics. In addition, such a framework forces us to be explicit about the benefits and costs of currency quality and ensures that the "demand," "supply," and "price" of currency quality are modeled in an internally consistent way.

[22] See Goodfriend and King (1988).

Many interesting aspects of currency policy are omitted here for the sake of clarity. Ignored are transportation costs, for example; they would modify the analysis in obvious ways. Also ignored is the fact that currency comes in distinct denominations that the public treats very differently; the model assumes a single denomination. Since the model is designed just to study the pricing of currency processing, the technology of public and private currency processing is taken as given.

Currency Processing

Currency, in this model, is either "fit" or "unfit." Currency is fit when first printed. While in circulation, a random fraction δ of fit notes deteriorate each period and become unfit. Thus, δ is both the probability that a given fit note becomes unfit in circulation in a given period and the fraction of fit notes that become unfit during the period.[23] The quality of the set of notes that circulate during period t will be denoted q_t, the fraction of the notes that are fit. Thus if notes of quality q_t begin circulating, their average quality at the end of the current period is $(1-\delta)q_t$.

There are a large number of identical unit banks, and at the beginning of each period they each receive a deposit of M units of currency from the public. At the end of the period they each must meet a demand for withdrawal of M units of currency. A representative bank can deposit notes directly at the Fed without processing them, or it can process the notes themselves, sending unsatisfactory notes to the Fed. The bank then withdraws currency from the Fed in order to meet their own withdrawal demand of M notes. For every note deposited at the Fed they need to withdraw a note later from the Fed.[24]

Fed processing sorts all deposited notes into fit and unfit notes and destroys all unfit notes. The cost of Fed processing is $f(x_t^F)$ units of labor time, where x_t^F is the number of notes processed by the Fed at time t. The destroyed unfit notes are replaced with newly printed notes. The cost of printing x_t^N new notes at time t is $g(x_t^N)$ units of labor time.

Bank processing is less accurate than Fed processing. When a bank sorts notes, the satisfactory notes, call them "clean," are mostly fit but mistakenly include some unfit notes. The unsatisfactory notes, call them "dirty" notes, are mostly unfit but mistakenly include some fit notes. The dirty notes are all sent

[23] I am implicitly assuming that there are an infinite number of notes. Variables denoting quantities of notes should thus be interpreted as fractions of the total currency outstanding.

[24] It is possible to modify the model to allow for heterogenous banks, some of which experience net currency inflows each period and some of which experience net currency outflows. Private processing would then involve some shipment of currency from inflow banks to outflow banks. Such a model could be used to assess cross-shipping restrictions; banks with net deposit inflows could deposit at the Fed rather than ship to other banks, and banks with net currency outflows could withdraw from the Fed rather than buy from other banks, but banks would be unable to both deposit and withdraw from the Fed. If transportation costs are positive, cross-shipping has different effects on differently situated banks.

to the Fed. Let η be the fraction of unfit notes that the bank classifies as clean, and let ζ be the fraction of fit notes that the bank classifies as dirty. Thus η and ζ represent "error rates" in bank processing.[25] Table A–1 displays the amount of notes in each possible category, assuming that the bank processes x_t^B notes itself at time t, and the quality of notes deposited at the bank is $\hat{q}_t = (1-\delta)q_{t-1}$. The cost of bank processing is $h(x_t^B)$, where x_t^B is the number of notes processed.

The quality of the currency disbursed by the bank is determined by the quality of the currency it receives and by the amount of currency it processes itself. All of the notes withdrawn from the Fed are fit. The only unfit notes the bank supplies to customers are those mistakenly classified as clean. Therefore the quality of notes supplied by the bank, q_t, is given by

$$q_t = 1 - \eta(1 - \hat{q}_t)x_t^B/M. \tag{1}$$

This expression relates the quality of notes supplied this period to the amount of notes processed by banks and $\hat{q}_t$, the quality of notes deposited at banks. The more processing done by banks, the more unfit clean notes slip through without being sent to the Fed for destruction. Thus q_t is decreasing in x_t^B.

