Pricing of Central Bank Payment Services

Edward J. Green*
The Pennsylvania State University

Draft: 2008.11.29

1 Introduction

The economic history of technologies is largely a history of low-fixed-cost, high-marginal-cost technologies being replaced by newly invented, decreasing-average-cost (approximately, high-fixed-cost and low-marginal-cost) technologies that come to be cost minimizing as demand for the product or service produced by the technology increases. The history of payment technology in the past century, and particularly in the last two decades, fits this description to a T. The adoption of decreasing-average-cost technologies in other industries (notably transportation and telecommunications) prompted the development of a body of economics regarding how use of these technologies ought to be priced.[1] The highlights of this theory of *non-linear pricing* are that often, when average cost is decreasing, (1) multi-part pricing must be used to support a total-surplus-maximizing allocation and (2) price discrimination can maximize total surplus, if cost must not exceed revenue and single-part pricing must be adopted.

My impression is that, although most central banks express concern about current or impending challenges in financing the operation of their payment systems, they tend to be more restrained than the operators of proprietary payment systems have been in taking full advantage of the opportunities that nonlinear pricing afford for generating revenue.[2] My goal in this paper is to present—mostly in informal terms—the main ideas of the pricing theory relevant to payment systems, to discuss the relationship of those ideas to some specific issues

*E-mail: eug2@psu.edu. Amy Ringel has provided considerable help. She held a Research Experience for Undergraduates award from Penn State University, funded by Bates White LLC.

[1] Wilson-1993 is a standard reference. The theory applies to multi-product technologies that exhibit economies of scope, as well as of scale, but I abstract from that complication here.

[2] Indirect evidence that central bankers tend to be reticent about using nonlinear pricing "aggressively" comes from several research reports in which central bank economists have collaborated with academic experts, such as BoltBumphrey-2005 and HolthausenRochet-2003, HolthausenRochet-2005 . Although these are careful, perceptive, and complete surveys overall, they fail to mention the possibilities of value-based pricing and "first-degree" price discrimination that I will discuss below.

that central banks operating payment systems face, and to suggest some concrete pricing strategies that the theory suggests deserve serious consideration by central banks.

A payment system (particularly, an electronic, account-based one) consists of a set of deposit accounts in which payments are settled, an arrangement for providing very short-term credit (implemented by crediting a payee's account before debiting the payor's account), a set of ancillary services (such as collateral management) for payors and payees, and a telecommunications system over which payment instructions and related messages are sent. I have suggested elsewhere (in GreenTodd-2001 and Green-2008) that providing the set of deposit accounts and offering short-term credit are the only operational payment-system roles that a central bank arguably has comparative advantage in performing. The cost of performing these narrow roles is a minor part of the overall cost of running contemporary payment systems. Thus, if a central bank were to play such a minimal operational role in payment system, there would probably be no issue about the costs thus entailed being subsidized rather than recovered from user fees. However, to the extent that central banks have decided to provide ancillary services as well, that issue has in fact arisen. Such a decision is efficient when there is a production complementarity between an ancillary service and the "core" service in which the central bank has comparative advantage. The scope of central bank payment services varies widely across countries, but the modal scope is strictly between the extremes. Except for the U.S. Federal Reserve System, the central banks of industrialized economies do not play prominent roles in the "retail" payment systems—those intended specifically for the making of "customer payments" (in which at least one ultimate transactor is a household or non-financial firm) rather than payments between financial intermediaries. Also, except for the Federal Reserve, transactors on those central banks' "wholesale" systems use the proprietary SWIFT telecommunications network. On the other hand, the Bank of Canada is unique in limiting its operational role in the wholesale payment system essentially to the provision of settlement accounts and short-term credit.

