Payment Scale Economies, Competition, and Pricing.

David B. Humphrey
Department of Finance
Florida State University, U.S.A.

May 2009

Joint European Central Bank-De Nederlandsche Bank Conference on
Retail Payments: Integration and Innovation
Frankfurt, Germany

Payment Scale Economies, Competition, and Pricing.

1. Introduction.

The topics of payment scale economies, competition in payment services, and per transaction pricing of payment instrument use are intertwined in an important way. First, bank scale economies are seen to be quite large when operating costs are related to expanding payment volume and the size of ATM and branch office networks needed to deliver services to depositors. Second, these economies have lowered average payment costs in Europe and one can ask where the lower unit costs have ended up, a question directly related to competition among banks in supplying payment services. This task would be relatively straightforward if payment use was priced on a transaction basis since these prices would likely be falling over time. However, transaction pricing is relatively rare.

Third, higher revenues to cover higher total payment operating costs, even as unit costs are falling, will largely come from payment prices mostly disconnected from changes in payment volume (such as fixed fees tied to deposit accounts, minimum balance requirements, low interest paid on deposits, etc.). This is why payment services are typically viewed as a "cost center" not a "profit center". Transaction pricing would benefit banks by tying revenues more closely to costs so management obtains a clearer picture of where their profits are generated, permitting more efficient use of internal investment resources. Transaction pricing would also benefit consumers allowing them to balance better the cost of using different payment instruments with their assessment of the benefits, saving resources for the economy as a whole.

Finally, the more that current bank service pricing is unrelated to payment volume and service delivery expansion, the more that these prices will likely rise over time to cover the expanded operating expenses even when strong scale effects are evident. While this may suggest the exercise of market power (c.f., European Commission, 2007), an alternative interpretation concerns the need to expand revenues as total costs rise since revenues are not closely tied to payment and service delivery volumes. When prices are tied to unit costs, revenues rise "automatically" with volume and prices should fall as scale economies are realized.

In what follows, payment scale economies are estimated for 11 European countries in Section 2. Estimates are obtained using different methods—some applied to individual countries at different points in time and others applied to all 11 countries over time in a panel framework. The results are rather similar, giving us confidence in the strong scale effects found. In Section 3, these scale estimates are used along with other bank cost and productivity indicators to assess the relative competitive efficiency of banks across the same 11 countries. The procedure developed borrows from the efficient frontier literature and estimates a competition frontier. These results are then compared to more standard measures of banking competition. Section 4 outlines the potential benefits to banks as suppliers of payment services and consumers as users from transaction-based pricing of payments. Since unit payment costs have been falling rather than being stable, this alters the cost-benefit trade-offs that need to be observed and made by both banks and consumers. But to respond efficiently to these changes, prices and revenues need to more closely mirror unit costs and thus changes in volume. Conclusions are presented in Section 5.

2. Payment Scale Economies.[1]

2.1 Payment Scale Economies: Different Countries and Different Methods.

Determining scale economies for financial institutions – much less payment scale effects – has been difficult due to a lack of appropriate data. Instead of measuring the flow of banking payment, deposit account maintenance, security transaction activity, and loan initiation and monitoring services directly, inferences on how costs may vary by size of bank and volume of service flow have been typically obtained by relating total operating plus interest expenses across banks and over time to the value of their stock of loans, securities and (sometimes) deposits, or some other combination of on- or off-balance-sheet positions in a regression-based cost function framework. In addition, information does not normally exist regarding the adoption of specific technical and other cost-saving innovations in banking and the default has been to assume that unknown technical change occurs linearly (or quadratically) with the passage of time and/or is somehow associated with (embodied in) the cost share or price of particular inputs.

An additional problem, especially in U.S. regression studies, has been that scale economy estimations using cross-section data with quadratic functional forms (e.g., translog) will understate the economies associated with large banks in the sample. This is because minimizing the sum of squared errors for all sizes of banks in a sample will be dominated by the fit estimated for the smaller banks in the sample since these banks often reflect the vast majority of observations. While the downward sloping portion of an estimated average cost curve will properly reflect the scale economies for smaller banks, it can imply that diseconomies are associated with the largest banks that are under-represented in the sample (c.f., McAllister and McManus, 1993). This problem is solved when a more flexible Fourier function form or a spline function is used in place of the translog form or when the sampled banks are more equally represented across the range of asset size-classes in the data set. Fortunately, the translog and Fourier forms give very similar values at the mean of the data but not for the very largest banks in the sample.

The payment scale economy results shown in Table 1 do not have the problems noted above either because they are not regression-based studies (the first four results) or because the potential bias from including small banks in the sample has been adjusted for (the fifth study). For example, in the first study card payment costs in Norway were observed to rise by 153% between 1994 and 2001 while card transaction volume rose by 355%. The ratio of these percent changes yields a scale economy estimate of 0.43 so that a doubling of card transaction volume implies that card operating costs rise by only 43%

[1] Material in this section is drawn from Bolt and Humphrey (2007).

so average cost falls by 29%.[2] However, these card payment costs include concurrent increases in labor expenses that should have been excluded to properly reflect scale effects alone. Since the cost of living index in Norway rose by 18% over this period, it is clear that holding the price of labor inputs constant would reduce the percent change in card costs so the true scale value would be lower than 0.43. This would indicate greater scale benefits since scale economies should be a function of the number of transactions, holding input prices and changes in technology constant.

Table 1: Payment Scale Economies: Four Countries, Different Methods.

	Cash	Card
Norway 1994-2001		
d(payment costs)/d(volume)		0.43
Netherlands 2002		
Marginal Cost/Average Cost	0.37	0.39
Belgium 2003		
Marginal Cost/Average Cost	0.25	0.39
U.S. 2005		
d(payment costs)/d(volume)		0.31 – 0.39
Netherlands 1997-2005		
Bank Data, Econometric Model		0.27 – 0.31

Sources: Gresvik and Øwre (2002); Brits and Winder (2005), Table 4.3; Quaden (2005), Table 3; First Annapolis Consulting (2006); Bolt and Humphrey (2008b).

Two other non-regression-based analyses of payments in the Netherlands (Brits and Winder, 2005) and Belgium (Quaden, 2005) estimate the average total cost and incremental variable cost of using cash and a debit card at the point of sale in 2002 and 2003. These estimates are based on data for a point in time, rather than over time, so there is no need to adjust for changes in labor input prices or technical change. As scale economies can be expressed as the ratio of marginal to average cost, this gives a point estimate of payment scale economies.[3] The implied scale economy for cash ranges between 0.37 in the Netherlands to 0.25 in Belgium while the implied scale economy for debit cards in both countries is 0.39. These card scale economies are not very different

[2] If payment costs and transaction volume are both initially at 100, average cost is 100/100 = 1.0. When volume doubles to 200 but costs only rise by 43% to 143, the new average cost is 143/200 = 0.715 and the percent reduction in average cost is (1.0 - .715)/1.0 = 29%.

[3] Scale economies (SCE) equal $(\partial OC/OC)/(\partial Q/Q)$ or $\partial\, ln\, OC/\partial\, ln\, Q$ where OC is operating cost and Q is a measure of output and values < (>) 1.0 represent economies (diseconomies). SCE can be rewritten as $(\partial OC/\partial Q)/(OC/Q)$ = marginal cost divided by average cost, where incremental cost approximates marginal cost. As bank interest expenses should be unaffected by changes in payment volume, operating cost – not total cost – is the appropriate cost measure to use.

from approximate PIN and signature debit card scale values of 0.31 and 0.39 respectively, for card issuing banks in the U.S. derived from data in the 2005 Issuer Cost of Payments Study (First Annapolis Consulting, 2006). Similar payment scale estimates were obtained for the Netherlands using confidential individual bank payment data in an econometric model, with results reported at the mean of the data (Bolt and Humphrey, 2008b).

