Abstract

Debit cards in Europe replace many medium value cash payments while stored value cards were designed to replace small value cash transactions. Unit cost considerations dictated a two-card (or single card) dual technology approach to cash replacement since unit costs were too high to use debit cards alone for low value transactions. We estimate payment scale economies and illustrate the change in processing volume likely needed to permit debit cards to cost-effectively replace a greater portion of smaller value cash transactions, allowing consumers and merchants to use increasingly a single cash replacement technology. (93 words)

Key Words: Debit cards, scale economies, cash replacement

JEL Classification Code: E41, C53

* This paper was presented at the 9th DNB Research Conference on "Recent Developments in Payment Economics", at the Nederlandsche Bank in Amsterdam, November 9-10, 2006. Comments by Santiago Carbó-Valverde, Pierre Lafourcade, and Leo Van Hove are gratefully acknowledged. We also thank Peter Keus for excellent research assistance.

Corresponding Author:
David Humphrey, Department of Finance, Florida State University, Tallahassee, FL 32306-1042; (850)-644-7899; dhumphrey@cob.fsu.edu
1 Introduction

Electronic payments are typically cheaper to produce and accept than paper-based instruments. For bill payments, this holds for both banks producing electronic giro payments and billers accepting them. At the point of sale, however, while electronic payments are cheaper for banks to produce than their paper-based substitutes (checks and cash over the counter), merchants often find it cheaper to accept cash, especially for smaller value transactions (Brits and Winder, 2005, Table 4.3; Garcia-Swartz, Hahn, and Layne-Farrar, 2006, Tables 2 and 3). Debit cards have largely replaced checks in many European countries (with France and the U.K. being the exceptions) and they continue to replace cash for medium value transactions. However, debit card costs have hindered their use for the replacement of cash for small value payments. This led banks and other suppliers to offer potentially lower cost stored value cards for small value transactions, a solution that required consumers and merchants or just merchants to adopt two different payment technologies.\footnote{The experience of the Netherlands is instructive. In the mid-1990s, commercial banks issued stored value cards separate from debit cards. Postal banks also issued a separate stored value card but it was incompatible with the commercial bank card. Merchants wishing to accept all cards had to have three terminals. Around 2001, commercial banks issued a single card that could handle both debit and stored value transactions. After commercial and postal bank stored value cards were made compatible, a single terminal was developed that could handle both types of transactions, although merchants were charged an extra fee.}

Consumer adoption and use of stored value cards seems stalled at a relatively low level of market penetration (Van Hove, 2006). Although data are incomplete, stored value cards account for only €1.2 billion in payments across 11 European countries in 2004. In contrast, the value of debit card transactions is estimated to be €1,146 billion while the value of cash withdrawals (a proxy for cash use) was €2,189 billion.\footnote{Debit and credit card data are combined here since separate data are often unavailable and credit card values are typically very small compared to the US. Cash use is understated as it is mostly the value of ATM cash withdrawals, although a few countries report cash withdrawn over the counter at financial institutions (but only 1 country reported "cash-back" at the point of sale).} Overall, card payments comprised 34% of the total, cash withdrawals accounted for 66%, while stored value payments were only 0.04%. As stored value payments are used to replace small value cash transactions, and thus would be expected to be a small portion of overall point of sale payments, their current small share is almost entirely due to their low level of market penetration.\footnote{Using a different measure, stored value transactions accounted for 1.2% of cash with-}
An important drawback of stored value cards is that consumers may have to carry two cards to replace cash—a debit card plus a stored value card—and the latter requires "filling" at terminals while the former does not. While convenience is enhanced if both technologies are on a single card and if merchants have a single terminal that can handle both types of transactions, banks often charge an extra fee to handle a stored value transaction. In 2002, the average total (fixed plus variable) bank plus merchant cost of a cash transaction at the point of sale in the Netherlands was € .30 while a debit card transaction was € .49 and that of a stored value transaction was € .93 (Brits and Winder, 2005, Table 4.3). The hope was that stored value transaction volume would rapidly expand and substantially lower average fixed costs since average variable costs for stored value transactions are the lowest of the three at € .033 per transaction versus € .176 for cash and € .197 for debit cards. This has not happened. As debit cards are the more mature product, already have a significant market penetration, and do not require consumers to access terminals to "refill" them, if debit card costs can be lowered sufficiently they could further reduce cash use and replace stored value cards for small value transactions.

As the replacement of cash by debit cards for smaller value transactions is importantly influenced by unit costs and unit costs are largely dependent on transaction volume, our goal is to try to determine payment scale economies in Europe, especially for debit cards. Estimates of scale economies, when combined with expected transaction growth within a country, or the consolidation of card processing operations across countries, permit future card unit costs to be approximated and the likelihood of debit cards replacing small value cash transactions assessed. Payment scale economies are considered to be the main economic driver behind the creation of a Single European Payment Area (SEPA), which entails the harmonization and standardization of retail payment instruments (especially payment cards, direct debits, and credit transfers) across the euro zone. SEPA aims to improve the efficiency of cross-border payments and "to develop common instruments, standards, and infrastructures in order to foster substantial economies of scale" (ECB 4th progress report, 2006).

drawal transactions in the Netherlands (Brits and Winder, 2005, Table 4.1) and 3.6% of the estimated number of cash transactions in Belgium (not cash withdrawals—Quaden, 2005, Table 2).

Credit card costs were € 3.59. Belgium carried out a similar payment cost study and, likely due to different transaction volumes, found that the total average cost of a POS transaction in 2003 was € 0.53 for cash, € 0.55 for a debit card, and € 0.54 for a stored value card. Credit card costs were € 2.62 per transaction (Quaden, 2005, Table 3).
Some payment scale economy estimates are available from Federal Reserve data on check processing, automated clearing house transactions, and wire transfers but none exist for cards as these payments are processed by private sector firms. Even if card scale estimates were available, the size of the U.S. market for card payments is much larger than that for individual European countries. As scale economies are typically volume dependent due to high fixed costs, the U.S. results would be interesting but perhaps of limited value for Europe. Consequently, we estimate payment scale economies using European data: first from public payment cost data for Norway over a short time period; second from recent point estimates of payment costs in the Netherlands and Belgium; and third using a panel of payment and banking data for 11 European countries over 18 years. In this last approach, both standard and mixed estimation procedures are used.

In what follows, in Section 2 we present information on card scale economies for three countries (Norway, the Netherlands, and Belgium). Data and estimation issues are discussed and our cross-country cost function model is specified in Section 3, along with an application of mixed estimation. Scale economy results for both point of sale and bill payment transactions are reported for each of 11 European countries in Section 4. These results are paired with approximate unit cost estimates of cash, debit card, and stored value card payments in Section 5 to determine the possible timeframe and context for debit card costs to fall sufficiently to be able to replace a portion of small value cash and stored value transactions. The timeframe differs depending on how debit card volume expands—overtime as a country’s internal market expands keeping the current set of processing centers or through consolidation of processing centers across countries. Our conclusions are presented in Section 6.

2 Payment Scale Economies: Three Countries

2.1 Norway

The total estimated bank-incurred cost and number of EFTPOS transactions for Norway (due almost exclusively to debit cards) are shown in Table 1, along with average cost per transaction. The 47% reduction in average card costs between 1988 and 1994 and the 44% reduction between 1994 and 2001 are substantial and imply scale economies of .49 and .43 (computed from (percent change in cost)/(percent change in volume)). However, the total and average cost figures are in nominal terms and the cost of living index in Norway rose by 19% and 18%, respectively, in these two periods so the
true scale values are likely lower, indicating greater scale benefits. This is because scale economies should be a function of the number of transactions, holding input prices constant.\textsuperscript{5}

Table 1: Norway: Bank Card Costs, Volume, and Scale Economy

<table>
<thead>
<tr>
<th></th>
<th>1988</th>
<th>1994</th>
<th>2001</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFTPOS Total Card Costs\textsuperscript{1}</td>
<td>NOK 51.85</td>
<td>NOK 325.35</td>
<td>NOK 822.75</td>
</tr>
<tr>
<td>EFTPOS Card Transactions\textsuperscript{1}</td>
<td>6.1</td>
<td>72.3</td>
<td>329.1</td>
</tr>
<tr>
<td>Average Cost per Transaction</td>
<td>NOK 8.50</td>
<td>NOK 4.50</td>
<td>NOK 2.50</td>
</tr>
</tbody>
</table>

Implied Card Scale Economy


\textsuperscript{1}Costs and transactions are in millions.
Sources: Gresvik and Øwre (2002) and Grete Øwre.