The amount of notes the bank deposits at the Fed is

$$\begin{aligned} x_t^F &= M - x_t^B + (1-\eta)(1-\hat{q}_t)x_t^B + \zeta\hat{q}_t x_t^B \\ &= M - [\eta + (1-\eta-\zeta)\hat{q}_t]x_t^B. \end{aligned} \tag{2}$$

The deposit consists of the notes the bank does not process, plus the notes that the bank processes but finds dirty. Note that the greater the volume of private processing, the smaller the volume of Fed processing.

Table A–1 Bank Note Processing During Period *t*

Note flow	Fit	Unfit
Deposit at the bank	$\hat{q}_t M$	$(1-\hat{q}_t)M$
Sent to the Fed	$\hat{q}_t(M - x_t^B)$	$(1-\hat{q}_t)(M - x_t^B)$
Processed by the bank	$\hat{q}_t x_t^B$	$(1-\hat{q}_t)x_t^B$
Clean	$(1-\zeta)\hat{q}_t x_t^B$	$\eta(1-\hat{q}_t)x_t^B$
Dirty, sent to the Fed	$\zeta\hat{q}_t x_t^B$	$(1-\eta)(1-\hat{q}_t)x_t^B$

M is the number of notes deposited at time t, $\hat{q}_t$ is the average quality of notes deposited at time t, x_t^B is the number of notes processed by the bank at time t.

[25] In the special case in which $\eta = \zeta = 0$, private processing is just as good as Fed processing at picking out unfit notes, and currency quality does not depend on the division of processing between banks and the Fed. In this case pricing has no effect on the quality of currency in circulation, although it does influence economic efficiency by affecting the division of processing between banks and the Fed. The exposition assumes that $\eta > 0$.

The number of new notes printed is equal to the number of unfit notes deposited at the Fed and is given by

$$\begin{aligned} x_t^N &= (1-\hat{q}_t)(M - x_t^B) + (1-\eta)(1-\hat{q}_t)x_t^B \\ &= (1-\hat{q}_t)(M - \eta x_t^B). \end{aligned} \tag{3}$$

The greater the volume of private processing, the larger the number of unfit notes escaping destruction each period, and so the smaller the number of new notes printed.

The General Equilibrium Environment

Currency deposit and withdrawal flows to and from the banking system are taken as given in the description of currency processing. These flows are determined within a general equilibrium model based on Lucas (1980). There are a large number of spatially separated locations. At each location there are representative banks like the one described above, along with a Federal Reserve Bank branch. In addition, there are a large number of households at each location. Each household consists of a pair of agents, a "shopper" that visits other locations and obtains goods, and a "worker" that stays at home to produce goods and sell them to visiting shoppers. (Households do not like to consume the goods produced at their own location.) The cost of reliably verifying a shopper's identity at another location is prohibitive, and this precludes the use of credit arrangements. Thus, shoppers carry currency to exchange for goods. At the same time the worker stays at home to sell goods to shoppers from other locations. The currency received by the worker cannot be used by the shopper during the same period; the shopper must use currency obtained earlier.

The worker produces a consumption good using a constant returns to scale technology in which one unit of labor time produces one unit of consumption good. (There is no capital.) The worker is endowed with L units of labor time, derives no utility from leisure, and so supplies labor inelastically. In addition to producing a consumption good for sale, the worker is employed outside of the house for a market wage of $w_t = 1$. Banks and Fed branches hire workers to process currency.

During the first part of each period, households deposit the currency accumulated the previous period at their local bank for safekeeping. Employment and household production then take place, and wages are credited to the household's bank account. At the end of the first part of the period, households withdraw currency for the shopper to use. In the second part of the period the shopper buys goods for currency, and the worker receives currency for goods. The shopper then returns home and consumption takes place. Currency is stored until the next period, and then deposited at the bank. All outstanding currency is deposited at banks at the beginning of each period; one can imagine that

thieves are at large during the first part of each period, making privately held currency risky.

Handling currency is time consuming for the worker, and the quality of currency affects the amount of time required for currency handling. I assume that handling fit notes takes no time at all but that handling unfit notes takes time. Specifically, handling $(1-q_t)M$ unfit notes requires $\phi[(1-q_t)M]$ units of labor time, where $\phi(\,)$ is an increasing function. This time cannot be used to produce goods or work outside the house. Currency quality has no other effect on shoppers or workers. Therefore, we can define the value of currency quality as $v(q_t) = L-\phi[(1-q_t)M]$, the labor time available for productive activities, net of the time spent handling currency.[26] Currency deteriorates after it has been handled by the worker, but before it has been deposited at the bank the next period.