While it may seem narrowly technical at first glance, the topic of this paper is actually the crux of the question of whether other central banks ought to adopt the Bank of Canada's model. Central banks seem to face a dilemma. Either to subsidize the fixed-cost component of the payment system or else wholeheartedly to adopt nonlinear pricing seem are the efficient ways to cover the fixed cost. However, central banks are either legislatively prohibited or informally constrained by political considerations from subsidizing the payment system, and they are unwilling to subtract the "political capital" that they would expend in dealing with lobbying from various payment-system users regarding the distributive implications of nonlinear pricing from what they have available to manage the political aspects of monetary policy such that have been described by Sargent-1986 .[3] Unless a central bank possesses both the legal authority and the "political will" to fund a payment system efficiently, then there is a *prima facie* case that it ought to step aside, and to let a proprietary organization that does have the authority and the will to fund it efficiently (via nonlinear pricing, rather

[3]The Monetary Control Act of 1980 prohibits the Federal Reserve (with some qualifications) from subsidizing its payment systems. The political constraint on subsidy probably reflects opposition by politicians and their constituents to the use of seignorage to fund the subsidy, when the opportunity cost is that it is not rebated to the treasury

than by subsidy).

The theory of nonlinear pricing that serves as the basis for this discussion is not the ideal theory for the purpose, especially since it abstracts from network phenomena that actually underly the demand functions that are assumed to be exogenous. This is the theory that continues to be used predominantly by industry regulators and by central bankers who deal with payment-pricing issues in practice, though. Part of the goal of this paper (especially section 3) is to provide an assessment of what the theory does well despite its limitations in principle (just as Ptolemaic astronomy provides good advice for navigating without a GPS device) and of what features and implications need to be treated circumspectly because of those limitations. It is hoped that this aspect of the paper will help to economists studying networks to set a research agenda that will provide practitioners with better foundations—or with corrections—in the specific areas of nonlinear pricing theory that are most problematic.

2 Theory of efficient price discrimination

2.1 Definition and examples of price discrimination

Price discrimination refers to selling goods or services with identical production cost at different prices to different buyers.

Examples of price discrimination include the following.

- Tariffs specific to identified customer classes, unless there are service-cost differences across those classes.
- Negotiated prices that reflect individual purchasers' bargaining power.
- Volume discounts (in cases where the quantity purchased does not affect unit cost).
- Various other forms of market segmentation.
  - Time-of-day premia, seasonal discounts, or other time-related pricing rules that do not reflect costs or constraints of providing time-specific capacity.
  - Product bundling (e.g., historically, bundling of IBM computers and punch cards).

Price differences that reflect underlying cost differences of providing a service (or related services) to different customers do not meet the definition of price discrimination, however. Examples of non-discriminatory price differences include the following.

- Time-of-day pricing that rations limited capacity (e.g., for generating electricity) at distinct levels of *aggregate* (not customer-specific) demand.
- Pricing that accurately incorporates the costs of mitigating different levels of risk, to which distinct purchasers expose the producer.

– Actuarially fair pricing of automobile insurance based on a driver's location, age, etc.

– Conceivably, a payment-service price premium that would be charged to banks with prudential shortcomings.

2.2 Price discrimination, technology, and welfare

The results in this section and the next one concern a market for a single good or service. When there are several commodities, some analogous results hold under appropriate assumptions about economies of scope in producing them. In particular, payments of low and high value will be treated as separate goods in a later section

'Welfare' and 'surplus' (used as synonyms here) refer to consumers' aggregate willingness to pay, net of production cost. When a consumer inelastically demands one unit of the good, willingness to pay is the "reserve price" that characterizes the consumer's demand. When the quantity of the good that a consumer demands is a function of price, willingness to pay is gross surplus: the area under the demand curve (to the left of the quantity demanded). The price determines how this gross surplus is split between the consumer and the producer. Adhering to the partial-equilibrium tradition in the study nonlinear-pricing, 'efficiency' is synonymous in the body of the paper with maximization of aggregate surplus.

Price discrimination *cannot* increase welfare if average cost is increasing in quantity supplied (regardless of demand elasticity), or if demand is inelastic and there is some nondiscriminatory price at which revenue covers the total cost of production (or else the shortfall of revenue from production cost can be covered by a subsidy, funded by nondistortionary taxation). In contrast, price discrimination can enhance welfare if all of the following conditions are satisfied.

- Demand is inelastic;
- Providing a subsidy is infeasible, or would require too large a tax distortion;
- Total cost of production cannot be covered by the revenue obtainable from nondiscriminatory pricing; and
- Resale can be prohibited.

Consider a formal example of a market with lumpy production and inelastic demand. There is an indivisible good (commodity or service). Two units of this good can be produced at cost ¥10. It is impossible to produce a single unit at lower cost. There are two consumers, each of whom has inelastic demand for one unit.