If we knew the share of labor expenses in total bank card costs and assumed (as is likely) that labor expenses rose in parallel with the rise in the cost of living index, we could then subtract the estimated rise in labor costs from total card costs and derive a more accurate scale estimate. In principle, this should be done for capital and material inputs as well. Although this is not possible with the data at hand, we can conclude that card scale economies in Norway are greater than those shown in Table 1 and the true scale statistics are less than .49 or .43. A complicating factor is that scale economies can vary as processing volumes become large, lowering average fixed costs, and Norway with a population of 4.6 million will likely experience a different scale economy than larger European countries with greater volume.

2.2 Netherlands and Belgium

Two recent detailed analyses of payments in the Netherlands (Brits and Winder, 2005) and Belgium (Quadern, 2005) have developed estimates of bank and merchant costs of providing and accepting different payment instruments. The average total cost and incremental variable cost of using cash, a debit card, and a stored value card at the point of sale are presented in Table 2. As economies of scale can be expressed as the ratio of marginal to average cost, the cost figures in the table provide an implied estimate of

\textsuperscript{5}Alternatively, if the goal is to illustrate how all card costs have changed over time in order to approximate the trend in future costs and perhaps prices, then all nominal costs should be included. In either case, bank card costs were determined using activity-based costing which allocates direct as well as joint costs (Gresvik and Øwre, 2003).
scale economies for these three payment instruments. The implied scale economy for cash ranges between .37 in the Netherlands to .25 in Belgium. The implied scale economy for debit cards in both countries is .39 while that for stored value cards is quite low at .04 to .16, indicating that average fixed cost here is quite high as transaction volume is low.

It is perhaps reassuring that the implied scale economy estimates for three countries (Norway, Netherlands, and Belgium) are not too dissimilar. Even so, these three countries are small relative to Germany, France, and the U.K.—the other countries we cover below—and our cross-country econometric scale estimates may well differ. As well, since the bank portion of payment costs likely includes some allocated expense from maintaining a branch network, as opposed to the back office processing of payments, scale economies associated only with payment processing (holding branches constant) are likely to be different. This is not an issue if these scale values were used to approximate future unit costs as payment volume expands within a given country (since branch and ATM networks may change as well). It would be something to consider if the way that volume grows and unit cost falls is through the establishment of consolidated processing centers among countries which could occur over a much shorter time period.

Table 2: Total Average Cost, Incremental Cost, and Implied Scale Economies: the Netherlands and Belgium

<table>
<thead>
<tr>
<th></th>
<th>Cash</th>
<th>Debit Card</th>
<th>Stored Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netherlands, 2002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Total Cost</td>
<td>.300</td>
<td>.489</td>
<td>.931</td>
</tr>
<tr>
<td>Incremental Cost</td>
<td>.112</td>
<td>.190</td>
<td>.033</td>
</tr>
<tr>
<td>Implied Scale Economies</td>
<td>.37</td>
<td>.39</td>
<td>.04</td>
</tr>
<tr>
<td>Belgium, 2003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Total Cost</td>
<td>.53</td>
<td>.55</td>
<td>.54</td>
</tr>
<tr>
<td>Incremental Cost</td>
<td>.133</td>
<td>.214</td>
<td>.084</td>
</tr>
<tr>
<td>Implied Scale Economies</td>
<td>.25</td>
<td>.39</td>
<td>.16</td>
</tr>
</tbody>
</table>

Sources: Brits and Winder (2005), Table 4.3; Quaden (2005), Table 3.

\[ \text{Scale economies (SCE) equal } \left( \frac{\partial OC}{\partial Q} \right) \bigg/ \left( \frac{\partial Q}{Q} \right) \text{ where } OC \text{ is operating cost and } Q \text{ is a measure of output. Thus } \text{SCE} = \left( \frac{\partial OC}{\partial Q} \right) \bigg/ \left( \frac{OC}{Q} \right) = \text{marginal cost divided by average cost, where incremental cost in Table 3 approximates marginal cost.} \]
3 Payment Scale Economies: A Cross-Country Framework

Determining scale economies for financial institutions 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 are 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 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.

We adopt a different approach which we feel can provide more solidly-based scale economy estimates. Specifically, we relate bank operating (not total) costs to measurable physical characteristics of banking output associated with payment processing and service delivery levels and mix. In this manner we focus 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 allows us to determine 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 from which scale economies may be approximated.\(^7\) In this regard our approach represents an alternative and more specific way to identify the likely effect on costs from technical change in banking.\(^8\) As point of sale and bill payment transactions are jointly processed in the deposit

\(^7\) The provision of deposit and loan services not directly associated with the number of payments or ATMs should be associated with the number of banking branches in a country. While this is similar to the common assumption made in the literature where the stock of the value of deposits, loans, or assets is assumed to reflect the underlying flow of associated banking services, it is necessitated by a lack of data. The R\(^2\) between the number of (standardized) branch offices and the value of total banking assets across our 11 countries and over time is .74.

\(^8\) To circumvent the impossibility of separating scale effects from technical change with time-series data, it has been common practice to use panel data so that the cross-section component identifies scale while the time-series component identifies technical change. Our panel data set is composed of 18 years of cross-section observations across 11 countries.
accounting function while aspects of service delivery are jointly produced via branches and ATMs, these two activities can be considered functionally separable.\footnote{About the only interaction would be consumers and businesses depositing (a declining number of) checks at a branch office and perhaps, on a one-time basis, filling out documents to pay recurring bills by electronic giro or applying for a debit/credit card. After establishing a giro account, bill payments occur automatically, as do all card payments, without branch or ATM intervention.}

The limitations of this approach are that our scale and technical change results are based on historical data so the most recent, and likely most productive, technologies need not be represented. Indeed, our long time period covers payment technologies that initially were partially manual but end up being close to fully electronic (Leinonen, 2004). As a result, we will likely understate cost reductions available with the very newest technology. Finally, costs that are statistically allocated to bank branch operations will cover a level and mix of non-payment services that will differ across countries (so branch scale effects should be viewed with caution).

### 3.1 Cross Country Data

Payment volume (or value) data either do not exist for individual banks or are not publicly available (Norway is the only exception). However, annual aggregate payment information by country is available from European Central Bank and European Monetary Institute ("Blue Book") as well as Bank for International Settlements ("Red Book") publications. Thus our data set consists of country-level panel data.\footnote{Data were originally collected by Magnus Willesson and used in an earlier study (Humphrey, Willesson, Bergendahl, and Lindblom, 2006). The data are updated to 2004 and adjusted as revised data for earlier periods has become available.}

Important trends in these data are illustrated in Table 3.\footnote{Due to data collection difficulties, largely solved today, in many countries initial information about certain payments was of necessity an educated guess. In early editions of the two data sources just cited, some countries reported the total volume of giro transactions but were unable to divide this into their paper and electronic components. Recent publications now provide such a division but the distinction between paper and electronic only goes back a few years. In some cases, we attempted an allocation ourselves based on earlier fragmentary data. Even so, it is safer to combine paper and electronic giro data as we do in our analysis.} As the euro did not exist for most of our time period, all value data are in U.S. dollars using purchasing power parity exchange rates (International Monetary Fund, 2006). The countries in Table 3 are ranked according to the value of their banking sector’s total operating cost in 2004 (Column 1—the ranking
is similar for 1987). The percent change in the ratio of operating cost (OC) to total assets (TA) between 1987 and 2004 is shown in Column 2 and is suggestive of how the cost of providing banking services has fallen over time (e.g., an apparent 40% fall for Germany but no reduction for France).\textsuperscript{12} For all countries together, shown at the bottom of the table, the ratio of operating cost to total assets fell by a third (-34%).\textsuperscript{13} At the same time, point of sale transactions rose by 140%, bill payments by 151%, ATMs by 434%, but branches only by 9.8%. One thread tying these changes together is that ATMs, compared to branches, are a more cost-efficient way to deliver cash to depositors while debit cards reduce the need to provide cash in the first place. Both developments reduce bank costs.