The number of notes in circulation is M and is held constant throughout the analysis. The Federal Reserve Banks perform the processing described earlier, perhaps collecting fees. Any deficit is covered by levying lump-sum taxes on households.

Optimal Currency Quality

The optimal currency quality maximizes the lifetime utility of a representative household subject to the resource constraints. Lifetime household utility is $\sum_{t=0}^{\infty}\beta^t u(c_t)$, where $u(c_t)$ is the utility of total consumption c_t at time t, and $0<\beta<1$. The economy is constrained each period by the L units of labor time per household available to be divided between bank and Fed processing, new note printing, handling unfit currency, and the production of consumption goods. For convenience, define the function $x^B(\hat{q},q)$ as the number of notes processed by the representative bank when the quality of deposited currency is $\hat{q}$ and the quality of withdrawn currency is q. Define $x^F(\hat{q},q)$ and $x^N(\hat{q},q)$ similarly for Fed processing volume and the number of new notes printed. Using (1), (2), and (3), we have

$$\begin{aligned} x^B(\hat{q},q) &= \frac{(1-q)M}{(1-\hat{q})\eta} \\ x^F(\hat{q},q) &= M-[\eta+(1-\eta-\zeta)\hat{q}]\frac{(1-q)M}{(1-\hat{q})\eta} \\ x^N(\hat{q},q) &= (q-\hat{q})M. \end{aligned} \tag{4}$$

[26] The dependence of $v(q_t) = L-\phi[(1-q_t)M]$ on M is suppressed here because M, the total number of notes, is held constant throughout the analysis. Note that a nonneutrality is built into the environment here by having available labor time depend on the number of unfit notes; scaling up the number of notes at each date reduces available resources, independent of the price level. The optimal policy would seem to be making the number of notes as small as possible. This obviously runs up against the (unmodeled) constraint that the real value of a note be no greater than the smallest transaction. Since the money stock is held constant, this nonneutrality is inconsequential.

The following "social planner's problem" finds the optimal currency quality. Choose sequences $\{\hat{q}_t\}_{t=0}^{\infty}$, and $\{q_t\}_{t=1}^{\infty}$ to

$$\max \sum_{t=0}^{\infty} \beta^t u(c_t)$$

such that

$$\begin{aligned} c_t &\leq L - \phi[(1-q_t)M] - f[x^F(\hat{q}_t, q_t)] \\ &\quad - g[x^N(\hat{q}_t, q_t)] - h[x^B(\hat{q}_t, q_t)] \\ \hat{q}_{t+1} &= (1-\delta)q_t \end{aligned} \tag{5}$$

This programming problem maximizes the representative household's discounted lifetime utility, subject to the constraint that consumption can be obtained with the labor not devoted to currency handling, currency processing or currency printing. The first-order condition for this optimization problem is

$$\begin{aligned} v'(q_t) - h'(x_t^B)x_2^B(\hat{q}_t, q_t) - \beta(1-\delta)h'(x_{t+1}^B)x_1^B(\hat{q}_{t+1}, q_{t+1}) \\ = f'(x_t^F)x_2^F(\hat{q}_t, q_t) + \beta(1-\delta)f'(x_{t+1}^F)x_1^F(\hat{q}_{t+1}, q_{t+1}) \\ + g'(x_t^N)x_2^N(\hat{q}_t, q_t) + \beta(1-\delta)g'(x_{t+1}^N)x_1^N(\hat{q}_{t+1}, q_{t+1}), \end{aligned} \tag{6}$$

where, for example, $x_1^B(\hat{q}_t, q_t) \equiv \partial x^B(\hat{q}_t, q_t)/\partial \hat{q}_t$, $x_2^B(\hat{q}_t, q_t) \equiv \partial x^B(\hat{q}_t, q_t)/\partial q_t$, and so on. The first term on the left side is the marginal social benefit of currency quality, the savings in productive resources from higher average currency quality. The second and third terms are the marginal reduction in private processing costs associated with higher currency quality—together these terms are generally positive because higher quality requires more Fed processing and thus less private processing. The right side is the marginal social cost of currency quality—marginal Fed processing costs plus marginal printing costs. The optimality condition, as usual, is that the total marginal social benefit should be equated to total marginal social cost. Note that an increase in q_t, quality at date t, increases total costs at date t, but reduces total costs at date $t+1$. The better the quality of currency at date t, the better the quality of the currency deposited at the beginning of period $t+1$.

