Queen's Economics Department Working Paper No. 1380

Where are the economies of scale in Canadian banking?

Robert McKeown
Queen's University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

4-2017

Where are the economies of scale in Canadian banking?

Robert J. McKeown*

April 17, 2017

Abstract

Using a new data set from the Office of the Superintendent of Financial Institutions, I conduct an in-depth study on cost efficiency and returns to scale (RTS) in Canadian banking. I estimate a transcendental log cost function for the six largest Canadian commercial banks which account for approximately 90% of chartered bank assets over the 1996-2011 sample period. The minimal amount of firm entry and exit simplifies many difficulties in the analysis, and the panel dynamic ordinary least squares estimator (PDOLS) provides less biased results than the fixed-effect OLS. Departing from previous studies in banking, I calculate whether the estimated cost function satisfies the microeconomic properties of a monotonicity and price concavity. To my knowledge, this is the first paper to find evidence of constant RTS among the Canadian banks. The result is robust to a number of different asset and price specifications. Furthermore, there is little evidence to suggest cost inefficiencies among the large Canadian banks. This is true whether the Greene (2005) true fixed effects ML estimator is estimated or a distribution-free approach is measured. Combining these two results, the large Canadian banks managed costs efficiently and minimized costs from 1996 to 2011.

Acknowledgments

I have benefitted from comments by Frank Milne, Jason Allen, Taylor Jaworski, Gregor Smith, Charles Beach, and the 3rd Annual Doctoral Workshop in Applied Econometrics. All responsibility for errors and the opinions expressed are my own.

*Queen's University, PhD Candidate

1 Introduction

Using new data from the Office of the Superintendent of Financial Institutions, I estimate a transcendental log cost function using the six largest Canadian commercial banks from 1996 to 2011. The Canadian banking system, on average, operated at constant returns to scale (RTS). To my knowledge, this is the first paper to find constant RTS among the largest Canadian banks as previous studies found increasing RTS of varying sizes. Increasing RTS implies that if a bank increases assets by 10 percent then total costs will increase by less than 10 percent. Furthermore, my specification allows me to estimate higher-order terms that produce a curved, or u-shaped, cost function. I find that some banks operated with increasing or even decreasing RTS. Additionally, I find that there is no, or very little, inefficiency in the use of inputs. Taking these results together, I conclude that the Canadian banking system took full advantage of changing economic conditions to operate at the minimum efficient scale. My work differs from previous studies on Canadian bank RTS in three fundamental ways: (i) variables are aggregated in such a ways as to avoid multicollinearity, (ii) prices are calculated to avoid mismeasurement, and (iii) the estimates are tested for monotonicity and price concavity – a prerequisite of a cost function. Chua et al. (2005) observe that if a cost function fails to satisfy these conditions, then interpretation of the coefficients becomes dubious. Lastly, the time period in question plays a role. In section 7.1, I estimate the Allen and Liu (2007) model with more recent data and find RTS has declined. After my own specification, it declines further. It is entirely possible that the cost structure was more stable from 1996 to 2011 than it was from 1983 to 2003.

Haldane (2010) states that research on RTS in banking fall into two categories: those that study a cross-section in one period of time and those that use difference-in-difference techniques to evaluate performance before and after an event such as a merger or acquisition. Since then Allen and Liu (2007), Davies and Tracey (2014), Wheelock and Wilson (2012) and Almanidis et al. (2015) estimate RTS using multiple period and cross-sectional data. Following Allen and Liu (2007), I use the panel dynamic OLS estimator to estimate a cost function with a short cross-section (N) and a long time dimension (T). This method has less bias in finite samples than OLS. Estimation is greatly simplified by the fact there were few

entries and exits in the Canadian banking system from 1996 to 2011. Figure 1 illustrates the consistency with which the Big Six controlled 90% of Canadian bank assets.