Generally, those countries with the largest reduction in the number of branches experienced the largest reductions in unit operating cost. Italy is an exception since it expanded its branches by 133% but its unit operating cost still fell by 29%. If Italy was excluded from the analysis, the number of branches in the remaining 10 countries would have fallen by 2% between 1987 to 2004 rather than rising by 9.8%. As we note below, large differences in the growth of payment transactions versus ATMs versus branches will influence our interpretation of scale economy results.

### 3.2 Estimation Issues

There are at least three estimation issues that need to be addressed in determining payment scale economies. First, cost function specifications used in these analyses—translog or Fourier—are sensitive to the size of an independent variable’s share in the cost of the dependent variable (operating cost). Variables with small cost shares are associated with unreliable results, a pattern well-documented in scope economy studies (Röller, 1990; Pulley and Humphrey, 1993) but the problem applies to scale analyses as well. This is of particular interest in our case since in the early part of our time period the volume of electronic giro payments in many countries and

\textsuperscript{12} For an individual bank, the denominator of the OC/TA ratio can expand via purchased funds and generate a smaller rise in operating costs than if an asset expansion was funded by produced deposits using branches. As our data includes the operating cost of those banks producing the deposits that are sold as purchased funds to other banks, this potential bias does not exist. However, a rise in the share of assets held as securities as a bank grows (lowering the loan share) will lower the OC/TA ratio. This is a common occurrence over the business cycle and typically reverses itself over time.

\textsuperscript{13} This reflects the sum of OC across all 11 countries divided by the sum of TA for 1987 vs. the same ratio of sums for 2004. It is not a simple average of separate country ratios (which would be -38%).
Table 3: Percent Changes for 11 European Countries Between 1987 to 2004

<table>
<thead>
<tr>
<th>Country</th>
<th>Operating Cost (2004, Mil PPP)</th>
<th>OC/TA</th>
<th>Point of Sale</th>
<th>Bill Payment</th>
<th>ATMs</th>
<th>Branches</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>$82,850</td>
<td>.02%</td>
<td>78%</td>
<td>185%</td>
<td>280%</td>
<td>1.4%</td>
</tr>
<tr>
<td>Germany</td>
<td>77,247</td>
<td>-40</td>
<td>501</td>
<td>115</td>
<td>601</td>
<td>14</td>
</tr>
<tr>
<td>U.K.</td>
<td>63,972</td>
<td>-52</td>
<td>117</td>
<td>214</td>
<td>160</td>
<td>-25</td>
</tr>
<tr>
<td>Italy</td>
<td>50,204</td>
<td>-29</td>
<td>121</td>
<td>117</td>
<td>809</td>
<td>133</td>
</tr>
<tr>
<td>Netherlands</td>
<td>34,157</td>
<td>-33</td>
<td>330</td>
<td>128</td>
<td>1,593</td>
<td>-50</td>
</tr>
<tr>
<td>Spain</td>
<td>32,120</td>
<td>-23</td>
<td>714</td>
<td>390</td>
<td>858</td>
<td>22</td>
</tr>
<tr>
<td>Belgium</td>
<td>12,070</td>
<td>-50</td>
<td>136</td>
<td>98</td>
<td>802</td>
<td>-48</td>
</tr>
<tr>
<td>Sweden</td>
<td>5,637</td>
<td>-38</td>
<td>685</td>
<td>8</td>
<td>70</td>
<td>-33</td>
</tr>
<tr>
<td>Denmark</td>
<td>4,112</td>
<td>-39</td>
<td>206</td>
<td>333</td>
<td>522</td>
<td>-38</td>
</tr>
<tr>
<td>Finland</td>
<td>2,783</td>
<td>-59</td>
<td>1,057</td>
<td>136</td>
<td>11</td>
<td>-46</td>
</tr>
<tr>
<td>Norway</td>
<td>2,160</td>
<td>-60</td>
<td>757</td>
<td>67</td>
<td>70</td>
<td>-38</td>
</tr>
</tbody>
</table>

All Countries Together: -34% 140% 151% 434% 9.8%

Note: All column values, except those for Column 1, reflect percent changes.

electronic card transactions in others were very small and generated little cost. As well, toward the end of our time period, the volume of checks and paper giro transactions are zero or close to it in a number of countries.\textsuperscript{14} Even a composite cost function (c.f., Pulley and Braunstein, 1992) which is better able to deal with this issue, since the data are not logged as in the translog or Fourier forms, experiences difficulties here.

Instead of specifying four separate payment categories where some transaction volumes are close to or at zero at different points of time, we sum card transactions with checks to represent point-of-sale transactions and aggregate paper and electronic giro transactions to reflect bill payments. This addresses the low cost share issue but at the expense of less specificity in the scale economy estimates since the result is effectively a weighted average of the scale effects of the underlying composition of transactions. As we are primarily interested in card payment scale economies, the fact that debit cards accounted for 90% of non-cash point of sale transactions over 2002-04 suggests that the point of sale scale economy estimate for the latter half of our time period may be a good approximation for debit cards. Similarly, since 86% of bill payments over 2002-04 were electronic giro transactions, the bill payment elasticity for the latter half of our time period should be

\textsuperscript{14}The Baumol, Panzar, and Willig (1982) average incremental cost measure derived from a multiproduct cost function (Chapter 4) also suffers from this problem as it is based on the evaluation of an estimated function at or close to the point where the production of an output is zero.
close to that for electronic giro payments.

A second estimation issue concerns how to incorporate some limited prior information on payment scale economies. The fact that production functions involved in processing payments are very similar across countries suggests that prior information could be a useful adjunct to statistical estimation. We incorporate limited prior information on payment scale economies using a mixed estimation procedure to determine the "strength" or robustness of scale estimates based only on our cross-country data.

A third estimation issue concerns autocorrelation in the time-series component of our panel data set as our payment, ATM, and branch data are strongly trended. Indeed, three out of seven of our specified variables are non-stationary. Since our variables are not cointegrated, we investigate the effect on our results from (a) adding time-specific indicator variables (one for each time period) as well as (b) transforming the data using a differencing parameter \( \rho \) estimated from the residuals from equations (1) and (2) below.

### 3.3 Cost Function Specification

The variation of operating cost (\( OC \)) across 11 European countries annually over 1987-2004 is used in a translog and also a Fourier cost function to derive scale economies for point of sale transactions, bill payments, as well as for ATM and branch office networks. The first two lines of the cost function (1) represents the translog form while the whole equation is the Fourier form. The cost function (1) is estimated jointly with cost share equations (2) for labour:

\[ \text{stationary variables are the logs of the number of card transactions, the number of ATMs, and the two input prices, which are specified in the following section. Non-stationary variables are logs of operating costs, the number of giro transactions, and the number of branch offices.} \]
\[
\ln OC = \alpha_0 + \sum_{i=1}^{4} \alpha_i \ln Q_i + 1/2 \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_{ij} \ln Q_i \ln Q_j + \sum_{i=1}^{4} \sum_{k=1}^{2} \delta_{ik} \\
\ln Q_i \ln P_k + \sum_{k=1}^{2} \beta_k \ln P_k + 1/2 \sum_{k=1}^{2} \sum_{m=1}^{2} \beta_{km} \ln P_k \ln P_m \\
+ \sum_{n=1}^{4} [\tau_n \cos(\ln Q^*_n) + \omega_n \sin(\ln Q^*_n)] \\
+ \sum_{n=1}^{4} \sum_{q=n}^{4} [\tau_{nq} \cos(\ln Q^*_n + \ln Q^*_q) + \omega_{nq} \sin(\ln Q^*_n + \ln Q^*_q)] \\
+ \sum_{n=1}^{4} [\tau_{nnn} \cos(\ln Q^*_n + \ln Q^*_n + \ln Q^*_n) \\
+ \omega_{nnn} \sin(\ln Q^*_n + \ln Q^*_n + \ln Q^*_n)] \\
S_k = \beta_k + \sum_{m=1}^{2} \beta_{km} \ln P_m + \sum_{i=1}^{4} \delta_{ik} \ln Q_i
\]  

where:

\( OC \) = total operating cost, composed of all labor, materials, outsourcing, and capital consumption costs (but no interest expenses);\(^\text{16}\)

\( Q_i \) = 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 (\( ATM \))\(^\text{17}\), and size standardized branch offices (\( BR^{STD} \)).\(^\text{18}\) The standard-

\(^{16}\)As is usual, fixed costs are reported as capital consumption expenses rather than the current value of all fixed assets. Depending on ATM ownership and access agreements, in some countries OC will include the capital expense of a bank’s own ATMs plus costs incurred by their customers who use other banks’ ATMs. This generates some unknown amount of double counting since the revenue received does not reduce the reported cost of the other banks’ ATMs.