I restrict attention to steady state equilibria. In a steady state equilibrium, all variables are constant over time. The condition for optimality in a steady state equilibrium can be written

$$\begin{aligned} v'(q^*) - h'(x^{B*})[x_2^B(\hat{q}^*, q^*) + \beta(1-\delta)x_1^B(\hat{q}^*, q^*)] \\ = f'(x^{F*})[x_2^F(\hat{q}^*, q^*) + \beta(1-\delta)x_1^F(\hat{q}^*, q^*)] \\ + g'(x^{N*})[x_2^N(\hat{q}^*, q^*) + \beta(1-\delta)x_1^N(\hat{q}^*, q^*)], \end{aligned} \tag{7}$$

where $x^{B*} = x^B(\hat{q}^*, q^*)$, $x^{F*} = x^F(\hat{q}^*, q^*)$, $x^{N*} = x^N(\hat{q}^*, q^*)$, and $\hat{q}^* = (1-\delta)q^*$. q^* is the optimal steady state currency quality.

The determination of the optimal currency quality is illustrated in Figure 1. The left side of (7) is plotted as a function of steady state quality q and is

labeled MSB. The right side of (7) is labeled MSC. The intersection determines the optimal currency quality q^*.[27]

Optimal Price, No Market Failure Case

As a reference point, I first derive the optimal currency policy when there is no "market failure," in the sense that the willingness of a bank to pay for a higher-quality note is identical to the total social benefits of a higher-quality note. I assume that consumers are willing to pay a bank for the value to them of the quality notes they withdraw. In order for this to hold, consumers must believe that they will be paid for the quality of notes that they pass on to others, including banks. Therefore, I need to assume that workers pay shoppers for the quality of the currency received and that banks pay depositors for the quality of notes deposited.

Let $\hat{\rho}_t$ be the price paid by banks (over and above par) for fit notes in period t, so that deposits of M notes of quality $\hat{q}_t$ at t cost $\hat{\rho}_t \hat{q}_t M$.[28] Let ρ_t be the price paid to banks (over and above par) for quality notes withdrawn at t, so that a bank receives $\rho_t q_t M$ for withdrawals of M notes of quality q_t.

A shopper pays $\rho_t q_t$ per note for a withdrawal of notes of average quality q_t at t, and then exchanges the notes for consumption goods. Since a shopper is unaffected by currency quality, the premium for currency quality paid by the worker must equal the premium paid by the shopper at the bank. Thus the worker pays $\rho_t q_t$ per note for currency received from visiting shoppers. A fraction δ of the fit notes deteriorate and become unfit during period t, so the average quality of the notes is $(1-\delta)q_t$ at the beginning of period $t+1$. The bank will pay $\hat{\rho}_{t+1}(1-\delta)q_t$ for these notes at the beginning of next period. If the worker faces a market real interest rate of $r = \beta^{-1} - 1$, then the present value of selling the currency next period is $\beta\hat{\rho}_{t+1}(1-\delta)q_t$. Thus the "rental price" of notes with average quality q_t is $[\rho_t - \beta(1-\delta)\hat{\rho}_{t+1}]q_t$. This expression is exactly analogous to the user cost of a durable good, as one would expect. Given the rental price, the worker chooses a utility-maximizing note quality which satisifies[29]

$$v'(q_t) = \rho_t - \beta(1-\delta)\hat{\rho}_{t+1}. \tag{8}$$

[27] Certain mild conditions on the functions v, h, f, and g are required for there to be a unique interior optimal currency quality.