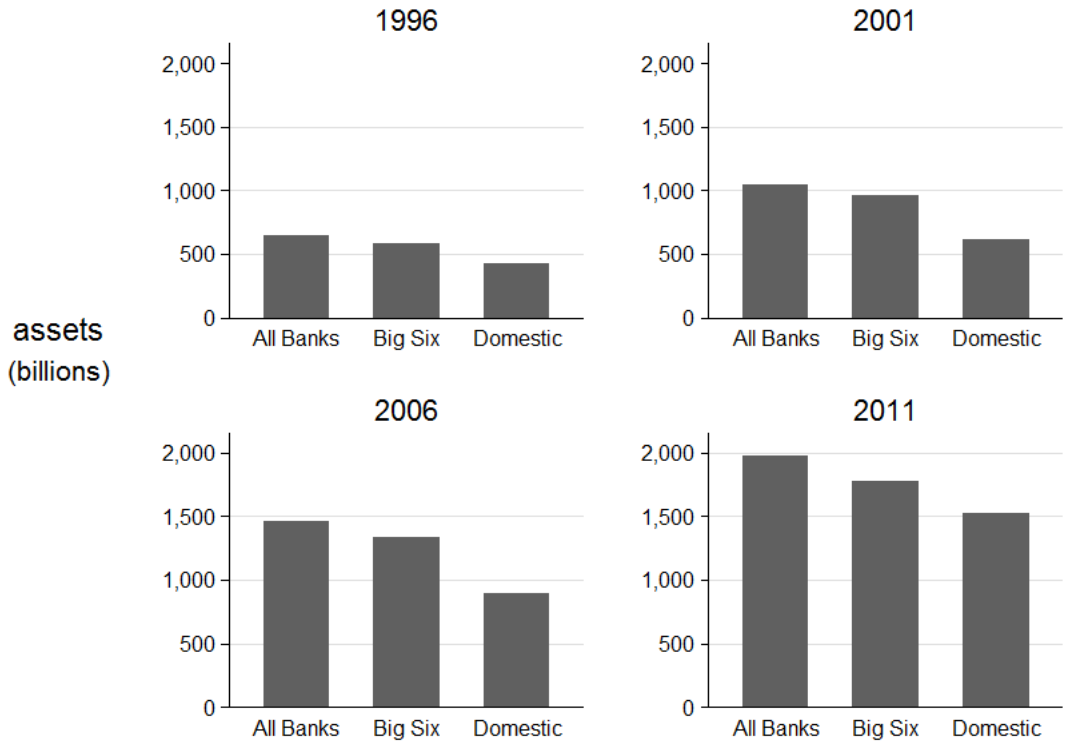


Figure 1: Total bank assets

Note: Assets are shown in three categories: the total reported to OSFI, those owned by the Big Six, and bank assets associated with a Canadian address, entitled 'domestic'. Assets are denominated in nominal values. Sources: OSFI and CANSIM.

Why measure returns to scale in Canadian banking? The first reason is domestic bank regulators and investors want to observe where along the average cost curve banks operate. Assuming a bank faces a u-shaped average cost curve, if banks operate to the right of the average cost curve nadir, then there are decreasing RTS . A policy that encourages competition through more banks might have a positive effect on costs and may improve bank profitability. Conversely, if banks operate to the left of the nadir, then there are increasing RTS and a policy that encourages expansion might be beneficial.

The second reason follows closely from the first. At times in the past, the 'Big Six' banks[1] have argued that they need to become larger to be more efficient and compete

[1]The 'Big Six' refers to Bank of Montreal (BMO), Canadian Imperial Bank of Canada (CIBC), Toronto Dominion (TD), Bank of Nova Scotia (BNS), Royal Bank of Canada (RBC) and the National Bank of Canada (NB).

internationally. In the late 1990's, the Canadian banks were keen on amalgamation. This was unpopular among the public and many academics, so that eventually the idea was quashed by the Canadian Federal Government. If banks operate with decreasing or constant RTS, this decision is in the best interest of consumers. Under increasing economies of scale, the answer is not so obvious. There is no guarantee mergers that lead to larger banks would be welfare improving; market power and pricing effects need to be considered. If the larger bank raises prices then it is possible a deadweight loss could reduce welfare. Kumar (2013) observes that studies on banking often confound increasing RTS with market power – larger banks decrease deposit rates, and this is picked up as scale economics. However the reality is often that the banks have some kind of market power that allows them to collect deposits at a lower cost. Hence estimates of RTS are often biased upwards. As I find constant RTS, this implies that either the Canadian banks had little market power on deposit rates or they operated with decreasing RTS. Most likely however, I posit that each enjoyed a similar quantity of market power irrespective of size. Each offered the same suite of financial products and a similar geographic branch coverage, so as they became larger, relative market power was held constant. Average interest expense rates between Big Six Canadian banks varied little. In McKeown (2017a), I compare the U.S. and Canadian banks and confirm that there were significant differences between the rates at larger and smaller U.S. commercial banks. Hence it is unlikely that market power is influencing RTS estimates. Otherwise, I would expect to observe more heterogeneity in rates such as that observed in the U.S.

Third, measuring efficiency at the banks might shed light on which institutions, if any, are of concern. I find no evidence that one bank is more cost efficient than another. Fourth, the Canadian Big Six banks were universal banks including retail and commercial banking, capital markets, wealth management providers and even insurers. They competed against each other to offer services both online and in brick-and-mortar branches[2]. There remain a number of rivals to the Big Six, but including them in a cost function is problematic. Two of their rivals, ATB Financial and Desjardins,[3] were not publicly traded companies. ATB

[2]Regulation prohibits Canadian banks from selling insurance products in a branch.