\(^{17}\)The number of ATMs in a country seems to be partly influenced by the number of ATM networks (Snellman and Viren, 2006). This variable is not consistently reported for our 11 countries and, when available, is typically constant over time and so would reflect country-specific differences. When 11 country-specific dummies were added to our model (below), there was little change in our scale economy results as our cross-country variables already reflect country-specific influences.

\(^{18}\)EFTPOS terminal availability is associated with the volume of electronic card payments—a variable we already use—and thus is not separately specified in the model.
The procedure is explained below:

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

\[ S_k = \text{the cost share for the labor input (the "mirror image" capital/materials input cost share is deleted to avoid singularity).} \]

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. The Fourier form we use adds sin and cos terms to the translog cost function shown in the first two lines in the equation. As our main concern is to allow for greater flexibility in the local identification of output effects on operating costs, the sin and cos terms are applied to our two payment and two service delivery measures (Q). The Fourier form is a globally flexible approximation since the respective sin and cos terms are mutually orthogonal over the [0, 2\pi] interval.

### 3.4 Comparing Branch Offices Across Countries

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. Although the average number of workers per branch office across our 11 countries was 15.9 it

---

19 No information is available for banks’ weighted average cost of capital (WACC) or a relevant risk premium on equity. Even though the nominal market interest rate should be lower than WACC, it is unlikely to deviate too much from it as banks have high leverage, a low beta value, and often a large share of low or zero interest deposits. The alternative of specifying a real interest rate had little effect on the payment scale economy values (changing them by 1 to 2 percentage points).

20 The terms in lines three to six in (1) are \( \ln Q^* = \ln Q \cdot YQ + ZQ, YQ = (1.8 \cdot 2\pi)/(\max \ln Q - \min \ln Q), ZQ = .2\pi - \min \ln Q \cdot YQ, \) and \( \pi = 3.141593 \ldots \), so that \( \ln Q^* \) is essentially expressed in radians (Mitchell and Onvural, 1996; Berger and Mester, 1997). Our Fourier specification follows Berger and Mester.

21 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 cash.
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. It is thus necessary to standardize branch office size according to some benchmark to make them more comparable across countries. 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 average of 16.1 workers per office over our 18-year period (close to our overall mean), was selected as the branch benchmark and other countries were adjusted accordingly. This adjustment, while necessary given the heterogeneity of the cross-country branch data, is a weakness in our study.

3.5 Mixed Estimation

Payment production functions are quite similar across countries. The processing of card, check, or giro payments in Europe require the same types of machines, labor skills, software, computers, and telecommunications that are used to process card, check, or automated clearing house (ACH) payments elsewhere. Thus payment scale economies should be similar as well, if processing center transaction volumes are not too different. As seen above,

\[ BR^{STD} = BR \left( \frac{L}{BR} \right) / 16.1, \]

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.

There is no improvement if a different adjustment were adopted. In related work, results using a deposit/branch or total asset/branch ratio as an alternative standardization metric produced unrealistic scale economy values (Humphrey, Willesson, Bergendahl, and Lindblom, 2006).
we have information on card payment scale economies for Norway (.43 over 1994-2001) and the Netherlands and Belgium (.39 during 2002-3). There also exist scale economy estimates for U.S. Federal Reserve check processing (averaging .74 across different models) and automated clearing house operations (.48) from Bauer (2002) and Bauer and Ferrier (1996), respectively. This prior information can be a useful adjunct to our cross-country statistical estimation and forms the basis for mixed estimation.26

The scale economy for the $i$th payment service ($SCE_i$) derived from the Fourier cost function (1) is:

$$\frac{\partial \ln OC}{\partial \ln Q_i} = \alpha_i + 1/2 \sum_{j=1}^{4} \alpha_{ij} \ln Q_j + \sum_{k=1}^{2} \delta_{ik} \ln P_k + \frac{\partial FT}{\partial \ln Q_i}$$

where, for simplicity, $FT =$ all of the sin and cos terms in lines 3 to 6 in equation (1) and $i, j =$ point of sale transactions ($POS$) and bill payments ($BP$), number of ATMs ($ATM$), and number of branches ($BR$). Scale economies for the translog would set $\partial FT/\partial \ln Q_i$ to zero.

During 2002-2004 across our 11 countries, 90% of non-cash point of sale transactions were from cards while 10% were from checks. Using the payment scale information just cited, an expected point of sale card scale economy value could be $SCE_{POS}^* = .39$. A random number generator is used to generate a normal frequency distribution with 198 observations (the size of our panel data set) giving the column vector $SCE_{POS}^*$ with a mean of .39 and a standard deviation ($\sigma$) of .10.27 With $\sigma_{POS} = .10$, the 68% confidence interval ranges from .49 to .29 but could be made tighter or looser by varying $\sigma$. A similar procedure is applied to bill payments: $SCE_{BP}^* = .48$, obtained from a study of U.S. automated clearing houses, and using $\sigma_{BP} = .10$.28

Mixed estimation is achieved by estimating the relationship in (4) for $i = POS, BP$ jointly with our cost function (1) and share equation (2):

26 The terminology "prior information" is suggestive of Bayesian statistics and the difficult problem of determining the appropriate posterior distribution of the estimated parameters (for hypothesis testing). However, the technique used here is grounded in classical statistics via the concept of fiducial probability (Paulus, 1975; Barnett, Kopecky, and Sato, 1981) and hypothesis testing proceeds as it would otherwise.

27 Alternatively, a random number generator could generate a uniform distribution over a given specified range which suggests a weaker set of prior information but one where the emphasis is on the range rather than the mean.

28 If we had prior information on scale effects associated with ATM or branch office networks this could be accommodated in a similar manner.
\[ SCE_i^* = \alpha_i + 1/2 \sum_{j=1}^{4} \alpha_{i,j} \ln Q_j + \sum_{k=1}^{2} \delta_{i,k} \ln P_k + \frac{\partial F T}{\partial \ln Q_j} \] (4)

Cross-equation equality restrictions among the same parameters in equations (1), (2), and (4) enforce the mixing of the prior and panel information. By varying the standard deviation in the random number generating process, the "strength" or robustness of the panel data in determining payment scale economies can be illustrated.\(^{29}\) In what follows, we first present scale economy results from our cross-country model without any prior information and then again when prior information is mixed with information in our panel data set. In this manner, we can see how different payment scale economies from other studies may be from the results obtained from our cross-country analysis and illustrate the apparent robustness of the cross-country results when combined with the prior information.

4 Payment Scale Economies: Results

4.1 Scale Economies by Country

Figure 1 illustrates how the ratio of predicted operating cost to asset value (Y-axis) changes as the log of asset value rises (X-axis). Cubic splines were used to graph the average cost curves in Figure 1 and is suggestive of how each country’s banking sector average operating cost varies with the asset size of this sector. The curves are based on a translog cost function but results using the more complex Fourier form are virtually identical. Norway, Finland, Denmark, and Sweden form the first part of each curve while the middle reflects Belgium, the Netherlands, Spain and the U.K. In 1987 and also in 1995, Spain and the U.K., experienced a higher ratio of operating cost (OC) to total assets (TA) than did a number of countries that preceded it (Belgium and the Netherlands) and this accounts for the upward slope in the curve for these two years. The last part of each curve covers Italy, France, and Germany.