- Consumer H is willing to pay up to ¥8.
- Consumer L is willing to pay up to ¥4.

If the price to both consumers is set at $p \leq ¥4$, then the good cannot be produced without subsidy. Both consumers will demand the good at such a price. The result will be that revenue is $p \cdot 2 < ¥10$, which is below the total cost of production.

If the price to both consumers is set at $¥4 < p \leq ¥8$, then the good also cannot be produced without subsidy. Only consumer H will demand the good at such a price. Revenue will be $p < ¥10$ which, again, is below the total cost of production. Obviously price cannot be set above ¥10.

However, if the good is sold to L at ¥3 and to H at ¥7, then the good can be produced and each consumer receives ¥1 surplus. That is, the price to each consumer is ¥1 less than his willingness to pay and revenue is ¥10, which covers the cost of production.

Note that resale makes price discrimination infeasible because of arbitrage.[4] Consumers who can purchase from the producer at a low price, will resell to those whom the producer charges a high price. For example, in the preceding example consumer L would purchase 2 units at price ¥3 and resell 1 unit at price slightly below the producer's price of ¥7. The producer's revenue would be ¥6, which fails to cover the cost of production.

2.3 Elastic (i.e., price-sensitive) demand

Demand elasticity matters for welfare, both for some reasons that apply to goods in general, and also for some reasons specific to payments.

In general, elasticity provides a further channel through which pricing affects welfare. When demand is inelastic, all prices that enable revenue to cover production cost generate equal surplus—only the *division* of surplus between producer and consumers is affected by a small price change (unless the price offered to some consumer moves across that person's reservation price). In contrast, when demand is elastic, price changes also affect surplus through the relationship between the producer's and the consumers' marginal conditions for optimization. As a consequence of the considerations just described, if the revenue effect of a price change is exactly offset by a tax/subsidy transfer from the consumers to the producer or *vice versa,* then it will have no welfare effect if demand is inelastic, but it may affect welfare if demand is elastic. Thus "nonlinear" pricing, such as a combination of a membership (or subscription) fee to join a payment network and a fee for each transaction on the network, can exploit elasticity in a way that affects welfare.

Regarding payment systems specifically, elasticity of demand represents (albeit imperfectly) the fact that transactors can respond to high prices by using netting and other forms of "bypass" to reduce their payment volume over the system in question. That is, when transactors have offsetting obligations to one another (for example, when A owes B ¥1 and B owes A ¥2), then they have the option either to pay one another separately through the payment system or else to make a single payment (or at least, when there are more than two transactors, a reduced number of payments) to settle their net obligations (for example, making a single payment of ¥1 from B to A).

[4] If only some consumers are able to resell to others, then prices above those offered to those potential resellers are subject to arbitrage.

The ensuing discussion of payments pricing will make reference to the following results regarding elastic demand.

1. If a subsidy to the producer can be financed by nondistortionary taxation of the consumers, then the only efficient pricing policy is to set the variable component of price equal to marginal cost for all consumers, and to subsidize any shortfall of revenue (which may include revenue from fixed fees charged to consumers) below total cost.

2. If resale is infeasible (or if it can be prohibited), then efficiency is achieved by a two-part pricing policy.

   - A constant price (e.g., membership or subscription fee) is charged to each consumer, and can vary between consumers.
   - All consumers are charged a variable fee, linear in demand, that is set equal to marginal cost (at the level of aggregate demand).

3. If resale is infeasible and pricing is constrained to be linear—that is, multi-part pricing is ruled out—then in some situations, price discrimination can enable production costs to be recovered (analogously to the example in section 2.2).[5] However, welfare will be lower than two-part pricing would achieve.

4. If the producer charges two-part prices and resale is feasible, then there are two possibilities. Either pricing must be linear and nondiscriminatory, or else a subset of consumers will resell to the others. At most one consumer will pay a subscription fee.

3 How does the theory apply to actual markets for payments?

A prominent way of applying of nonlinear pricing to payment systems has been to identify the producer with a central bank. This identification has several limitations in principle.

One limitation is that both the payor and the payee benefit from the transaction, and that both might be charged for it, is ignored. This issue does not seem very significant. It is plausible that, except in the case of small "retail" transactions, the payor and receiver can typically achieve efficiency by negotiating to share the cost of payment.