2.2 Payments versus Service Delivery: An Output Characteristics Approach.

A different approach which can provide more solidly-based scale economy estimates than traditional analyses which uses balance sheet asset value data to represent banking physical "output", relates bank operating (not total) costs to measurable physical characteristics of banking output associated with payment processing and service delivery levels and mix—a set of four banking output characteristics. Here the focus is on those activities and expenses directly associated with the provision of payment services. Interest expenses paid to depositors and with a mark-up charged to borrowers are functionally separable from these activities. This approach determines how the level and mix of payment activities, along with the number of ATMs and bank branches, are directly associated with the size of a bank and its labor, capital, and materials operating cost which reflects scale economies. In this regard, our approach represents an alternative and more direct way to identify the effect of scale and technical change on costs in banking.[4] As point of sale and bill payment transactions are jointly processed in the deposit accounting function while aspects of service delivery are jointly produced via branches and ATMs, these two activities can be considered functionally separable.

The variation of operating cost (*OC*) across 11 European countries annually over 1987-2004 is used in a translog cost function to derive scale economies for point of sale transactions, bill payments, as well as for ATM and (standardized) branch office networks. All cross-country value/price data are in terms of purchasing power parity U.S. dollars since the euro did not exist until 1998. A translog function (1) and a more complex Fourier cost function (not shown) were both estimated for robustness. As the scale results were very similar and the figures of predicted unit payment (seen below for the translog) are almost identical, only the translog results are reported here. The translog cost function is estimated jointly with cost share equations (2) for labor:[5]

$$\begin{aligned}\ln OC = \alpha_0 + \sum_{i=1}^{4} \alpha_i \ln Q_i + 1/2 \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_{ij} \ln Q_i \ln Q_j + \sum_{i=1}^{4} \sum_{k=1}^{2} \delta_{ik} \ln Q_i \ln P_k \\ + \sum_{k=1}^{2} \beta_k \ln P_k + 1/2 \sum_{k=1}^{2} \sum_{m=1}^{2} \beta_{km} \ln P_k \ln P_m + \ln e_{OC}\end{aligned} \quad (1)$$

[4] Our panel data set is a cross-section of 11 countries over 18 years. The standard way to identify scale effects from technical change is to presume that cross-section variation identifies scale while time-series variation identifies technical change. A problem is that payment volume tends to rise at a fairly constant rate over time while the simple passage of time itself is used as the (unknown) index of technical change. This collinearity problem is reduced when the time pattern of important technical changes is known (which rarely occurs).

[5] The standard coefficient symmetry and linear homogeneity in input price restrictions are imposed in estimation.

$$S_k = \beta_k + \sum_{m=1}^{2} \beta_{km} \ln P_m + \sum_{i=1}^{4} \delta_{ik} \ln Q_i + \ln e_S \qquad (2)$$

where:

$OC =$ total operating cost, composed of all labor, materials, outsourcing and capital consumption costs (but no interest expenses);

$Q_i =$ four output characteristics (i = 1, ..., 4) composed of point of sale (*POS* = card and check) and bill payment (*BP* = electronic and paper giro) transactions along with the number of automated teller machines (*ATM*) and size standardized branch offices (BR^{STD}). The standardization procedure is explained below;

$P_k =$ two input prices (k = 1, 2) denoting the average labor cost per bank employee and an opportunity cost approximation to the price of bank physical capital and materials inputs represented by each country's market interest rate; and

$S_k =$ the cost share for the labor input (the "mirror image" capital/materials input cost share is deleted to avoid singularity). It is expected that operating costs not directly associated with the type of payment or mode and level of service delivery will be represented in the intercept term.

It is clear that a single payment transaction in one country, whether by card, check, giro or cash withdrawal from an ATM, is equivalent to a single payment transaction in another country. While there is a great deal of size homogeneity among banking offices within a single country, this is not the case for banking offices across countries.[6] Although the average number of workers per branch office across our 11 countries was 15.9 it was only 6.7 in Spain but 26.2 in the U.K. Clearly, each branch office in Spain (or Belgium with 10.7 workers per branch) is providing a different level of banking service output than occurs in the U.K. (or Germany with 15.8 workers per office). An even greater dispersion exists for the value of assets generated per branch.[7] It is thus necessary to standardize branch office size according to some benchmark to make them more comparable across countries.[8] Our view is that the production function relationship reflected in the "labor/capital" ratio – tying the labor input to branch payment processing, cash access, loan origination and monitoring – reflects better the service flow produced by branch offices than the unadjusted number of branches themselves. France, with an

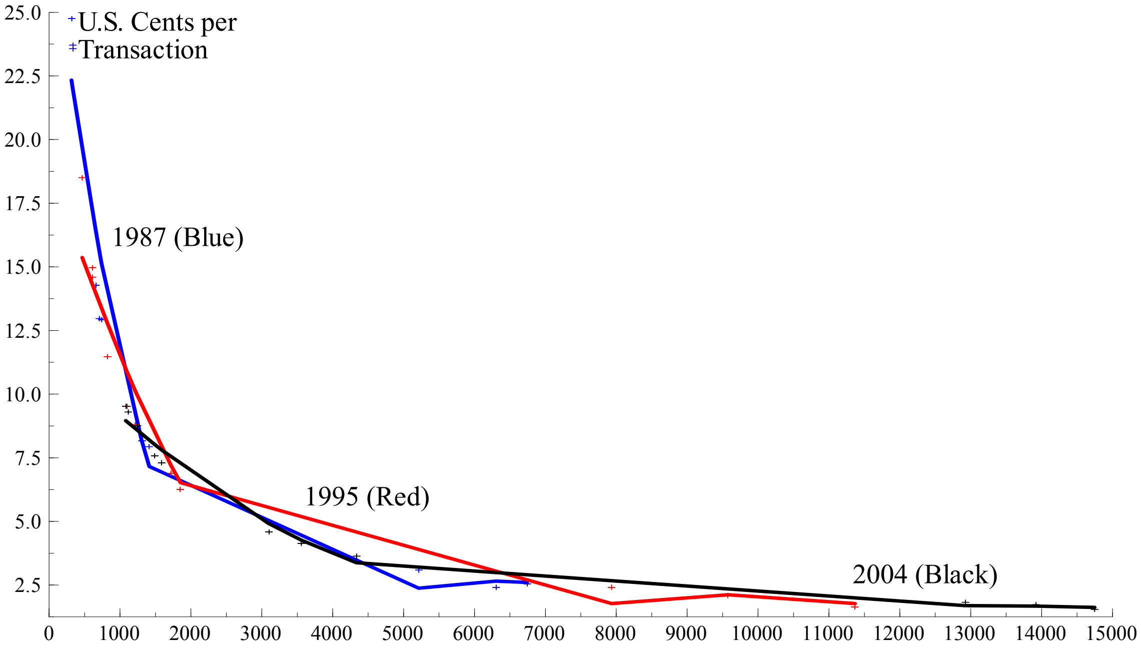
[6] When new branches are established in a country they tend to expand in parallel with the growth of their local market. Once a branch reaches a given size, further bank growth occurs via establishing additional new branches (or through mergers). The large differences in branch office size across countries is probably due to differences in population density, earlier norms developed when (prior to ATMs) cash could only be obtained from a branch office and greater reliance in some countries (France and the U.K.) on checks versus cards or cash.

[7] The average total asset/branch ratio ranges from \$30 to \$193 million with an overall average of \$96 million.

[8] In partial support of the adjustment, the level of operating cost (or total assets) varies more closely with the standardized branch measure (R-squared = 0.76) than it does with the unadjusted number of branches (R-squared = 0.57).

average of 16.1 workers per office over our 18-year period (close to the overall mean), was selected as the branch benchmark and other countries were adjusted accordingly.[9]

Figure 1: Predicted Unit Payment Costs versus Payment Transaction Volume.