Shifts in these curves between the three time periods shown illustrate the degree of cost reduction from changes in the composition of payments (small effect), the shift away from branches to ATMs (the major effect),

\(^{29}\) A zero standard deviation effectively imposes an exact restriction equal to the mean of the prior information while a very large \(\sigma\) value yields results close to what may be obtained without prior information.
Figure 1: Predicted unit operating cost (OC/TA) vs. ln(Total Asset Value) under a Translog cost function

and other technical change.\textsuperscript{30} In predicting total operating cost (OC), only input prices were held constant at their mean values. As noted below, this is the closest we can get to a standard average cost curve.

Figure 2 shows how an approximation to unit payment cost varies by the total number of payment transactions. 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 2 are not average cost curves as the level of these curves (in U.S. cents per transaction) is too high.\textsuperscript{31} However, the slopes shown are a fair reflection of

\textsuperscript{30} We do not have quality adjusted input prices for bank use of computers and telecommunication facilities. Thus cost reductions here will be incorporated in and shown as decreases in predicted payment operating cost in Figure 1. When quality adjusted input prices exist and are specified, technical change can be separated into the portion derived from the cost function and the portion embodied in input prices. In either case, operating cost falls.

\textsuperscript{31} 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 constant, these mean values and their associated
how payment unit costs change with payment volume. As seen, these slopes (and their associated scale economies) vary with payment volume but are quite similar for the three years shown, suggesting that a changing payment mix or technology seemingly adds little to the operating cost reductions shown over time in Figure 1.

Table 4 presents our payment-related scale economy (SCE) or cost elasticity estimates.\textsuperscript{32} In Column 1, the countries are ranked according to their total payment volume in 2004. Note that in 2004 total 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% costs add to the level of the payment costs being predicted. The inability to obtain average costs from a translog (or Fourier or composite) multi-output cost function was noted in Baumol, Panzar, and Willig (1982). The exception is for Figure 1 where total operating cost is being predicted, not the portion of this cost associated with payments (or ATMs or branches) separately.

\textsuperscript{32}Since scale economies are defined as $SCE_i = \partial \ln OC / \partial \ln Q_i$ for $i =$ point of sale transactions, bill payments, ATMs, and branch offices, values $< (>) 1.0$ represent scale economies (diseconomies).
Table 4: Non-Cash Payment Volumes and Translog Scale Economies by Country

<table>
<thead>
<tr>
<th>Country</th>
<th>Payment Volume 2004</th>
<th>Average Payment SCE</th>
<th>Point of Sale SCE</th>
<th>Bill Payment SCE</th>
<th>ATM SCE</th>
<th>Branch SCE</th>
<th>Total Realized SCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>14,748</td>
<td>.23</td>
<td>.06</td>
<td>.17</td>
<td>.22</td>
<td>.59</td>
<td>1.09</td>
</tr>
<tr>
<td>France</td>
<td>13,926</td>
<td>.30</td>
<td>.08</td>
<td>.22</td>
<td>.31</td>
<td>.36</td>
<td>.97</td>
</tr>
<tr>
<td>U.K.</td>
<td>12,919</td>
<td>.35</td>
<td>.11</td>
<td>.24</td>
<td>.36</td>
<td>.27</td>
<td>.99</td>
</tr>
<tr>
<td>Spain</td>
<td>4,335</td>
<td>.30</td>
<td>.10</td>
<td>.20</td>
<td>.23</td>
<td>.48</td>
<td>1.04</td>
</tr>
<tr>
<td>Netherlands</td>
<td>3,563</td>
<td>.17</td>
<td>.09</td>
<td>.09</td>
<td>.24</td>
<td>.65</td>
<td>1.06</td>
</tr>
<tr>
<td>Italy</td>
<td>3,094</td>
<td>.21</td>
<td>.05</td>
<td>.16</td>
<td>.17</td>
<td>.62</td>
<td>.99</td>
</tr>
<tr>
<td>Belgium</td>
<td>1,594</td>
<td>.20</td>
<td>.10</td>
<td>.10</td>
<td>.26</td>
<td>.59</td>
<td>1.05</td>
</tr>
<tr>
<td>Sweden</td>
<td>1,488</td>
<td>.33</td>
<td>.18</td>
<td>.15</td>
<td>.39</td>
<td>.37</td>
<td>1.09</td>
</tr>
<tr>
<td>Finland</td>
<td>1,244</td>
<td>.35</td>
<td>.19</td>
<td>.16</td>
<td>.40</td>
<td>.34</td>
<td>1.09</td>
</tr>
<tr>
<td>Norway</td>
<td>1,117</td>
<td>.34</td>
<td>.19</td>
<td>.15</td>
<td>.40</td>
<td>.34</td>
<td>1.08</td>
</tr>
<tr>
<td>Denmark</td>
<td>1,081</td>
<td>.24</td>
<td>.12</td>
<td>.12</td>
<td>.28</td>
<td>.52</td>
<td>1.04</td>
</tr>
<tr>
<td>Average</td>
<td>5,374</td>
<td>.27</td>
<td>.11</td>
<td>.16</td>
<td>.30</td>
<td>.47</td>
<td>1.04</td>
</tr>
</tbody>
</table>

Total 11 country non-cash transactions 59.1 billion in 2004 (U.S. $45 bln)

1Millions of non-cash transactions. All data are rounded. *Calculated separately.

more non-cash transactions per year than individuals in Europe, indicating greater cash replacement. The average scale economy for all payments for each country using the translog cost function is shown in Column 2. The average across countries is .27 (for the Fourier form it is .29) and indicates that substantial scale effects would be expected as payment volume rises. The point of sale and bill payment SCEs appear quite low and, at .11 and .16, respectively, are considerably lower than independent estimates for Norway, the Netherlands, or Belgium using different data sources.

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 results indicated that the simpler translog form gave very similar results to the more complex Fourier form, that the average payment SCE results were not dramatically 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.\(^{33}\) 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).\(^{34}\) Finally, a differencing para-

\(^{33}\)The translog average payment SCE is .27 over the entire period and is .24 using data for only the first half (1987-1995) and .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 .30 while for the 6 largest countries it is .36. Estimation with 11 country-specific dummies raises the average payment SCE to .34 so country identification explains some of the reduction in operating cost, leaving less to be associated with our four outputs.

\(^{34}\)Estimation with 18 time-specific dummies gives an average payment SCE of .20. Here
meter ($\rho = .85$) 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.\textsuperscript{35} Thus it appears that a fixed effects estimation framework would not offer much improvement. As well, fully adjusting for positive autocorrelation yields anomalous results.\textsuperscript{36} Considering these results, we focus on the scale economy estimates in Table 4 which applies a translog cost model using unadjusted levels data using the entire time period and all 11 countries together (in panel econometrics jargon, a "total pooled regression").

As noted above for Figure 2, the payment scale effects shown there are not the source of the cost reductions between time periods illustrated in Figure 1. This source appears to be associated with the shift away from expensive branch offices toward ATMs for the delivery of cash acquisition and other banking services (probably along with lower computer and telecommunication input expenses and other technical changes we can not directly identify). This is seen from Figure 3 where the ratio of predicted operating cost associated with ATMs and branches to the value of banking assets for each country (Y-axis) is arrayed against the log of asset value (X-axis). For the same reasons noted in the discussion of Figure 2, the cost curves shown in Figure 3 are not average cost curves: their level is too high but their slopes and shifts over time are indicative of scale effects and cost reductions in delivering banking services to depositors.

### 4.2 Realized Scale Economies

The payment SCE in Table 4 is the simple sum of the point of sale plus bill payment scale economies (Columns 3 and 4).\textsuperscript{37} Equally weighting and

\textsuperscript{35}Setting $\rho = .75$ 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).