Another limitation is that the ability of correspondent banks and clearinghouses partially to bypass the central bank by payment netting is only implicitly and inexactly captured by the representation of willingness to pay and resale.

A third limitation is that there is a dilemma regarding how the consumers of the nonlinear-pricing theory should be modeled: as individuals (households and non-financial firms), or as

[5] Consider demand functions $D_L(p) = 8 - p$ and $D_H(p) = 16 - p$, and cost function $C(q) = 58 + 2q$. The revenue-maximizing unit price is ¥7, which raises only ¥50 in excess of the variable cost of meeting demand. Setting the price ¥5 to consumer L and ¥9 to consumer H raises the full fixed cost, ¥58, in excess of the variable cost of meeting demand.

banks?[6] If as consumers, then the theory overstates the degree to which the central bank can discriminate. The central bank can actually discriminate among banks, but its settlements of the transactions of all of the bank's customers are priced identically. Moreover, to the extent that banks pass the central bank's payment prices through to their customers, and that those prices affect consumers' welfare, a consumer can change banks to get a better price—an option that the theory ignores. To relate this point to a point made in the preceding section, banks re-sell central bank payment services to their customers, and such re-sale may prevent the central bank from implementing price discrimination that would increase welfare if it were feasible.

On the other hand, if the consumers in the model are interpreted to be banks, then to the extent that actual banks are imperfectly competitive, their derived demand functions for payment services (which determine their reservation prices in the theory) diverge from the actual consumers' demand functions with which, conceptually, welfare analysis ought to be concerned.[7] Given the huge markups that banks charge their customers for access to central bank payment services, customers' demand surely is almost completely inelastic to the price charged by the central bank. Thus, to the extent that a central bank sees any demand response to its pricing policy, that response must be due predominantly to change in banks' use of netting and other bypass strategies, rather than to change in end-user demand.[8]

My personal assessment is that, despite its limitations, nonlinear-pricing theory lends credibility to three main conclusions:

1. Nondiscriminatory, marginal-cost pricing, with the central bank's revenue shortfall from cost recovery being subsidized by a tax, might well yield higher welfare than the best non-subsidized, discriminatory pricing policy can provide.

2. If subsidy is infeasible (as currently, in the United States, by law), then there might be circumstances in which full cost recovery would be impossible except for price discrimination.

3. Nevertheless, because of the arbitrage and bypass activities that are represented by demand elasticity and resale in the model, the scope for two-part pricing (i.e., comprising a subscription fee and a per-payment fee that does not depend on the value of the payment) is small.

Regarding the last conclusion, because arbitrage and bypass activities require technological and institutional changes that take some time to implement, price discrimination may provide some extra revenue in the short run. However, likely this extra revenue will mostly disappear in the long run. It is even possible that, once the arbitrage and bypass activities

[6]Throughout the ensuing discussion, 'bank' refers to any direct participant in the payment system.

[7]Actually, a bank's customers include firms as well as households. In the case of corporate transactions, there is an additional degree of separation between the theoretical construct and the conceptually correct basis for welfare analysis.

[8]Some people believe that such bypass strategies can raise issues of financial stability. Such issues are completely outside the scope of this theory.

stimulated by price discrimination have gone into effect, central bank revenue might decrease from its initial level.

Beyond the three basic implications just enumerated, I do not consider the model useful or trustworthy as a guide to central bank policy. For example, the Federal Reserve has sometimes attempted to justify price discrimination by conducting econometric studies suggested by this theory and related theories. In view of the limitations just mentioned, such studies are of dubious scientific value. The goal of conducting such studies—to ensure that price discrimination is used only for good reasons of public policy—is laudable but, in practice, the strategy has been burdensome and ineffective. My impression is that, when there have in fact been good reasons, the substantial time expended in conducting and vetting a study has delayed implementation of some pricing arrangements that bankers and other members of the public immediately recognized that it passes an intuitive "sniff test" for good public policy. On the other hand, some subsequent econometric studies did not provide sufficient discipline to prevent the Federal Reserve from using volume-based pricing to prolong its operation of some retail payment systems for which (in my view) there is not a good rationale for central bank involvement.