2.3 Payment Scale Economies for 11 European Countries.

Based on the cost model presented above, Figure 1 shows how predicted bank payment cost varies by non-cash transaction volume across countries. Norway, Finland, Denmark and Sweden form the first part of each curve while the middle reflects Belgium, the Netherlands, Spain and the U.K. The last part of each curve covers Italy, France and Germany. The curves are cubic splines of the ratio of predicted payment costs (for both point of sale and bill payments together) divided by the total number of transactions (Y-axis) and arrayed against total transactions (X-axis). Importantly, the curves in Figure 1 are not average cost curves as the level of these curves (in U.S. cents per transaction) is too high.[10] However, the slopes shown are a fair reflection of how payment unit costs

[9] Specifically, the number of each country's banking offices (*BR*) was adjusted as follows: $BR^{STD} = BR[(L/BR)/16.1]$, where *L*/*BR* is the observed labor/branch ratio in each country for each year and 16.1 workers per office is the standardized size of each office. This gives the number of standardized, size-adjusted branches (BR^{STD}) which is used for each country in the estimations, not *BR*. For example, the average *L*/*BR* for the U.K. was 26.2 workers per office so dividing by the French benchmark gives 26.2/16.1 = 1.63 which increases by 63% the number of "standard" U.K. branches used in the analysis. In contrast, since Spain had an average *L*/*BR* of 6.7 workers per branch office, dividing by the French benchmark gives 6.7/16.1 = 0.42 which reduces the number of "standard" Spanish offices by close to 60%. This was done for each country for each year.

[10] This because predicted payment operating costs are obtained by evaluating the estimated cost function by holding input prices and the number of ATMs and branches constant at their mean value. Although

change with payment volume. These slopes--and their associated scale economies--vary with payment volume but are quite similar for the three years shown. This suggests that a changing payment mix (reflecting possible scope economies) or disembodied technical change adds little to the operating cost reductions shown over time.

The payment-related scale economy (*SCE*) or cost elasticity estimates using the translog form are presented in Table 2. The scale economy for the i^{th} payment service (SCE_i) from the cost function (1) is $\partial \ln OC / \partial \ln Q_i = \alpha_i + 1/2 \sum_{j=1}^{4} \alpha_{ij} \ln Q_j + \sum_{k=1}^{2} \delta_{ik} P_k$, where: *i,j* = point of sale transactions (*POS*) and bill payments (*BP*), number of ATMs (*ATM*), and number of branches (BR^{STD}). Countries are ranked according to their total non-cash payment volume in 2004 (Column 1). Note that in 2004, non-cash payment volume in our 11 European countries was 59 billion transactions while in the U.S., it was 85 billion. On a per person basis, the U.S. makes 74% more non-cash transactions per year than individuals in Europe, indicating a greater degree of cash replacement.

Table 2: Operating Cost to Total Assets, Non-Cash Payment Volumes, and Scale Economies by Country.

	Payment		Average	Point	Bill			Total
	Volume		Payment	of Sale	Payment	ATM	Branch	Realized
	2004, Mil	OC/TA	SCE	SCE	SCE	SCE	SCE	SCE
Germany	14,748	-40%	0.23	0.06	0.17	0.22	0.59	0.31
France	13,926	+1	0.30	0.08	0.22	0.31	0.36	0.47
U.K.	12,919	-52	0.35	0.11	0.24	0.36	0.27	0.54
Spain	4,335	-50	0.30	0.10	0.20	0.23	0.48	0.45
Netherlands	3,563	-33	0.17	0.09	0.09	0.24	0.65	0.24
Italy	3,094	-29	0.21	0.05	0.16	0.17	0.62	0.30
Belgium	1,594	-23	0.20	0.10	0.10	0.26	0.59	0.26
Sweden	1,488	-38	0.33	0.18	0.15	0.39	0.37	0.21
Finland	1,244	-59	0.35	0.19	0.16	0.40	0.34	0.20
Norway	1,117	-60	0.34	0.19	0.15	0.40	0.34	0.23
Denmark	1,081	-39	0.24	0.12	0.12	0.28	0.52	0.37
Average	5,374	-34%	0.27	0.11	0.16	0.30	0.47	0.40

The change in bank operating costs as a ratio to asset value, an approximate indicator of the unit cost of producing banking services, is shown in Column 2. Reductions ranged from -23% for Belgium to -60% for Norway but was +1% for France. For all 11 countries, the average was -34% over 1987-2004.

constant, these mean values and their associated costs add to the level of the payment costs being predicted. The inability to obtain average costs for a subset of outputs in a multi-output cost function was noted in Baumol, Panzar and Willig (1982).

The average scale economy for all payments for each country using the translog cost function is shown in Column 3. The average across countries is 0.27 and indicates that substantial scale effects would be expected as payment volume rises.[11] The point of sale and bill payment SCEs appear quite low and at 0.11 and 0.16 respectively, are considerably lower than independent estimates for Norway, the Netherlands or Belgium using different data sources. The independent estimates include some branch office expenses while the econometric analysis holds branches (and their associated expenses) constant in order to focus on only payment processing costs and this likely accounts for most of the difference.

To gauge robustness, the estimation process was repeated using data from earlier versus later time periods, segmenting countries by smaller versus larger payment volumes and adding 11 country-specific indicator variables to the model. The average payment SCE results were not markedly different by earlier or later time periods nor for countries having smaller versus larger payment volumes compared to using all 11 countries together for the entire time period.[12] Also, adding 18 time-specific indicator variables yielded little reduction in the positive autocorrelation evidenced in the Durbin-Watson statistic for the panel data and had no effect on the number of significant parameters (as 17 out of the 20 that were common were significant).[13] Finally, a differencing parameter ($\rho = 0.85$) was estimated from the residuals from equations (1) and (2) and used to transform the data. This generated a D-W value close to 2.00 but the cost function concavity condition was not met and branch scale economies were negative, suggesting that operating costs fall absolutely for countries with larger branch networks.[14] Based on these results, a fixed effects estimation framework would not offer any improvement in estimation. More importantly, fully adjusting for positive autocorrelation yields anomalous results.[15] Consequently, we focus on the scale economy estimates in Table 2

[11] Average scale economies of 0.27 were also obtained using cost data specific to 8 European payment processor operations over 1990-2005 in a translog econometric model (Beijnen and Bolt, 2009, Model 3b).

[12] The translog average payment SCE is 0.27 over the entire period and is 0.24 using data for only the first half (1987-1995) and 0.35 for the second half (1996-2004). Restricting the estimation to the 5 countries with the smallest payment volumes gives an average payment SCE of 0.30 while for the 6 largest countries it is 0.36. Estimation with 11 country-specific dummies raises the average payment SCE to 0.34 so country identification explains some of the reduction in operating cost, leaving less to be associated with our four outputs.

[13] Estimation with 18 time-specific dummies gives an average payment SCE of 0.20. Here, all the time dummies are positive indicating that the passage of time raises (not lowers) operating cost. This anomalous result likely occurs because we already apparently identify the banking outputs associated with cost curve shifts over time in our originally specified variables.

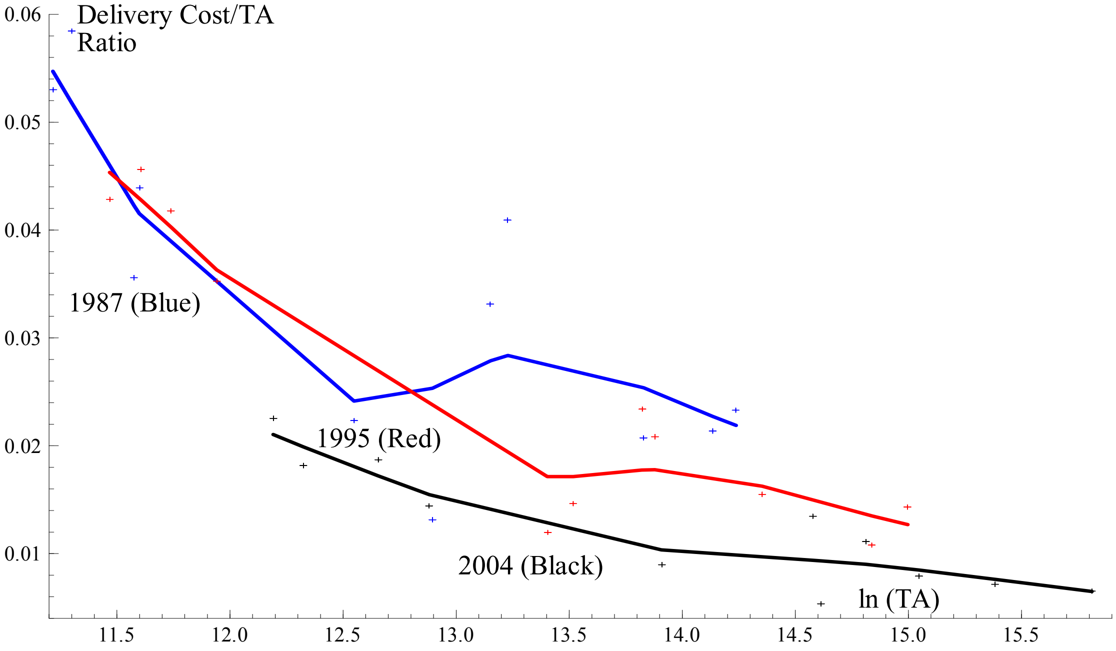
[14] Setting $\rho = 0.75$ met the concavity condition but branch scale economies were still negative. Although the time sequence of observations is fixed for each country, the ordering of countries in the panel data is arbitrary and changes here did lead to some reduction in positive autocorrelation (but of course, had no effect on parameter or scale results).