\textsuperscript{36}Indeed, a grid search using values of $\rho = 0.25, 0.50, 0.75$, to 1.0 to transform the data indicated that the average payment SCE (Column 2) was fairly stable up to the point where $\rho = 0.50$ but completely fell apart when $\rho = 1.0$.

\textsuperscript{37}Numbers may not add up properly due to rounding.
adding these two payment SCEs together is reasonable in this case since the average percent changes in point of sale transactions (140%) and bill payments (151%) shown at the bottom of Table 3 over 1987-2004 are very similar. Scale economies associated with ATMs are on average lower at .30 while economies associated with branch offices are lower still at .47. Individually, all four of the specified characteristics of banking output—two classes of payments, ATMs, and branches—experience scale economies. It is common to report the simple sum of all individual SCEs as representing the overall scale effect. This is shown in the next to last column in Table 4 under the heading "Total SCE" and seems to suggest that overall operating cost economies reflect constant returns to scale or even slight diseconomies.

This apparent result could be rationalized by noting that collecting deposits using branch offices is (depending on the interest rate cycle) typically a lower cost method than purchasing funds from other banks or issuing other debt instruments. Thus, although constant returns or slight diseconomies may seem to exist overall for operating costs, profits could still be improved relative to a greater reliance on purchased funds which affects interest expenses. While this reasoning is correct, we prefer a different interpretation—
one that suggests that the simple sum of underlying scale economy estimates (giving a Total SCE) is appropriate only if all of our four underlying output characteristics experienced the same rate of growth. As seen in the last row of Table 3, this clearly is not the case. As point of sale transactions expanded by 140%, bill payments by 151%, ATMs by 434%, and branches by 9.8%, the Total SCE value typically reported in the literature and in Table 4 is not valid here.

What is needed is a measure of overall scale economies that have been actually realized. If all banking outputs have grown by the same percentage amount, then the scale economies actually realized would be:

\[
\text{RealizedSCE} = 1.0(.11) + 1.0(.16) + 1.0(.30) + 1.0(.47) = 1.04
\]

from \( \sum_i SCE_i \) for \( i = \text{POS, BP, ATM, BR} \) using the average scale economies at the bottom of Table 4. What if the number of branch offices, on average, did not rise by 9.8% overall (as shown in Table 3), but instead was constant while the other three outputs all expanded as the same rate? If the number of offices did not change overall, the scale economy estimated from the cross-country data where in some countries it rose while in others it fell should have a zero weight since the overall effect on operating cost should also be zero. This gives:

\[
\text{RealizedSCE} = 1.0(.11) + 1.0(.16) + 1.0(.30) + 0.0(.47) = .57.
\]

Finally, if only ATMs had expanded and nothing else, then the realized scale economy should be .30. This progression suggests that in general the output that experienced the greatest growth should have the largest weight for its scale economy while other outputs experiencing less growth should have smaller weights. A reasonable measure of realized scale economies should be able to generate the same numerical values for the previous three

\[\text{Mathematically, "Total SCE" is linked to the individual SCEs through a specific weighting scheme. In particular, one can show that } SCE_{tot} = \sum_i w_i SCE_i, \text{ with } w_i = s_i^C g_i^T / \left( \sum_i s_i^C g_i^T \right), \text{ where } s_i^C \text{ denotes the (average) cost share of output } i, s_i^T \text{ the (average) transaction share, and } g_i^T \text{ the average growth rate of transactions. However, since the output cost shares are unknown, an alternative weighting arrangement is developed below.}\]

\[\text{In the standard banking literature, where "outputs" are specified from balance sheet data on loans (as a total or distinguished by type), security holdings, and/or deposits (as a total or by type), reporting a Total SCE is less of a problem since these categories of outputs tend to expand at similar rates and rarely is the information so detailed that reductions are observed.}\]
simple examples, handle situations where the growth rates of the various outputs are not all the same, and weight the output with the largest growth the highest. A weighted average is called for but not the usual type where the sum of the weights equals 1.0.

One way of meeting these conditions (there may be others) would be to define the scale economy weights as ratios of each output’s growth rate divided by the growth rate of the output that experienced the greatest expansion, whether this be 10%, 100%, or in our case 434% for ATMs giving:

\[
\text{RealizedSCE} = \frac{1.40}{4.34} \cdot 0.11 + \frac{1.51}{4.34} \cdot 0.16 + \frac{4.34}{4.34} \cdot 0.30 + \frac{0.98}{4.34} \cdot 0.47 = 0.40.40
\]

If all the outputs had expanded by 434%, all the numerators would be 4.34 yielding weights of 1.0 and we obtain the Total SCE value of 1.04. If the first three outputs expanded by 434% but branches were constant (a zero in the numerator), we get 0.57 while if only ATMs expanded, we obtain 0.30.

Our scale economy estimates are derivatives which are appropriate for small changes in output while our growth figures are very large and cover 18 years. Thus the Realized SCE figures shown by country in Table 4 are approximations. However, if instead of using growth rates over 18 years we had used average annual (compound) growth rates, our Realized SCE would equal 0.47.\(^{41}\) Thus on an annual average basis or over 18 years, realized operating cost scale economies are less than half the value of the simple sum of derivatives and overall operating cost scale economies are apparently quite large. Although leasing costs, depreciation schedules, and prices paid for surplus offices will differ across countries, all other things being equal, banks that reduce their branch operations the most should realize the greatest overall scale benefit as payments and ATMs expand. This holds in the future even if branch networks no longer contract but remain constant.

### 4.3 Scale Effects Using Mixed Estimation

Joint estimation of equations (1), (2), and (4) allow the mixing of prior scale economy information with our panel data and is used to illustrate the "strength" or robustness of the panel data in determining payment scale

\(^{40}\)The Realized SCE using scale estimates from a Fourier cost function is 0.43.

\(^{41}\)The compounded annual growth rates are, in the same order as the numerators in the text equation, 4.98%, 5.25%, 9.75%, and 0.52%.
Table 5: Effect of Prior Information on Cross Country Scale Economy Estimates

<table>
<thead>
<tr>
<th>Prior Information:</th>
<th>Average Payment SCE</th>
<th>Point of Sale SCE</th>
<th>Bill Payment SCE</th>
<th>ATM SCE</th>
<th>Branch SCE</th>
<th>Total SCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Prior:</td>
<td>.27</td>
<td>.11</td>
<td>.16</td>
<td>.30</td>
<td>.47</td>
<td>1.04</td>
</tr>
<tr>
<td>$\sigma = .10$</td>
<td>.67</td>
<td>.29</td>
<td>.39</td>
<td>.11</td>
<td>.27</td>
<td>1.05</td>
</tr>
<tr>
<td>$\sigma = .15$</td>
<td>.54</td>
<td>.21</td>
<td>.34</td>
<td>.18</td>
<td>.33</td>
<td>1.06</td>
</tr>
<tr>
<td>$\sigma = .20$</td>
<td>.47</td>
<td>.17</td>
<td>.30</td>
<td>.21</td>
<td>.37</td>
<td>1.05</td>
</tr>
<tr>
<td>$\sigma = .50$</td>
<td>.36</td>
<td>.13</td>
<td>.23</td>
<td>.27</td>
<td>.42</td>
<td>1.04</td>
</tr>
</tbody>
</table>

The first row in Table 5 repeats the average cross-country estimated scale values shown in the last row of Table 4. The cross-country point of sale scale economy (.11) is lower than the mean prior information (.39) derived from estimates of incremental and average card payment cost per transaction for the Netherlands and Belgium. Similarly, the cross-country bill payment scale value (.16) is lower than the econometric estimate (.48) for Federal Reserve automated clearing houses (ACH) in the U.S. There are several reasons why our prior information could be different from our cross-country estimates but, as we are unable to determine their empirical importance, they are noted in a footnote.42

When the normal distribution of our prior information is very "tight", so 68% of these scale values fall within .29 to .49 for point of sale and .38 to .58 for bill payments, the combined model generates the scale estimates shown in the row labeled $\sigma = .10$. In this case, prior information raises the two restricted scale estimates up to the value of $SC_E^*$ but correspondingly reduces the unrestricted ATM and branch scale estimates, leaving the