4 Some ideas regarding how central banks should price payment services

4.1 Irrelevance of the inverse-elasticity rule

Central banks are considering price discrimination as a way to raise the minimum required operating revenue. As explained above, the best solution from an economic perspective may not be to engage in discriminatory pricing, but rather to subsidize the payment service, if the subsidy can be financed by a reasonably non-distortionary tax.

If the central bank is unable or unwilling to subsidize its payment service, then its problem is to extract revenue from the transactors with the highest *total* willingness to pay. It is clear from the simple example in section 2.2 that total willingness to pay, rather than marginal willingness to pay, is the relevant aspect of a consumer's demand. The econometric studies undertaken by central banks (to which I briefly referred in the preceding section) are aimed at estimating the elasticity of demand—a statistic having to do with *marginal* willingness to pay—of various classes of payment-service customer. Moreover, an inverse-elasticity rule is optimal subject to a constraint that the producer must charge a single-part price, so its relevance is questionable if the central bank has latitude to implement multi-part pricing. A further issue is that the relevant elasticity from a welfare-economic perspective is that of *end-user transactors' demand,* not of their banks' demand. Transactors' demand elasticity, with respect to the price charged by the central bank to their banks, is arguably low, as was explained in section 3,. This can be true even when *banks'* estimated elasticity of demand is high. The upshot is that an inverse-elasticity rule may not provide very useful guidance to a central bank that needs to meet a cost-recovery target.

Rather, a central bank should keep in mind that its important problem is to find a

way to finance provision of a service that is a vital part of financial-market infrastructure. The relevant practical problem is simply to identify the banks with high total willingness to pay for that service (and the specific transactions for which they have high willingness to pay) and to design a pricing scheme that extracts the required level of revenue from them. Once that problem has been solved, the incremental welfare gain to fine-tuning prices would be of secondary importance, especially if the elasticities relevant to doing so are low. (When demand is inelastic, the welfare cost of departure from the inverse-elasticity rule is small.) In particular, it is doubtful that there would be significant welfare gain from using using inverse-elasticity prices, relative to using other prices that meet the central bank's cost-recovery requirement.

4.2 Subsidy as an alternative

Many central bankers are skeptical of the prospect to subsidize their payment services. Reserve requirements (as opposed to clearing-balance requirements tied closely to a bank's payment-system activity) and subscription to central bank stock are traditional, implicit, taxes that central banks levy on banks for that purpose (among others). Lower inflation rates and increased competition between national banking and financial systems have reduced the real income that those traditional taxes provide. Those traditional funding options are not the only ones available, however. For example, Governor Mervyn King-1999 of the Bank of England has proposed that central banks might consider requesting legislative authority to fund themselves by levying an explicit tax on banks. He argues that such a funding arrangement would not compromise central bank independence—a legitimate concern that has led central bankers mostly to stick with their traditional revenue sources.

If a central bank payment service is to be subsidized, then it is important to provide the public with an explicit, cogent rationale. Such a rationale is implicit in the economic theory surveyed above, as follows. The central bank provides public infrastructure, available to all who require it, for a public purpose.[9] Subsidizing the service and pricing at marginal cost is the welfare-maximizing arrangement.

Some operators of proprietary payment services argue that they should also receive a subsidy, if the central bank's service with which they compete receives a subsidy. However, that argument presupposes that it is efficient to have several payment systems actively compete with one another. Because a payment service is a natural monopoly (for reasons of both scale economies and network effects), the argument is unsound. Given that the service is operated prudently, accessibly, and cost effectively, there is no rationale for subsidizing additional (privately provided) payment services.

A noteworthy feature of this natural-monopoly argument for treating the central bank's payment system asymmetrically from other payment systems, is that the argument does not rely on any alleged superiority of the central bank's payment system with respect to minimizing systemic risk or other performance criteria. All major private payment systems

[9]The availability of the payment system to all financial intermediaries that reasonably should use it, is a "core principle" stated by the CPSS .

in G-10 countries make substantial investments (including design decisions that would not otherwise be cost effective) to mitigate systemic risk. An economist would be unable to make an objective, convincing case, on the basis of publicly available evidence, that those private systems, as operated and supervised today, carry material systemic risk. Indeed such a case, if it could be made regarding some system, would be an indictment of the central bank that oversees the system. Absent a convincing case case that a payment system operated by a central bank is intrinsically superior to a proprietary system, the logical conclusion of an argument for subsidizing the payment systems least exposed to systemic risk would be that private large-value systems should receive the same subsidy that the central bank's system receives. The natural-monopoly rationale does not share that conclusion.