[28] Specifically, if banks pay p_t^u in real terms per unfit note and p_t^f per fit note, then p_t^u is par and $\hat{\rho}_t = p_t^f - p_t^u$. One could carry out the analysis in terms of the numbers of fit and unfit notes bought and sold by various parties, rather than in terms of the number of total notes and the quality of those notes. The latter parameterization allows us to separate quality choice from the choice of total money holdings.

[29] It is straightforward to derive this condition from the household's maximization problem, omitted here for brevity.

In other words, consumers equate the rental price of currency quality to the marginal real resource savings from currency quality.

I only consider currency policies that involve deposit and/or withdrawal charges—there is no quantitative rationing. The Fed pays p^D for deposits at all dates, which is equal to the par value of the currency minus any deposit charges. The Fed charges p^W for withdrawals of currency, par plus any withdrawal charges. Therefore, the bank pays a net amount of $(p^W - p^D)$ to send a note to the Fed for processing and then withdraw a note to replace it.

A typical bank essentially solves a static problem each period, maximizing receipts from currency sales minus outlays for purchased currency, processing costs, and Fed charges. If the bank buys currency of quality $\hat{q}$ and wants to supply currency of quality q (suppressing time subscripts), it must process $x^B(\hat{q}, q)$ notes itself, and send $x^F(\hat{q}, q)$ notes to the Fed. (These functions were defined above in [4].) The maximization problem for a typical bank during a typical period, therefore, is

$$\max_{q,\hat{q}} \rho q M - \hat{\rho}\hat{q}M - h[x^B(\hat{q}, q)] - (p^W - p^D)x^F(\hat{q}, q). \tag{9}$$

The bank maximizes the value received from customers for supplying currency of quality q, minus the amount paid for currency deposits, minus the bank's processing costs, and minus the net cost of depositing and withdrawing currency from the Fed. The bank's decision problem reflects the social value to consumers of higher-quality currency through ρ, the market value of supplying higher-quality currency. The bank bears its own processing costs, but does not directly bear the cost of Fed processing and new note printing. The first-order condition for this problem is

$$\begin{aligned} \rho - \beta(1-\delta)\hat{\rho} - h'(x^B)[x_2^B(\hat{q}, q) + \beta(1-\delta)x_1^B(\hat{q}, q)] \\ = (p^W - p^D)[x_2^F(\hat{q}, q) + \beta(1-\delta)x_1^F(\hat{q}, q)]. \end{aligned} \tag{10}$$

The bank equates the market value of an improvement in currency quality $[\rho - \beta(1-\delta)\hat{\rho}]$, plus the bank's marginal processing cost savings, to the marginal cost of Fed processing. Given $p^W - p^D$, the first-order condition (10) can be combined with (8) to determine the steady state equilibrium value of q consistent with bank profit maximization and consumer utility maximization.

$$\begin{aligned} v'(q) - h'(x^B)[x_2^B(\hat{q}, q) + \beta(1-\delta)x_1^B(\hat{q}, q)] \\ = (p^W - p^D)[x_2^F(\hat{q}, q) + \beta(1-\delta)x_1^F(\hat{q}, q)], \end{aligned} \tag{11}$$

where $\hat{q} = (1-\delta)q$. This shows how the price of Fed currency processing, $p^W - p^D$, determines the equilibrium level of currency quality. Under standard assumptions, the left side of (11) is decreasing in q, so that an increase in $p^W - p^D$ causes equilibrium currency quality to decrease.

The optimal price of Fed processing, call it $\hat{r}^*$, is the value of $p^W - p^D$ that induces an equilibrium level of quality equal to q^*, the optimal quality. If

$q = q^*$, then the left side of (11) is equal to the left side of (7). Therefore, $\hat{r}^*$ must equate the right side of (11) with the right side of (7). This implies that,

$$\hat{r}^* \equiv f'(x^{F*}) + g'(x^{N*})\frac{[x_2^N(\hat{q}^*, q^*) + \beta(1-\delta)x_1^N(\hat{q}^*, q^*)]}{[x_2^F(\hat{q}^*, q^*) + \beta(1-\delta)x_1^F(\hat{q}^*, q^*)]}, \tag{12}$$

where $\hat{q}^* = (1-\delta)q^*$. Equation (12) states that in the absence of market failure the optimal price of Fed processing is the marginal cost of Fed processing plus the marginal cost of printing new notes. The latter is the marginal printing cost per note times the marginal number of notes printed per note processed by the Fed.