42 First, average transaction volumes underlying the cross-section and prior information data are different and this can affect the scale estimates. The volume of card transactions in Norway, the Netherlands, or Belgium are lower than the average for our 11 countries while U.S. ACH volume far exceeds that of giro transactions in any one of our 11 countries. Second, the cross-country point of sale scale estimate is derived holding non-payment related costs associated with branches constant within a statistical framework while costs associated with card transactions in the Netherlands and Belgium use accounting data which may allocate branch costs and other overhead expenses differently. Third, an ACH transaction in the U.S. would likely involve more telecommunications expense than a giro transaction in (geographically smaller) European countries. Finally, we had to adjust the cross-country branch data to make it more homogeneous and this adjustment can affect the estimated branch scale value and, in turn, the other scale values as well.
simple sum of the four values almost unchanged. As $\sigma$ is progressively increased, the "strength" of the prior information is reduced allowing for greater overlap of the normally distributed prior information with the parameter distributions of the cross-country model. However, only when the prior information is very weak ($\sigma = .50$) are the two restricted scale economy values much above the illustrative value of $SCE - \sigma$, suggesting the cross-country data yields scale estimates that are themselves relatively "tight" and not easily dominated by the prior information. From this perspective, it is concluded that the cross-country scale estimates and the additional independent (prior) information are different and not easily dominated by the other. Had one set of estimates easily dominated the other, this would have suggested that the dominant set of estimates would likely be the more reliable. As it is, we are effectively left with two sets of scale estimates to evaluate in the following section.

5 Replacement of Cash and Stored Value Cards

The average value of a cash transaction at the point of sale in the Netherlands and Belgium (respectively, €9 and €18) is much smaller than for debit card (€44 and €50). As seen earlier in Table 2, at these transaction values the average total cost of using cash (€.30) is €.19 cheaper than a debit card (€.49) in the Netherlands but only €.02 lower in Belgium.\textsuperscript{43} Cash is also cheaper than debit cards for both retailers and banks in the Netherlands, as illustrated in Table 6 (rows 2 and 3).\textsuperscript{44} However, if the value of a transaction is around €10 to €11, the average variable cost (not total cost) of cash and debit card transactions are essentially equalized in these two countries. As these cost differences are not reflected in explicit per transaction prices, users have no real price incentive to use one instrument over the other and relative use is driven by apparent consumer preferences and trade-offs, such as the universal acceptability of cash or the convenience of using a debit card that does not require "refilling" at an ATM or other terminal. While underlying cost does not directly influence consumer use, banks and retailers who pay these expenses are influenced

\textsuperscript{43}Estimates for the U.S. focus on marginal costs and suggest that the retail (bank) cost of a cash payment is $.30 ($.07) while that of a PIN debit are $.57 ($.27). This suggests that the difference in marginal costs for a low value transaction averaging $11.52 favors cash by $.47, considerably more than that for the Netherlands or Belgium (Garcia-Swartz, Hahn, and Layne-Farrar, 2006, Table 3).

\textsuperscript{44}For Belgium, retail (bank) costs of cash and debit cards are, respectively, €.27 and €.28 (€.26 and €.27).
Table 6: Illustration of Debit Card Cost Reduction

<table>
<thead>
<tr>
<th></th>
<th>Netherlands, 2002&lt;sup&gt;4&lt;/sup&gt;</th>
<th>Cash</th>
<th>Debit Card</th>
<th>Stored Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Total Cost:</td>
<td>€ .30</td>
<td>€ .49</td>
<td>€ .93</td>
<td></td>
</tr>
<tr>
<td>Retail Sector Unit Cost:</td>
<td>.16</td>
<td>.24</td>
<td>.15</td>
<td></td>
</tr>
<tr>
<td>Banking Sector Unit Cost:</td>
<td>.14</td>
<td>.25</td>
<td>.78</td>
<td></td>
</tr>
<tr>
<td>Scale Economy Value:</td>
<td>SCE Volume Growth Years</td>
<td>Netherlands/Belgium SCE</td>
<td>64%</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>Average Payment SCE</td>
<td>.27</td>
<td>53%</td>
<td>6.6</td>
</tr>
<tr>
<td></td>
<td>Average Debit Card SCE</td>
<td>.11</td>
<td>44%</td>
<td>5.5</td>
</tr>
</tbody>
</table>

<sup>4</sup>Source: Brits and Winder (2005), Table 4.3, data are rounded.

by them along with the revealed preferences of users.<sup>45</sup> Considering the growth of debit card use in Europe, banks and retailers would benefit if the expense of debit card use were reduced. One way to do this would be to directly price cash and debit card use so users would have an incentive to shift to cash, which—according to Table 6—is cheaper. The alternative is to accommodate consumer preferences and consider ways to realize more fully the scale economies associated with debit card use to lower bank expenses. If successful, this could permit banks to reduce fees charged merchants for accepting a debit card and equalize better merchant costs of cash and cards for low value transactions where cash now has an advantage. This may also lead banks to replace stored value cards so that a single technology could be used for both larger and smaller value transactions.

Using scale economies computed for the Netherlands and Belgium as well as from our cross-country estimation (shown in the lower half of Table 6), it is possible to illustrate the required debit card volume growth and time it may take at current rates of expansion to reduce the average total cost of debit card use (€ .49) down to a level close to that of cash (€ .30), a potential reduction of € .19 per transaction. Given the illustrative scale economies (SCEs) shown in Table 6, the expansion in debit card transaction volume apparently needed to reduce the average total cost of a debit card by € .19 is quite large and, at the annual growth rate of transactions in the Netherlands over 2002-2004 (around 8%) could take from 6 to 8 years to be realized.<sup>46</sup> This is likely too optimistic, however, since debit card volume...

---

<sup>45</sup>Banks introduced ATMs to reduce the costs of supplying cash to depositors over the counter and the benefits were twofold. A cash withdrawal was cheaper at an ATM than at a branch office and, in addition, ATMs could be substituted for more expensive branch offices and still provide the same (or greater) level of cash access for depositors.

<sup>46</sup>The relationship between changes in transaction volume ($\partial \ln Q$) and reductions in
growth in the Netherlands has been declining year-to-year over 2000-2004, as would be expected when adoption of a "new" payment technology has seemingly passed its inflection point on a logistic S-curve (Bolt, Humphrey, and Uittenbogaard, 2005). If the expected domestic growth in the near future only averaged 6% annually, the illustrative reduction in debit card unit costs would take from 7 to 11 years. And what if the S-curve projections are correct in forecasting even lower growth in the future as card use approaches "maturity"?

A more viable alternative would be to consider merging card processing centers among European countries to achieve potentially larger scale economy benefits over a shorter time period—benefits less dependent upon volume growth within a single country. Importantly, around 31% of bank and 35% of retailer total debit card expenses—those associated with front office operations (Brits and Winder, 2005, Table 4.2)—may be relatively constant since the scale-related cost reductions would be due to the pooling of card volumes across countries, not the expansion of volume within a home country. Indeed, the pooling of card volumes suggests that the appropriate SCE value is lower that the .39 value implied from cost analysis in the Netherlands and Belgium since this value includes front and back office expenses while the two lower values in Table 6 have attempted to separate branch costs that do not vary with payment volume from payment processing expenses.