A sound policy that involves subsidy must address the issue that the central bank thereby becomes insulated to some extent from competitive pressure for good, cost-effective management. My personal perception is that good management in this case has two major requirements.

- To design and implement stringent, publicly visible, management controls over payment operations.
- To safeguard against misallocation of subsidy by senior executives/policy makers.

The first of these requirements is largely a standard exercise in corporate governance. The part that perhaps goes beyond what is standard, is the idea that the controls should be publicly visible. In the case of a private corporation, audit controls and their implementation are visible to the corporate directors, who represent the all owners. In the case of a central bank, there is not a comparable (notably, independent of management) internal constituency. The government should not act as a surrogate, since an arrangement to do so might be abused to violate central bank independence in monetary policy. The alternative is to facilitate scrutiny by the public, especially through the agency of the press, academic researchers on payments and central banking, and banking- and payment-industry lobbying associations.

The risk of misallocation of resources by senior executives and payment-policy makers within the central bank mas much more to do with "mission creep," than with any deliberate malfeasance.[10] A central bank can address this risk by clearly identifying, and publicly announcing, a tightly circumscribed mission for central bank payment operations. It should take measures to ensure time consistency such as adopting a public-comment process for payment-service expansions, and perhaps additional procedures that can enable bankers and private payment-system operators—the citizens with the strongest interest in monitoring and constraining the central bank—to be politically effective in opposing "mission creep" that might be a future temptation.

[10]Deterring and potentially detecting such malfeasance by senior management is within the purview of the audit controls envisioned in the first requirement.

4.3 Don't adopt a retail business plan for a wholesale system

Unless the fixed-cost component of a payment system can be subsidized fully, a central bank must make a decision regarding the strategies that it will use to raise revenue from users of the system. One such strategy is to encourage the use of the central bank payment system for relatively small transactions, although those transactions are not the rationale for central bank operation of the system, nor are they the transactions that the architecture of the system was designed to optimize. I believe that such a strategy is neither desirable nor effective, though.

Essentially by definition, parties to small transactions have rather low willingness to pay to make them, so a central bank relying on revenue from those transactions would have to charge a low price but to attract high transactions volume. Over time, however the central bank will face progressively stiffer competition from commercial payment systems for this business. Those commercial systems will integrate payment processing *per se* with "value-added" services, such as transmission of remittance information, that integrate payment with other business processes. Competing in that dimension would be costly for the central bank. Design and implementation of value-added services is not a central bank's comparative advantage. In the context of a highly secure, automated system, outsourcing the provision of value-added services to a third-party contractor would raise significant security issues. Some highly valued add-ons (such as translation of data to formats supported by major ERP systems) would involve changes to the design of the computer software of the payment system. The benefit of higher revenue resulting from such changes would have to be weighed against various costs, including the cost that making software changes carries some operational risk. For these reasons, the long-term prospect for small-value transactions as a revenue source is uncertain.

4.4 Be more entrepreneurial about price discrimination

An alternate strategy is to extract much more revenue from parties to large-value transactions, if (as I suspect) they have higher willingness to pay than do the small-value transactors. There is no good reason of public policy not to shift the costs of operating a central bank's payment system to the parties to the systemically important transactions, with which the rationale for central bank operation of a payment system is predominantly concerned.

One such strategy is value-based pricing. It is plausible that fees could be raised high enough to generate substantial revenue without appreciably shifting behavior of the largest-value transactors (for example, consolidation of very large payments into huge payments, once the flat part of the price schedule had been reached, to avoid paying multiple fees). Those transactors have already used considerable ingenuity to minimize their fees for central bank payment services. Their options for doing more are likely very limited. I do not know (and probably no one knows) whether high price on settlement of netted payments through the central bank—perhaps even at a price of something like \$0.10 per \$1M value of payment (62.5 times the sum of origination and receipt fees for a \$100M Fedwire payment between high-volume Fedwire customers today)—would cause clearinghouses to find ways to increase

their netting ratios significantly beyond their current levels, or to find other ways to "bypass" use of the central bank's payment system.