[15] Indeed, a grid search using values of $\rho = 0.25, 0.50, 0.75$ and 1.0 to transform the data indicated that the average payment SCE (Column 3) was fairly stable up to the point where $\rho = 0.50$ but completely fell apart when $\rho = 1.0$.

which applies a translog cost model using unadjusted levels data for the entire time period for all 11 countries together (in panel econometrics jargon, a "total pooled regression"). To correct for heteroskedasticity, Robust-White standard errors were calculated and are shown with the estimated parameters in Bolt and Humphrey (2007).

Strong scale economies were also estimated for ATM and branch office networks and were 0.30 and 0.47, respectively (Table 2). These scale effects are illustrated in Figure 2 where the predicted ATM and branch office delivery costs for different years are summed and expressed as a ratio to bank asset value (Y-axis) and arrayed against the log of asset value (X-axis). As was the case for Figure 1, the cubic splines in Figure 2 are not average costs but their slopes do reliably reflect the scale economies associated with delivering banking services to users.

Figure 2: Predicted Delivery Cost/Total Assets versus ln (Total Asset Value).



The last column in Table 2 illustrates the scale effects associated with both payments and service delivery activities together and is an indicator of total bank scale economies. It is a weighted average of the four scale measures in the table since a simple sum of these four scale values would be inappropriate as the quantity expansions of point of sale transactions (140%), bill payment transactions (151%), ATMs (434%) and branch offices (9.8%) are not equal. Indeed, the number of branches in all but 4 of the 11 countries decreased absolutely rather than expanded.[16] Overall, if a bank doubles its size over time

[16] A simple sum of individual scale elasticities is only appropriate when the percent changes in output quantities are not very different. This condition is probably close to being met when balance sheet values are used to approximate banking output quantities, the usual practice in the literature, as opposed to the output characteristics approach used here which reflect actual quantities and where the percent changes are markedly different.

or through a merger, its operating costs is estimated to only rise by 40% so its overall average costs will fall.

2.4 Payment Scale Economies: Volume Growth versus Back Office Consolidation.

Payment processing center volume can rise in two different ways: (1) each bank's payment volume in a country rises as users in that country make more payments over time; or (2) each bank's payment volume in a country is combined with the payment volume of banks in other countries in a single cross-border payment processing center at a point in time.

In (1), rising payment volumes at each bank leads to lower unit payment costs but generates higher total operating costs. Since per transaction pricing is not common in Europe or the U.S., total bank revenues do not rise "automatically" with rising operating costs and payment prices that do exist need not fall as unit cost falls due to realized scale economies (a movement along a downward sloping average cost curve as each bank's payment volume rises). For this reason, payment activities are often considered a bank "cost center" rather than a "profit center".

In (2), back office consolidation of payment processing centers within or across borders can lower each bank's unit costs as well as decreasing each bank's total operating cost since, at least initially, each bank's payment volume is constant and the volume expansion that generates the scale economies comes from the consolidation (a downward shift in the average cost curve for all banks even as their individual payment volumes remain constant). This can be one benefit of a single euro payments area (SEPA).

To illustrate, consolidating debit card volume in Belgium (671 million in 2004) with that of the Netherlands (1,247 million) would expand volume at a cross-border Netherlands processing center by 54%, suggesting that unit processing cost may decline by 33% if the Netherlands/Belgium scale economy of 0.39 in Table 1 is the appropriate metric.[17] If the average payment (0.27) or average point of sale (0.11) SCEs in Table 2 were used, the change in unit processing costs could be -39% or -48%, respectively. These reductions in unit processing costs would, however, not translate directly into reductions in total unit costs since additional telecommunication expenses would need to be factored in (as well as the amortized cost of the expanded investment in processing equipment).

Consolidation of existing processing arrangements is clearly more difficult to arrange than merely piggy-backing the currently small amount of cross-border card volume onto current national processing center operations. Even so, Interpay, the Dutch processor of debit card and giro payments, just completed a merger with its German equivalent Transaktionsinstitut (TAI) at the end of 2006. The merged company – called Equens – will double its yearly processed volume from around 3.5 billion transactions each to 7

[17] Since $\partial \ln AC / \partial \ln Q$ can be re-expressed as SCE – 1, the percent change in average operating cost is $\partial \ln AC = \partial \ln Q$ (SCE - 1) = 54% (0.39 – 1) = -33% when payment volume expands by 54% and the SCE = 0.39. AC is average operating cost.

billion in total, making it one of the largest payment processing centers in Europe. In a recent press release (January 23, 2007), Equens announced that the volume growth from their merger will generate large scale benefits and by 2010, they could "realize cost savings of approximately 25 percent, which include a reduction of around 400 full-time positions and plan to increase the yearly processed payments transactions to 10 billion".[18]

In sum, there are strong scale economies associated with the provision and processing of payment transactions by banks as well as in how banks deliver these and other services to users through their ATM and branch office networks. As payment volume expands over time or through cross-border consolidation of processing centers, average operating costs should fall as they have in the past. Where this reduction in average cost ends up—in lower loan rates, higher deposit rates, lower service prices, or higher profits--is an issue of some interest to competition authorities. However, there are difficulties in determining the level of banking or payment market competition, an issue discussed next.

3. An Indirect Assessment of Bank and Payment Market Competition.[19]

European banks are estimated to have saved some $32 billion in operating costs over 1987 to 1999 due to the realization of scale economies as payment volume expanded combined with the technology-associated shift from paper-based payments to cheaper electronic methods and away from expensive branches toward ATMs for cash acquisition (Humphrey, Willesson, Bergendahl, and Lindblom, 2006). Cost savings are also implied from the observed one-third reduction in the average ratio of operating costs to asset value over the last 18 years (Table 2). If European banking markets are reasonably competitive, cost reductions should be correlated over time with lower revenue flows from loan-deposit rate spreads and non-interest income. Indeed, this would be one way to assess the competitive efficiency of banking markets within and across countries.

3.1 Payment Competition: Among Instruments or Banks?

While payment instruments "compete" with each other for use at the point of sale or for bill payments, this competition concerns the inherent characteristics of the different payment instruments, such as safety, convenience, availability, etc. These characteristics are not priced and users choose those instruments that best meet their needs, such as using cash, cards, or checks at the point of sale or giro versus checks for bill payments. This is not the competition we wish to measure since neither regulators nor antitrust authorities have any real control over the availability or mix of these instrument-specific payment characteristics.

[18] In a recent interview (in Dialogue, Q4 2006, p.10), Ben Haasdijk – former chairman of Equens – stated "scale was the main driver for the merger that gave birth to Equens" and that "we will be passing on benefits of scale we achieve".

[19] Material in this section is drawn from Bolt and Humphrey (2008a).

When the focus is on suppliers of a specific payment instrument (such as credit cards) or where specific payment prices are largely transaction based (such as card interchange fees), it is possible to examine the competitive behavior of these suppliers and/or the prices they charge. However, such an investigation is necessarily partial and would not address the question of payment market competition more generally. A broader approach would examine payment competition in terms of providing cash, check, card, and giro payment services all together.