Mergers are clearly more difficult to arrange than merely expanding existing processing centers slowly over time and extra telecommunication expenses will offset some of the scale benefits from consolidating processing centers. Even so, it would be instructive to study the experience of the U.S.—which operates large card processing centers—to get a better idea of the cost and benefits of a possible cross-border consolidation, much in the way that the consolidation of Fedwire operations in the U.S. suggested that similar cost benefits may be obtained from cross-border consolidation of Target (the European large value payment network).

\[
\frac{\partial \ln AC}{\partial \ln Q} = \text{SCE} - 1.
\]

Setting \( \frac{\partial \ln AC}{\partial \ln Q} = -39\% \) (from -€.19/€.49), the implied growth in transaction volume shown in Table 6 is \( \frac{\partial \ln Q}{\partial \ln Q} = -39\%/(\text{SCE} - 1) \).

\[47\] The year-to-year growth rates over 2000-2004 were, respectively, 19.0%, 12.1% 8.2% and 7.8%.
6 Summary and Conclusions

Analysis of the average total bank plus merchant costs of making a cash or debit card transaction at the point of sale indicate that cash is cheaper for low value transactions—those less than ten to eleven euros. Although the variable cost of a stored value card transaction is by far the lowest of the three, its average total cost is the highest due to large fixed costs spread over a small volume—a volume that seems stalled at a relatively low level of market penetration (Van Hove, 2006). The current composition of payment instrument use at the point of sale in European countries is primarily determined by consumer preferences and trade-offs such as the universal acceptability of cash and the convenience of debit cards which do not need to be "refilled" at terminals or banking offices (unlike cash or a stored value card). Explicit pricing per payment transaction, used broadly in Norway, is not the norm and explicit payment prices that do exist in Europe are typically piecemeal, need not reflect the underlying relative costs of different instruments, and in some cases induce users to use the more expensive instrument (c.f., Sveriges Riksbank, 2004).

In 2002, the average total (fixed plus variable) bank plus merchant cost of a cash transaction at the point of sale in the Netherlands was € .30 while a debit card transaction was € .49. While the unit cost of cash was € .19 less than that for a debit card in the Netherlands, it was only € .02 lower in Belgium. As the replacement of cash by debit cards for smaller value transactions is importantly influenced by unit costs and unit costs are largely dependent on transaction volume (since the majority of expenses are fixed), our goal was to try to determine payment scale economies in Europe, especially for debit cards. When combined with expected transaction growth within a country, or the consolidation of card processing operations across countries, scale economy estimates permit future card unit costs to be roughly approximated and the likelihood of debit cards replacing small value cash transactions assessed.

Information on card scale economies for three countries (Norway, the Netherlands, and Belgium) ranged from .43 to .39 and was compared to economies derived from a cost function using a panel of payment and banking data for 11 European countries over 18 years. Estimated scale economies for point of sale transactions (primarily debit cards) averaged .11 while bill payment (paper and electronic giro) transactions averaged .16. For all payments together, the scale economy was .27 (the simple sum of point of sale and bill payment economies).

These scale estimates were used to illustrate the possible context and
timeframe for debit card costs to fall to a level similar to that for cash, potentially enabling debit cards to replace cash for small value cash and stored value transactions as well as lowering unit costs for higher value transactions. The timeframe differs depending on how debit card volume expands—over time as a country’s internal market expands while maintaining current within-country processing arrangements (perhaps taking seven to ten years) or through consolidation of processing centers across countries where volume growth would be much larger over a shorter time period. A consolidation approach would also have an upper limit to processing volume that would be a multiple of that associated with a single country.

This study provides preliminary scale economy information that may be helpful in outlining possible benefits from a Single European Payments Area (SEPA) arising from the consolidation of electronic payment processing centers across the Euro zone. If this approach were pursued, the experience of the U.S. which consolidates card processing across states, may serve as a useful example concerning the realized costs and benefits, as well as likely implementation issues of cross-border consolidation. However, for SEPA to become a success, both consumers and merchants need to accept the harmonized pan-European payment instruments. This will most likely translate into payment pricing conditions—relative to the (incumbent) domestic payment instruments—that must be met. This interplay between scale benefits, emerging payment market structure, and pricing requirements is complex and has not been previously addressed.
7 Bibliography


Bank for International Settlements, (various years). Statistics on Payment Systems in the Group of ten Countries, Basle, Switzerland. Also called the "Red Book".


European Monetary Institute (various years). Payment Systems in the European Union, Frankfurt, Germany. Early editions of the "Blue Book".


The LSQ nonlinear procedure in TSP (Time Series Processor, Version 4.5) was used in estimation while Doornik and Hendry’s GiveWin program (Version 2) was used for the graphs. The number of observations = 198 and the Log likelihood = 546.821. Standard errors were computed from a heteroscedastic-consistent matrix (Robust-White). Parameter estimates and t-statistics are shown below for the system estimation of equations (1) to (2). The cost function was concave.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter</th>
<th>Estimation</th>
<th>T-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>$\alpha_0$</td>
<td>2.1562</td>
<td>4.54</td>
</tr>
<tr>
<td>$\ln Q_{CARD}$</td>
<td>$\alpha_1$</td>
<td>.2117</td>
<td>3.44</td>
</tr>
<tr>
<td>$\ln Q_{GIRO}$</td>
<td>$\alpha_2$</td>
<td>-.4530</td>
<td>-5.77</td>
</tr>
<tr>
<td>$\ln Q_{ATM}$</td>
<td>$\alpha_3$</td>
<td>.2192</td>
<td>2.30</td>
</tr>
<tr>
<td>$\ln Q_{BR}$</td>
<td>$\alpha_4$</td>
<td>1.2095</td>
<td>13.49</td>
</tr>
<tr>
<td>$1/2(\ln Q_{CARD})^2$</td>
<td>$\alpha_{11}$</td>
<td>.0063</td>
<td>.34</td>
</tr>
<tr>
<td>$1/2(\ln Q_{GIRO})^2$</td>
<td>$\alpha_{22}$</td>
<td>.0084</td>
<td>.24</td>
</tr>
<tr>
<td>$1/2(\ln Q_{ATM})^2$</td>
<td>$\alpha_{33}$</td>
<td>.0668</td>
<td>5.04</td>
</tr>
<tr>
<td>$1/2(\ln Q_{BR})^2$</td>
<td>$\alpha_{44}$</td>
<td>.3614</td>
<td>12.63</td>
</tr>
<tr>
<td>$\ln Q_{CARD} \ln Q_{GIRO}$</td>
<td>$\alpha_{12}$</td>
<td>.0267</td>
<td>1.57</td>
</tr>
<tr>
<td>$\ln Q_{CARD} \ln Q_{ATM}$</td>
<td>$\alpha_{13}$</td>
<td>.0476</td>
<td>3.52</td>
</tr>
<tr>
<td>$\ln Q_{CARD} \ln Q_{BR}$</td>
<td>$\alpha_{14}$</td>
<td>-.1076</td>
<td>-7.17</td>
</tr>
<tr>
<td>$\ln Q_{GIRO} \ln Q_{ATM}$</td>
<td>$\alpha_{23}$</td>
<td>.0684</td>
<td>4.06</td>
</tr>
<tr>
<td>$\ln Q_{GIRO} \ln Q_{BR}$</td>
<td>$\alpha_{24}$</td>
<td>-.0695</td>
<td>-3.72</td>
</tr>
<tr>
<td>$\ln Q_{ATM} \ln Q_{BR}$</td>
<td>$\alpha_{34}$</td>
<td>-.1998</td>
<td>-9.91</td>
</tr>
<tr>
<td>$\ln Q_{CARD} \ln P_L$</td>
<td>$\delta_{11}$</td>
<td>-.0450</td>
<td>-6.64</td>
</tr>
<tr>
<td>$\ln Q_{GIRO} \ln P_L$</td>
<td>$\delta_{21}$</td>
<td>-.0819</td>
<td>-8.45</td>
</tr>
<tr>
<td>$\ln Q_{ATM} \ln P_L$</td>
<td>$\delta_{31}$</td>
<td>-.1007</td>
<td>-11.22</td>
</tr>
<tr>
<td>$\ln Q_{BR} \ln P_L$</td>
<td>$\delta_{41}$</td>
<td>.2120</td>
<td>18.19</td>
</tr>
<tr>
<td>$\ln P_L$</td>
<td>$\beta_1$</td>
<td>1.2916</td>
<td>25.95</td>
</tr>
<tr>
<td>$1/2(\ln P_L)^2$</td>
<td>$\beta_{11}$</td>
<td>.1826</td>
<td>37.94</td>
</tr>
</tbody>
</table>

Cost function equation (1): $R^2 = .97$, Durbin-Watson = .43
Cost share equation (2): $R^2 = .42$, Durbin-Watson = .44