Another pricing strategy would be to switch from a business model of earning revenue predominantly from small fees on many small transactions to one of earning revenue from large fees on large transactions. One such strategy would be to engage in explicit, "first-degree," price discrimination by charging a premium for payments to settle transactions that are netted through a clearinghouse. There are two polar versions of such a strategy. A potential drawback is that such an arrangement might discourage banks from settling through the clearinghouse, reducing the amount of surplus that they receive from the ability to net payments and from the ancillary services that the clearinghouse provides. One possibility is that, from a welfare point of view, such a reduction of netting should not be cause for concern today. The farther back one goes in history (especially, to the days when gold had to be transferred by stagecoach to settle payments), the higher were the real, variable costs of gross settlement, and the correspondingly greater was the social value of netting payments. Today, however, when the price of making a payment is approximately the average cost of operating the central bank's payment system, and when that average cost is orders of magnitude higher than the marginal cost, the social benefit of netting is much smaller than the magnitude of resultant cost shifting from participants who can net payments among themselves to others who, for one reason or another, cannot do so. It seems equitable for the central bank to price its service in such a way that large banks that are clearinghouse members pay a share of central bank operating costs that is proportional to their sizes, and specifically that joining the clearinghouse should not be a "bypass strategy" that enables them to evade making that contribution.

An argument that the reduction of netting carries a welfare cost might be made on grounds of "systemic stability." Although many central bankers believe that using an RTGS system minimizes the risk of harm from a payor's default, some academic researchers have made a cogent rejoinder. According to that view, netting reduces the maximum value of payments that have to be made (or, at least, shifts the distribution of payment values to a stochastically dominated distribution), and consequently it has the potential to avoid situations in which a payor would have to default. That is, in some cases a payor might be able to make the payment that it owes in a netting system, but not the larger payment that it would owe in an RTGS system. This beneficial effect has to be weighed against the isolation of direct exposure to a default that an RTGS provides when a default does occur, in making an overall comparison between the stability of the two systems. If the advantage of netting in avoiding defaults dominates the advantage of RTGS in isolating defaults when they occur, then a netting system is preferable. In that case, a central bank must take care not to raise the price to clearinghouses so high that netting of large value transactions would become unprofitable for banks. Similarly, clearinghouses do provide ancillary "value-added" services to their members, and a central bank should not set prices for its clearinghouse customers so high that they cannot profitably provide those services.

5 Concluding remark

The discerning reader will notice that I have discussed one polar case of a business model relying on clearinghouses for revenue, but not the other one. The other polar case would to discriminate *in favor of* a clearinghouse in the variable component of the fee, but to charge a very large fixed fee to a clearinghouse. In fact, this is the business model to which I think that central banks are likely to be forced eventually, unless they decide to put clearinghouses out of business.

A clearinghouse is a customer that re-sells central bank settlement services to other potential customers. Result 4 of section 2.3 implies that, in that situation, either the central bank must implement nondiscriminatory, single-part pricing (which many central bankers believe to be currently or imminently incapable of generating sufficient revenue to recover costs), or else have a single direct customer (which would be the clearinghouse).

The upshot of this business model is that some clearinghouse will become the "front end" for the central bank's settlement service. There is an inescapable question of what would be gained from this outcome. An answer might be that the clearinghouse will be able to use types of nonlinear-pricing arrangement that the central bank would be constrained (perhaps by political considerations) from using. Then, why is an arrangement in which the clearinghouse becomes a front end for the central bank (and banks are essentially forced to join a proprietary clearinghouse in order to have access to central bank settlement) preferable to one in which banks would obtain access to a settlement account and short-term credit at the central bank—likely on a subsidized basis—and would be free to purchase whatever ancillary services they need from private providers on a market basis. For which ancillary services are the economies of scope with the central bank's core services really so large that they outweigh the evident *prima facie* disadvantage of the front-end-clearinghouse outcome? Is there a tightly circumscribed set of ancillary services, by providing which the central bank can achieve the most significant economies of scope, while keeping the central bank's part of the total cost of operating the payment system at a low enough level so that economically efficient subsidy would not encounter political opposition? These are questions that central bankers need to consider while they are still able to make ends meet, rather than waiting until a cost-revenue "crunch" drives them into a front-end-clearinghouse solution by default. In doing so, scientific advice from academic experts about the issues sketched here will surely be extremely helpful.