3.2 Two Problems in Applying Current Measures of Banking Competition to Payments.

Standard indicators of banking competition are: (a) the structure-conduct-performance paradigm which focuses on the degree of banking market concentration (usually a Herfindahl-Hirschman index of deposit market concentration); (b) the Lerner Index which is a price mark-up measure as in (price – marginal cost)/price; and (c) the H-statistic which indicates the extent that input cost changes are associated with changes in price. The Lerner Index and the H-statistic typically rely on cost and pricing data but payment costs are not separately reported and payment instrument use is not typically priced on a per transaction basis. While these measures are used to reflect the competitive behavior of the "entire" bank, due to limitations on price data they are more reflective of traditional loan/deposit rate spreads. There are two additional problems.

Problem 1: Non-Proportionality of Unit Payment Revenues With Payment Costs. Payment services are typically indirectly priced by banks using fixed monthly account maintenance fees, minimum balance requirements, low or no interest paid on deposits, account opening/closing charges, etc. The revenue they generate is not closely tied to payment volume. While a rise in payment volume will increase total payment operating costs and unit costs will fall due to strong scale economies in payment processing, total payment revenues need not rise proportionally with total payment operating costs.

If per transaction pricing was in place, then proportional changes in costs and revenues would be observed unless competition was limited and revenues expanded more than proportionally to underlying costs. Currently, banks could well respond to higher payment volume by raising indirect fees (so expanding total operating costs are fully/partially covered with higher revenues) even while unit payment costs are falling due to scale economies. As indirect fees are not closely related to variations in payment volume while unit and total operating expenses are, relating prices (indirect fees) to unit payment costs via a Lerner Index or H-statistic could well show that prices remain constant or rise while unit costs are falling, suggesting an uncompetitive market. Hence the need to focus on overall revenues and not on unit revenues (prices). Fortunately, information on bank revenues is often broader and more available than data on prices.

Problem 2: Inconsistencies Among Bank Competition Measures. While there is disagreement about which of the three measures noted above may "best" reflect market competition, the expectation is that since they purport to measure the same thing they are all positively correlated. Unfortunately, this expectation is not always met. These measures are almost unrelated when compared across European countries over time and can be negatively related within the same country over time. If there was a consensus as to which of the indicators is indeed "best", this inconsistency would be mitigated but this is not the case so choice among these measures can influence the outcome.

To illustrate: with data on 14 European countries over 1995-2001 covering 1,912 banks, the R^2 between the Lerner Index and the H-Statistic was only .06. Similarly, the R^2 between the HHI concentration measure (from the structure-conduct-performance paradigm) and the Lerner Index and H-statistic was, respectively, .09 and .05 (Carbó, Humphrey, Maudos, and Molyneux, 2007).[20] In addition, looking at each of the 14 countries separately over time, the relationship between the Lerner Index and the H-statistic was positive in only 8 out of 14 countries. The relationship between the HHI and these two measures was positive in only 8 and 5 countries, respectively. Since the choice of an existing banking competition measure may affect the results obtained, a different procedure where choice among these current measures is not necessary may prove useful.

3.3 Measuring Competition as Residual Revenues after Accounting for Costs.

In broad terms, banking revenues are a function of underlying input costs, differences in factor productivity, and the degree of competition in the market for banking services. Consequently, after "subtracting" the value of banking revenues across banks within a country or across countries associated with input costs and factor productivity (including scale economies), the remaining differences in revenues are likely associated with differences in competition. This approach is broader than the typical approach used in applications of the H-statistic or the Lerner Index in that it does not require information on unit costs or unit revenues (prices) which, for payment and other non-loan activities, are simply not available (or if available are indirect fees rather than transaction-based prices).[21]

Banks are the main providers of the full set of payment services. Indeed, it is estimated that about 25% of bank revenues are associated with payment services (McKinsey and Company, 2005). A similar analysis finds that 20% to 30% of the operating revenues of the 25 largest U.S. bank holding companies is associated with payment activities (Radecki, 1999). Unfortunately, information on bank-provided payment service revenues and costs are not reported separately (at least not in the public domain).

Publicly available data typically only separates banking service revenues into those associated with (a) the loan-deposit rate spread which reflects traditional banking activities, (b) securities interest revenue, and (c) non-interest revenues associated with all other activities which reflect payment activities (such as fees on deposit accounts and

[20] In this analysis, the H-statistic was multiplied by -1.0 so that a larger value of the H-statistic, the Lerner index, and the HHI would all indicate less competition.

[21] The limited availability of pricing data is why the Lerner Index and the H-statistic use computed average loan and deposit rates along with deposit/funding average or (estimated) marginal costs.

other indirect fees) as well as other, newer activities such as trading. This separates spread revenues (which do not involve payments) from other revenues (which do include payments) and enables one to focus somewhat more directly on payment activities, albeit not as directly as one would like (but better than looking at the variation of all revenues from all activities). Securities interest income is determined in reasonably competitive national or international markets and can be neglected.

3.4 A Competition Frontier.

There are at least four ways to determine a competition frontier from bank data on revenues, costs, and productivity. The approach used here is the composed error Distribution Free Approach or DFA (Berger, 1993).[22] This approach assumes that averaging each bank's residuals from the relationship illustrated in (3) and (4) across separate cross-section regressions reduces normally distributed error to minimal levels leaving only average inefficiency (or the average effect of competition on revenues).

In applying frontier analysis to the measurement of competition, it is maintained that the three most important determinants of non-interest income revenues and loan-deposit spread revenues are the underlying unit operating costs of producing these services, the productivity of the factor inputs used to produce these services, and the existing level of market competition. Two unit revenue functions are specified. One is the ratio of non-interest income (*NII*) to operating cost (*OC*) and reflects how income from priced services (payment transaction fees, ATM fees, deposit account maintenance fees, account switching fees, loan fees, loan commitment fees, etc., as well as trading income) varies with costs (*NII/OC*). A second revenue function reflects the ratio of revenues from the loan-deposit rate spread times the value of deposits (*SPREAD*) to operating cost (*SPREAD/OC*). Interest revenues not associated with the loan-deposit rate spread (e.g., securities income) are excluded since these rates of return are set in competitive national and international markets. Revenues from non-interest income across the same 11 countries used above was 20% of NII plus SPREAD revenues in 1987 but has grown to comprise 44% in 2006. SPREAD revenues fell from 80% to 56% over this period.

The variation of each dependent variable is a function of four indicators of unit costs: the average price of labor (*PL*), the opportunity cost of investing in physical capital (*PK*, the market interest rate), an index of the unit cost of processing payment transactions (*PC*), and a similar cost index reflecting ATM network scale economies (*ATMC*). The payment cost index is based on estimated country-specific payment scale economies shown in Table 2 above and changes in observed non-cash payment volume. Similarly, the ATM cost index used country-specific ATM scale economies estimates (Table 2) and changes in the number of ATMs. In addition, the variation in each dependent variable can be influenced by the use and productivity of labor in producing banking services as

[22] An alternative Stochastic Frontier Approach typically assumes a half-normal distribution for inefficiencies (or in our case competition inefficiencies) in order to separate unknown inefficiencies from normally distributed error in a panel regression. Two other approaches concern Data Envelopment Analysis (DEA) and Free Disposal Hull. These are linear programming approaches that assume error is zero but have the advantage that no functional form is imposed to fit the data.

indicated by the labor/deposit ratio (*L/DEP*), an indicator of the substitution of ATMs for more expensive branch offices to deliver cash to depositors (*ATM/DEP*), and an indicator of the variation in demand for banking services associated with the business cycle (*GAP* a measure of the GDP output gap). In summary, our two equation translog functional form model in logs is:

$$\ln(NII/OC) = \alpha_0 + \sum_{i=1}^{5}\alpha_i \ln X_i + 1/2\sum_{i=1}^{5}\sum_{j=1}^{5}\alpha_{ij} \ln X_i \ln X_j + \sum_{i=1}^{5}\sum_{k=1}^{2}\delta_{ik} \tag{3}$$

$$\ln X_i \ln P_k + \sum_{k=1}^{2}\beta_k \ln P_k + 1/2\sum_{k=1}^{2}\sum_{m=1}^{2}\beta_{km} \ln P_k \ln P_m + \ln \mathrm{e}_{\mathrm{NII}} + \ln \mathrm{u}_{\mathrm{NII}}$$

$$\ln(SPREAD/OC) = \theta_0 + \sum_{i=1}^{5}\theta_i \ln X_i + 1/2\sum_{i=1}^{5}\sum_{j=1}^{5}\theta_{ij} \ln X_i \ln X_j + \sum_{i=1}^{5}\sum_{k=1}^{2}\psi_{ik} \tag{4}$$

$$\ln X_i \ln P_k + \sum_{k=1}^{2}\phi_k \ln P_k + 1/2\sum_{k=1}^{2}\sum_{m=1}^{2}\phi_{km} \ln P_k \ln P_m + \ln \mathrm{e}_{\mathrm{SPREAD}} + \ln \mathrm{u}_{\mathrm{SPREAD}}$$

where: $X_{i,j} = PC, ATMC, L/DEP,\ ATM/DEP,\ GAP$; $P_{k,m} = PL,\ PK$ and have been defined above.