The determination of the optimal price of processing is illustrated in Figure 1. The curve labeled MSB again corresponds to the left sides of (7) and (11). For a given price of processing, banks and consumers act so as to equate MSB to $(p^W - p^D)\left[x_2^F(\hat{q}, q) + \beta(1-\delta)x_1^F(\hat{q}, q)\right]$, the right side of (11). The optimal "price" of currency quality, r^*, is just $\hat{r}^*$, the optimal price of Fed processing, multiplied by $x_2^F(\hat{q}^*, q^*) + \beta(1-\delta)x_1^F(\hat{q}^*, q^*)$, the derivative of Fed processing volume with respect to steady state quality. If $p^W - p^D = \hat{r}^*$, then the equilibrium level of q is q^*, the optimal currency quality. Since there is no market failure in this case, marginal private willingness to pay is equal to marginal social benefits, and the optimal price is just marginal social cost.

Optimal Price, Market Failure Case

I now consider the case in which banks do not completely internalize the social benefits of currency quality. As discussed in the text, the market failure occurs because currency generally trades at par, with no adjustment for quality. Thus, banks and consumers do not obtain full compensation for the quality of the currency they pass on. I implement this case by assuming that $\rho = 0$, so that banks receive *no* compensation for the quality of the currency they supply. I show that even in this extreme case the optimal price is still strictly positive.

Banks solve the same profit-maximization problem as before, but with $\rho = 0$. The resulting first-order condition is just (10) with $\rho = 0$. As before, this condition can be combined with (7), the condition for the optimality of currency quality, to obtain an expression for the optimal price of Fed processing.

$$\begin{aligned}\hat{r}^{**} \equiv\ & f'(x^{F*}) + g'(x^{N*})\frac{[x_2^N(\hat{q}^*, q^*) + \beta(1-\delta)x_1^N(\hat{q}^*, q^*)]}{[x_2^F(\hat{q}^*, q^*) + \beta(1-\delta)x_1^F(\hat{q}^*, q^*)]} \\ & - \frac{v'(q^*)}{[x_2^F(\hat{q}^*, q^*) + \beta(1-\delta)x_1^F(\hat{Q}^*, q^*)]} \\ <\ & \hat{r}^*\end{aligned} \tag{13}$$

Because of the third term, the optimal price is now smaller than $\hat{r}^*$, the optimal price in the no market failure case. Banks are not compensated for the direct

value of currency quality to consumers, so the optimal price of Fed processing includes a subsidy proportional to $v'(q)$, marginal value of currency quality to consumers.

The determination of the optimal price of processing in this case is illustrated in Figure 2. Again, r^{**}, the optimal "price" of currency quality, is just $\hat{r}^{**}$ multiplied by $x_2^F(\hat{q}^*, q^*) + \beta(1-\delta)x_1^F(\hat{q}^*, q^*)$. Marginal private willingness to pay (MWTP) is the left side of (10) but with $v'(q)$ omitted; it now lies below the marginal social benefit curve because it does not include the value of currency quality to consumers. Optimality requires that the price of processing just equal MWTP at the optimal quality, q^*. Clearly, this requires that the price be below the price that is appropriate in the no market failure case, so we have $\hat{r}^{**} < \hat{r}^*$. In the market failure case the optimal price partially subsidizes processing. Without such a subsidy, banks will not send enough currency to the Fed for processing, and the quality of currency they supply will be below the optimum.

Note that the optimal price of Fed processing is still strictly positive in this case. Banks still weigh their own costs against Fed processing charges in deciding how much currency to process themselves. Fed pricing still must ensure that the division of processing between the Fed and the private sector is efficient.