Equations (3) and (4) are related in that banks may choose to increase revenues over time (in response to higher costs or weaker competition) by raising the fees they charge on various banking services (affecting NII) or they can instead increase revenues by altering their loan-deposit rate spread (raising loan rates and/or lowering deposit rates). Industry observers have suggested that the faster growth in revenues from non-interest income activities and the slower growth in revenues from loan-deposit rate spreads is a response to expanded competition for loans and deposits which banks, in turn, attempt to offset by pricing and pricing higher more non-interest income activities. Some of these activities previously appeared to be "free" as their costs were largely bundled in loan and deposit rates. Since errors in explaining the variation of non-interest revenues in (3) may be correlated with errors in explaining the variation of revenues from the loan-deposit rate spread in (4), these two revenue equations are estimated jointly in a seemingly unrelated regressions (SUR) framework.

In a composed error framework, the regression relationship (3) can for illustration be re-expressed as:

$$\ln(NII/OC) = \mathrm{f}\ (\ln \text{Cost}, \ln \text{Productivity}) + \ln e + \ln u \tag{5}$$

The total residual (ln e + ln u) reflects the unexplained portion of the revenue dependent variable remaining after cost and productivity influences have been accounted for. Here ln e represents the value of random error while the maintained hypothesis is that ln u represents the effect of competition on revenues. The DFA concept relies on the assumption that ln e will average to a value close to zero while the average of ln u_i will reflect the average effect of competition (ln $\bar{u}_i$).[23]

The i^{th} bank or country with the lowest average residual (ln $\bar{u}_{min}$) is also the bank or country where the variation in underlying cost and productivity explains the greatest amount of the variation in revenues. This minimum value defines the competition frontier and the relative competition efficiency (CE_i) of all the other i banks or countries in the sample is determined by their dispersion from this frontier:

$$CE_i = exp\ (\ln \bar{u}_i - \ln \bar{u}_{min}) - 1 = (\bar{u}_i / \bar{u}_{min}) - 1 \qquad (6)$$

As the term u_i is multiplicative to the dependent variable in (5), in an unlogged equation (5) the ratio $(NII/OC)_i$ = R (Cost, Productivity)$_i$ u_i. Thus the ratio $\bar{u}_i / \bar{u}_{min}$ is an estimate of the ratio *NII/OC* for the i^{th} bank or country, for a given level of underlying cost and service productivity, to the value of the ratio $(NII/OC)_{min}$ for the bank or country facing the greatest competition and having the same underlying cost and service productivity.[24]

If $CE_i = .25$, then $\bar{u}_i$ is 25% larger than $\bar{u}_{min}$ so the unexplained portion of the revenue dependent variable in (5) is 25% larger than its minimum value at another bank or in another country. This difference reflects the unspecified influence of competition. Thus the larger is CE_i, the weaker is the ability of market competition to restrain revenues.[25]

A limitation is that (6) only indicates the relative level of competition: it can not determine the absolute level of competition even for the most competitive bank or country. Consequently, it is important to also examine the fit of the estimating equation since, if the R^2 is very high (e.g., .90 or above), the difference in relative competition measured by CE may not be very economically significant since the residuals $\bar{u}_i$ and $\bar{u}_{min}$ are themselves absolutely small.[26]

[23] Our frontier competition efficiency measure is similar in concept to an indicator developed by Boone (2008). Boone relies on balance sheet data to compute the difference between reported total firm revenues and reported total variable costs, a spread that contains total fixed cost plus extra revenues associated with the degree of market competition. We are interested in revenues for particular subsets of banking services (which are reported, but not in as much detail as we would like) but rely on statistical cost analysis to identify the (unreported) associated variable and fixed costs, leaving the effect of competition on revenues as an average residual.

[24] The ratio $\bar{u}_i / \bar{u}_{min} = [(NII/OC)_i/\text{R (Cost, Productivity)}_i]/[(NII/OC)_{min}/\text{R (Cost, Productivity)}_{min}]$ and when evaluated at the same mean level of underlying cost and service productivity, the predicted values of R (Cost, Productivity)$_i$ and R (Cost, Productivity)$_{min}$ are equal as both are at the same point on the estimated revenue curve, leaving the ratio $(NII/OC)_i/(NII/OC)_{min}$.

[25] The cost efficiency literature reports efficiency (EFF) and inefficiency (INEFF) values. If efficiency is 80% (EFF = .80), then inefficiency is INEFF = (1 - .80)/.80 = .25 or 25%. In (6), CE reflects the relative weakness of competition in restraining revenues and is equivalent to INEFF which reflects relative weakness of cost efficiency.

[26] This qualification is not well-understood in the frontier literature. Absolute differences in residuals need to be considered along with their relative size, so goodness of fit should be an additional consideration (Carbó, Humphrey, and Lopez del Paso, 2007).

3.5 Preliminary Indications of Cross-Country Bank/Payment Competition.

Aggregate, country-level data were collected from publicly available sources over 20 years (1987-2006) for 11 European countries. Three separate panel cross-section SUR estimations of each regression in (3) and (4) were made and the three sets of residuals were then averaged for each country separately. The resulting competition efficiency values (CE) are shown in the first two columns of Table 3 (along with their efficiency ranks, 1 being the most competitively efficient).[27]

Relatively high R² values were obtained and indicate that the seven operating cost and productivity variables explain 81% to 95% of the cross-country variation in bank non-interest and loan-deposit spread revenues. Another way to illustrate this is to compute the average percent that the residuals are of the two dependent variables for the three cross-section estimations. These values were 4%, 6%, 9% (three times), and 13%. Thus the share of the value of revenues not explained, and maintained here to reflect the effect of competition, is rather small. It is smaller still when one considers that revenues need to exceed costs by some degree in order to earn a positive return on invested capital or equity.

Table 3: Competition Efficiency for 11 European Countries

	CE_{NII}		CE_{SPREAD}	
U.K.	.000	(1)	.091	(11)
Spain	.028	(2)	.035	(9)
France	.036	(3)	.008	(2)
Norway	.042	(4)	.026	(5)
Netherlands	.046	(5)	.034	(8)
Denmark	.054	(6)	.032	(7)
Finland	.057	(7)	.030	(6)
Germany	.065	(8)	.016	(3)
Italy	.069	(9)	.036	(10)
Belgium	.075	(10)	.017	(4)
Sweden	.140	(11)	.000	(1)

In terms of determining competition efficiency, this means that $\bar{u}_i$ and $\bar{u}_{min}$ are both relatively small so the computed CE_i value is also low. Consequently, CE values from country-level averaged residuals in Table 3 suggest that differences in the apparent level

[27] In terms of model performance the Durbin-Watson statistics varied from 1.24 to 2.52. As the time-series dimension of the three panel data sets used in the estimation of (3) and (4) is less than two-thirds of the cross-section dimension, data stationarity is not of great importance but, when tested, unit roots could be rejected in 75% of the cases.

of competition efficiency across countries are small. Indeed, the difference in averaged residuals for non-interest income activities (CE_{NII}) between the U.K.--which defines the frontier--and Sweden--the country with the greatest apparent inefficiency--is only 14%. In contrast, Sweden defines the loan-deposit rate spread (CE_{SPREAD}) frontier while the U.K. has the greatest apparent inefficiency at 9%. By the standards of the cost efficiency literature where inefficiency is commonly found to be around 25%, competition inefficiency appears to be much smaller.