APPENDIX B: DATA SOURCES

Table 1. Operating expenses for currency and coin are from Board of Governors of the Federal Reserve System, *1991 PACS Expense Report* (Board of Governors 1992): for currency, "total cost, currency service," p. 215; for coin, "total cost, coin service," p. 236. Operating expenses for reserve account balances are from the Pro Forma Income Statement for Federal Reserve Priced Services, by Service, 1991, in Board of Governors of the Federal Reserve System, *78th Annual Report*, 1991 (Board of Governors 1992), p. 235; operating expenses for funds transfer and net settlement ($69.3 million) plus commercial ACH ($49.5 million). Replacement costs for currency are calculated by multiplying 6,870,910,000, the number of new notes paid into circulation (*1991 PACS Expense Report*, p. 206), by $ .03686, the average cost of new notes delivered to the Federal Reserve in 1991. The latter was calculated by dividing the cost of currency printing in 1991, $261,316,379 (*78th Annual Report*, 1991, Statements of Revenues and Expenses and Fund Balance, "expenses for currency printing, issuance, retirement, and shipping," p. 239), by the number of notes delivered to the Federal Reserve Banks in 1991, 7,088,967,000 (Board of Governors database, provided by Ted Kelly, Federal Reserve Bank of Richmond).

Table 2. Board of Governors of the Federal Reserve System, *1991 PACS Expense Report* (Board of Governors, 1992): *processing costs,* pp. 199–218; *number of notes received,* p. 205. For *Printing costs,* see above.

REFERENCES

Barron, John M., Michael E. Staten, and John Umbeck. "Discounts for Cash in Retail Gasoline Marketing," *Contemporary Policy Issues,* vol. 10 (October 1992), pp. 89–102.

Board of Governors of the Federal Reserve System. *78th Annual Report,* 1991. Washington: Board of Governors, 1992.

_____. "The Federal Reserve in the Payments System," *Federal Reserve Bulletin,* vol. 76 (May 1990), pp. 293–98.

_____. "Currency Guidelines," *Federal Reserve Bulletin,* vol. 70 (February 1984), pp. 113–14.

_____. Press Release, December 31, 1980.

Booth, George. *Currency and Coin Responsibilities of the Federal Reserve: A Historical Perspective.* Cleveland: Federal Reserve Bank of Cleveland, 1989.

Bös, Dieter. "Public Sector Pricing," in Alan J. Auerbach and Martin Feldstein, eds., *Handbook of Public Economics.* New York: North-Holland, 1985.

Cash Services Strategic Planning Task Force. "A Proposal to Pilot Test the Market-Based Management of Currency Volumes." Federal Reserve Bank of San Francisco, internal memorandum, April 19, 1991.

Curtin, Richard T. "Payment Method Costs: Assessments by Retailers." Survey Research Center, University of Michigan, September 1983.

de Roover, Raymond. *Money, Banking and Credit in Medieval Bruges.* Cambridge, Mass.: The Medieval Academy of America, 1948.

Dotsey, Michael. "Cost Function for Federal Reserve Currency Handling," in Federal Reserve System, *A Comprehensive Assessment of U.S. Currency Quality, Age, & Cost Relationships.* September 1991.

Fama, Eugene F. "Financial Intermediation and Price Level Control," *Journal of Monetary Economics,* vol. 12 (1983), pp. 7–28.

Federal Reserve System. *A Comprehensive Assessment of U.S. Currency Quality, Age, & Cost Relationships.* September 1991.

Goodfriend, Marvin S., and Robert G. King. "Financial Deregulation, Monetary Policy, and Central Banking," Federal Reserve Bank of Richmond *Economic Review,* vol. 74 (May/June 1988), pp. 3–22.

Jevons, W. Stanley. *Money and the Mechanism of Exchange.* Henry S. King & Co.: London, 1875, reprinted New York: Garland, 1983.

Lucas, Robert E., Jr. "Equilibrium in a Pure Currency Economy," in John H. Kareken and Neil Wallace, eds., *Models of Monetary Economies.* Minneapolis: Federal Reserve Bank of Minneapolis, 1980.

Redish, Angela. "The Evolution of the Gold Standard in England," *Journal of Economic History,* vol. 50 (December 1990), pp. 789–805.

Supel, Thomas M., and Richard M. Todd. "Should Currency Be Priced Like Cars?" Federal Reserve Bank of Minneapolis *Quarterly Review,* vol. 8 (Spring 1984), pp. 6–7.