Table 4: Other Competition Measures for Same 11 Countries[28]

	H-Statistic		Profit/Revenue		CR-3	
U.K.	.76	(2)	.33	(6)	.51	(3)
Spain	.65	(4)	.40	(10)	.40	(1)
France	.61	(5)	.27	(5)	.57	(5)
Norway	.54	(8)	na		na	
Netherlands	1.01	(1)	.18	(2)	.88	(9)
Denmark	.22	(10)	.37	(7)	.75	(6)
Finland	.41	(9)	.39	(9)	.93	(10)
Germany	.70	(3)	.20	(3)	.55	(4)
Italy	.09	(11)	.22	(4)	.41	(2)
Belgium	.56	(7)	.11	(1)	.79	(8)
Sweden	.58	(6)	.38	(8)	.78	(7)

Using the same country ranking of Table 3, three other indicators of country-level banking competition are shown in Table 4. This concerns an H-statistic, a ratio of pre-tax bank profit to revenues for retail activities, and the share of retail banking income of the top three banks in estimated total retail income (CR-3, a revenue concentration ratio). While there is some consistency across these measures, it is sporadic. For example, Spain and Italy are the most competitive countries according to the revenue concentration ratio (CR-3) but the CE_{SPREAD} measure and the H-statistic both place Italy among the least competitive countries. Further, the profit/revenue ratio suggests that Spain (which has the highest ratio of retail profits to revenues) is ranked as the most uncompetitive but in terms of non-interest income activities (CE_{NII}), Spain appears to be very competitive. Overall, the R^2s between the H-statistic, the profit/revenue ratio, and the CR-3 revenue concentration measure range from .002 to .05 across the 11 countries. The R^2s are only slightly higher when these three measures are related to CE_{NII} or CE_{SPREAD} while the R^2 between these latter two measures is .53.

Although the differences in competitive efficiency measured here appear to be small across European countries, it is still of some interest to compare in Table 5 the rankings

[28] Sources: H-statistic (Bikker, J., L. Spierdijk, and P. Finnie, 2007, Table 2); Profit/Revenue and CR-3 (Economic Commission, 2007, Figures 5 and 1, respectively).

of the three most competitive countries with the three least competitive countries according to the five competition indicators. Overall, the most competitive countries would seem to include the U.K. (observed among the top ranked countries 3 out of 5 times) and Germany (also 3 out of 5).[29] The least competitive countries across all measures in Table 5 seem to be Italy and Finland (each ranked as least competitive 3 out of 5 times).[30] Both the U.K. and Germany remain the most competitive countries (and Italy and Finland the least competitive) even if an asset value HHI (European Central Bank, 2007) replaces the revenue concentration ratio (CR-3) in Table 5 (not shown).[31] These results for the U.K., Germany, and Italy are in general accord with conventional wisdom regarding country-level banking competition.

While CE_{NII} contains revenues from payment activities, specific information on payment-related revenues across banks and countries would be needed to make a more informed comparison concerning the competition efficiency of payment services. This would likely require changes in regulatory reporting procedures which are difficult to implement. Hence the emphasis here on indirect inference using CE_{NII}.

Table 5: Most and Least Competitive Countries

	CE_{NII}	CE_{SPREAD}	H-Statistic	Profit/ Revenue	CR-3
Most	U.K.	Sweden	Netherlands	Belgium	Spain
Competitive	Spain	France	U.K.	Netherlands	Italy
	France	Germany	Germany	Germany	U.K.
Least	Italy	Spain	Finland	Sweden	Belgium
Competitive	Belgium	Italy	Denmark	Finland	Netherlands
	Sweden	U.K.	Italy	Spain	Finland

4. Transaction Pricing of Payment Services.

The effect of changes in bank productivity, payment transaction and ATM scale economies, as well as labor and capital factor input costs across 11 countries on non-interest income and loan-deposit rate spread revenues (the two dependent variables) are illustrated in Table 6. The elasticities shown are averages across 11 countries for the entire 20 year period. In terms of productivity effects, the 6.6% average annual reduction in the labor/deposit ratio the countries experienced is associated with significant increases

[29] The Netherlands, France, and Spain were all ranked most competitive 2 out of 5 times.

[30] Sweden, Belgium, and Spain were all ranked least competitive 2 out of 5 times.

[31] An asset value HHI using Bankscope data, which does not include all banks, produces the same results.

in the ratios of both non-interest income and spread revenues to operating cost. However, the benefit of reducing the number of workers per value of deposits "produced" (similar to a lower input/output ratio) is to some degree offset by the negative elasticity for the price of labor, indicating that the 4.4% annual rise in unit labor cost for the 11 countries significantly lowers spread revenues. There is no significant labor price offset for non-interest income activities so the rise in labor productivity allows these revenues to be higher.

As noted earlier, marked differences in the size of branch offices across countries means that these offices are not comparable and could not be used as a variable in our analysis. Fortunately, an ATM in one country is quite similar to an ATM in other countries and the elasticity of the ATM/deposit ratio for spread revenues indicates that the 1.6% average annual growth in this "capital/output" ratio is associated with a relatively small reduction in spread revenues. This suggests the extra capital and maintenance costs of having "too many" ATMs per unit of deposits to attract and hold depositors by providing greater convenience can reduce spread revenues.

Table 6: Elasticities of the Cost Effect on Revenues

	Non-Interest Income Activity Revenues	Loan-Deposit Rate Spread Revenues
Productivity:		
Labor/Deposit ratio	-.42*	-.83*
ATM/Deposit ratio	.06	-.17*
Scale Economy:		
Payment Cost Index	.82*	-.02
ATM Cost Index	.06	-.24*
Factor Input Cost:		
Price of Labor	.01	-1.08*
Capital Opportunity Cost	-.53*	.70*

* Elasticity is significantly different from zero at p-value = .01.

Due to the existence of strong scale economies the reduction in unit payment cost is estimated to be 4.1% annually. Multiplied by the payment cost elasticity of .82 in Table 6, the reduction in unit payment cost is associated with a 3.4% yearly fall (not rise) in the ratio of non-interest activity revenues as a percent of operating cost. This seems odd since scale economies should reduce costs and presumably raise revenues. The explanation (as outlined in the Introduction) is that scale economies reduce unit costs while total payment operating expenses rise as payment volumes expand (rising by 6.1% annually).[32]

[32] This occurs when payment volume at each bank expands over time. However, when banks' payment processing centers are consolidated across borders (as envisioned by SEPA), a bank's given payment volume is combined with that of others so unit costs and total operating costs at the combined banks will both fall due to scale economies.

Total revenues could rise to offset the increase in total operating cost (the denominator in the non-interest income dependent variable) if payment transactions were directly priced but this is not common in Europe. Although more and more banking services are being priced and some transaction-based fees have been instituted, only Norway has implemented transaction-based pricing for all payment services. Thus a significant source of revenues is not being tapped, likely due to depositor views that payment services have been "free" and antitrust strictures against banks coordinating the implementation of pricing so as not to disrupt deposit market shares if only one or a few banks implemented transaction pricing.[33]

ATM networks also experience scale economies and unit ATM costs are estimated to fall by 5.5% annually. In contrast to non-cash payment services, banks do price some ATM cash withdrawals (usually those made by customers of other banks) and the expansion of ATM networks are viewed as an important element of strategic non-price competition to attract and hold depositors. Consequently, as ATM networks expand and unit costs fall, there is a significant rise in spread revenues as a percent of operating costs (offsetting the reduction in spread revenues associated with having a high ATM/deposit ratio).

The results of Table 6 support the proposition that, since banks do not generally price payment services on a per transaction basis, they may seek higher revenues to cover higher operating costs as payment volume expands by instituting and raising prices that are largely disconnected from changes in payment volume. Looked at in isolation, this suggests the exercise of market power (c.f., European Commission, 2007). However, such pricing behavior can also be associated with not tying prices to unit costs whereby revenues would rise “automatically” with volume growth as scale economies are realized.

In this regard transaction pricing would benefit banks by tying revenues more closely to costs so management obtains a clearer picture of where their profits are generated, permitting more efficient use of internal investment resources. Transaction pricing would also benefit consumers allowing them to balance better the cost of using different payment instruments with their assessment of the benefits, saving resources for the economy as a whole.

5. Conclusion.

We have demonstrated that strong scale economies are associated with bank payment activities as well as with the size of bank ATM and branch office networks. Indeed, relating bank operating cost to point of sale and bill payment transaction volumes across

[33] Norway coordinated the timing of when payment pricing would be implemented but there was no agreement on the price to be charged (and some banks did not initially price at all). The quid pro quo was the elimination of payment float--a benefit for depositors (Enge and Øwre, 2006). Pricing was efficient: it speeded up the adoption of lower cost electronic payments by 20% (Bolt, Humphrey, and Uittenbogaard, 2008).

11 European countries over 1987-2004 suggests that a doubling of payment volume increases operating expenses by only around 27%, so average payment costs could potentially fall by 37%. In an earlier analysis over a shorter period (1987-1999), bank operating expenses at European banks were estimated to be some $32 billion lower than they would otherwise have been due to the realization of payment scale economies, the shift from paper-based to cheaper electronic payments, and a corresponding shift away from expensive branch offices to ATMs for cash acquisition. These statistical estimates are seemingly supported by the observed fall in the ratio of bank operating cost to asset value in 11 European countries by 34% over the 1987-2004 period.

Considering the apparent reduction in bank operating expenses, it is of interest to see how competitive European banking markets may be since cost reductions are more likely to be passed on over time to users of banking services in a competitive market rather than mostly retained as higher profits. Unfortunately, the extant competition indicators that focus on market concentration, price mark-up over cost, or covariation of price with cost changes often give conflicting results. As found in a study covering 14 European countries over 1995-2001 covering 1,912 banks, the R^2 between the Lerner Index and the H-Statistic was only .06 while the R^2 between a HHI concentration measure and the Lerner Index and H-statistic was, respectively, .09 and .05. Even looking at each of these countries separately over time, the relationship between these three commonly used indicators was often negative when one would expect a positive correlation.

The apparent inconsistency among current indicators of market competition and our interest in trying to assess the competitiveness of bank payment activities, suggests that a different indicator may be useful. Bank revenue data is reported in more detail than pricing information and this permitted us to assess the degree of cross-country competition for three main service categories: traditional banking services associated with funding loans with deposits (loan-deposit rate spread), activities associated with non-interest income (which includes payment services), and securities trading activities which are important for liquidity. As securities trading takes place in competitive domestic and international markets, our revenue-based competition measures focus on the loan-deposit rate spread and non-interest income activities of the same 11 European countries over 1987-2006. While the procedure used to derive our cross-country competition measures borrows from the efficient frontier literature and estimates a competition frontier, the framework is not very different from the theoretically-based industrial organization approach of Boone (2008).

Our results suggest that cross-country differences in banking market competition are not large and that cost and productivity differences by themselves “explain” 84% to 95% of the variation in banking revenues for the six time periods our frontier model was estimated for our two measures during 1987-2006. This suggests that the large differences in individual banking service prices reported in a recent study (European Commission, 2007) do not seem to roll over into correspondingly large revenue differences, at least in the more aggregative loan-deposit rate spread and non-interest income activities the current data restricts us to.

It was also found that the estimated 4.1% annual reduction in unit payment cost is associated with a 3.4% yearly fall (not rise) in the ratio of non-interest activity revenues as a percent of operating cost while the expectation would be that realized scale economies lower unit costs and presumably raise revenues. This suggests that bank prices are not closely tied to unit costs where revenues would rise "automatically" as service/payment volume expands even as unit costs fall due to scale economies. Thus a significant source of revenues is not being tapped

In addition, efforts to raise bank prices to offset rising total operating costs when these prices are not closely tied to the volume of services being provided could well generate the wrong signal to regulatory authorities who may interpret the higher prices as the exercise of market power. More importantly, since unit payment costs have been falling, this alters the cost-benefit trade-offs needed to be observed and made by banks for profitably allocating internal investment funds and by consumers for making an informed choice among the banking services being offered. To respond efficiently to these changes, prices and revenues need to closely mirror unit costs and thus changes in volume. Finally, in countries where banks rely on the loan-deposit rate spread to cover a portion of their payment expenses, low interest rate environments—such as the one being currently experienced--reduce these margins but not payment expenses which can be more properly covered using transaction pricing.

Bibliography

Baumol, W., J. Panzar and R. Willig (1982). *Contestable Markets and the Theory of Industrial Structure.* Harcourt Brace Jovanovich: San Diego.

Beijnen, C., and W. Bolt (2009): “Size Matters: Economies of Scale in European Payments Processing”, *Journal of Banking and Finance*, 33: 203-10.

Berger, A. (1993): "'Distribution Free' Estimates of Efficiency in the US Banking Industry and Tests of the Standard Distributional Assumptions", *Journal of Productivity Analysis*, 4: 261-292.

Bikker, J., L. Spierdijk, and P. Finnie (2007): "The Impact of Market Structure, Contestability and Institutional Environment on Banking Competition", DNB Working Paper No. 156, November.

Bolt, W., and D. Humphrey (2008a): "Bank Competition Efficiency in Europe: A Frontier Approach", De Nederlandsche Bank working paper No. 194/2008, December.

Bolt, W., and D. Humphrey (2008b): “Scale Economies from Individual Bank Payment Data,” Working Paper, Florida State University.

Bolt, W., and D. Humphrey (2007): "Payment Network Scale Economies, SEPA, and Cash Replacement", *Review of Network Economics*, 6: 453-473.

Bolt, W., D. Humphrey, and R. Uittenbogaard (2007): "Transaction Pricing and the Adoption of Electronic Payments: A Cross-Country Comparison", *International Journal of Central Banking*, 4: 89-123.

Boone, J. (2008): “A New Way to Measure Competition”, *Economic Journal*, 118: 1245-1261.

Brits, H. and C. Winder (2005): “Payments Are No Free Lunch,” De Nederlandsche Bank, Occasional Studies, Vol. 3/Nr. 2 (preliminary version published as: De Nederlandsche Bank (2004). *The Costs of Payments: Survey on the Costs Involved in POS Payment Products*, Working Group on Costs of POS Payment Products, March).

Carbó, S., D. Humphrey, and R. Lopez del Paso (2007): "Opening the Black Box: Finding the Source of Cost Inefficiency", *Journal of Productivity Analysis*, 27: 209-220.

Carbó, S., D. Humphrey, J. Maudos, and P. Molyneux (2007): "Cross-Country Comparisons of Competition and Pricing Power in European Banking", *Journal of International Money and Finance*, forthcoming.

Enge, A., and G. Øwre (2006): "A Retrospective on the Introduction of Prices in the Norwegian Payment System", *Norges Bank Economic Bulletin,* 77: 162-172.

European Central Bank (2007): *EU Banking Structures*. October.

European Commission, (2007): *Report of the Retail Banking Inquiry*. Commission Staff Working Document, SEC (2007) 106, January 31.

First Annapolis Consulting (2006). *2005 Issuer Cost of Payments Study* (data used with permission from First Data Corporation).

Gresvik, O. and G. Øwre (2002): "Banks' Costs and Income in the Payment System in 2001," *Norges Bank Economic Bulletin*, 73: 125-33.

Humphrey, D., M. Willesson, G. Bergendahl and T. Lindblom (2006): "Benefits from a Changing Payment Technology in European Banking," *Journal of Banking and Finance*, 30: 1631-52.

McAllister, P., and D. McManus (1993): "Resolving the Scale Efficiency Puzzle in Banking", *Journal of Banking and Finance,* 17, 389-405.

McKinsey and Company (2005). European Payment Profit Pool Analysis. (Cited by G. Tumpel-Gugerell, speech at the 3rd International Payments Summit, Milan, October 27, 2008.)

Quaden, G. (2005). *Costs, Advantages and Drawbacks of the Various Means of Payment*, National Bank of Belgium, English Summary, December: 41-7.

Radecki, L. (1999): "Banks' Payment-Driven Revenues", Federal Reserve Bank of New York Economic Policy Review, July: 53-70.