[3]These two deposit-taking institutions are large enough that Allen et al. (2013) and Perez-Saiz and Xiao (2014) refer to them and the Big Six as the 'Big 8'. Neither submit filings to OSFI, so ensuring the data is

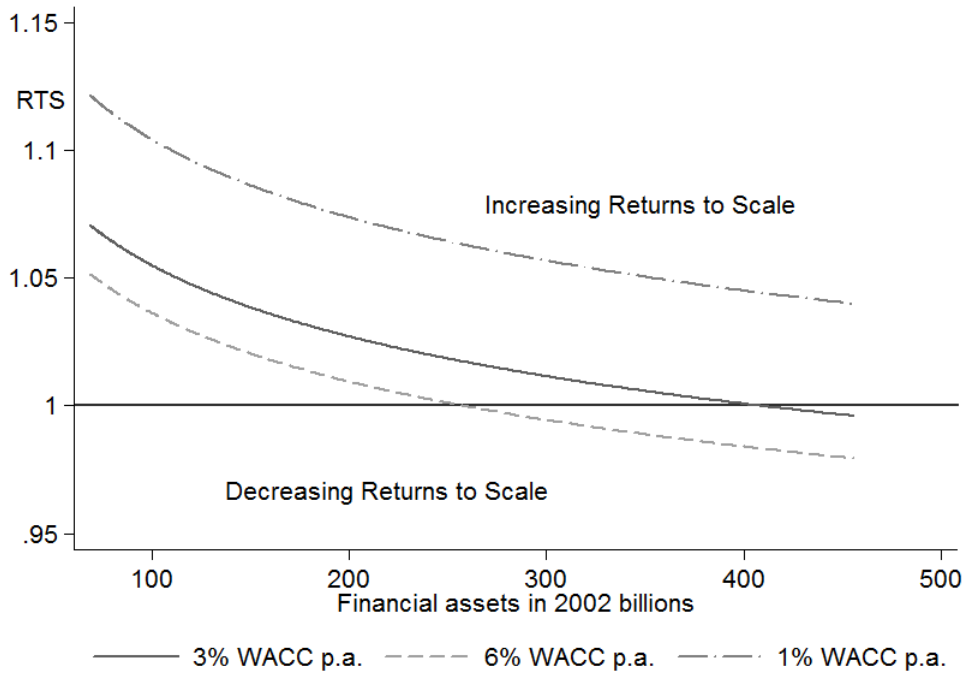


Figure 2: Short-run cost function RTS

Note: Using the coefficient estimates from table 8, RTS is evaluated at the median cost of labour and by proportionally increasing output for three different weighted average costs of capital (WACC). An increase in the WACC, which includes the implied cost of equity, shifts the RTS curve downward. Financial assets are total assets. See section 3.3 for details on calculations.

was a crown corporation and Desjardins was a caisse populaire (credit union) and both are domestic, regional providers that collected deposits in just one province. There were also a large number of smaller chartered banks. Some of these were owned by the Big Six and offered discount services, mostly online. Others were associated with the financing arm of a retailer such as the Canadian Tire Bank. Either of these entities could have costs paid by its parent, so it is best to remove them from the study. For independents, a difference in ownership structure or scope of operations could also confound the estimates relating to cost and size. The estimation is parametric so that each observation is equally weighted, and the difference in size from the Big Six to the next largest domestic, chartered, and independent bank is quite large. Including many small banks with few assets would create a range of asset values for which no observations are available. Considering the translog cost function is a linear approximation of a nonlinear function, a greater distance from one

comparable is not straight-forward.

observation to the next could reduce accuracy.

Internationally, Canada holds a number of interesting and, perhaps unique, features that merit it as a worthy object of study. Between 1996 and 2011, the aggregate Canadian banking system never suffered a quarter of negative profit. Even during the Great Financial Crisis (2007-09), the Canadian banking system performed well relative to their international peers. Of the twelve quarters between 2007 and 2009, CIBC reported three consecutive periods of negative profit while BMO and NB recorded one each. However, no other large Canadian bank reported a loss. This was in spite of the large exposure to the U.S. market where both the dot-com bubble and the financial crisis of 2007-09 originated. Studying the Canadian banking system provides a useful comparison to other jurisdictions such as Europe and Australia which have similarly large, universal banks. While the U.S. banking system is categorized as having many small, regional banks, American regulation allows for the creation of universal banks. Recently, the largest U.S. commercial banks own an increasing share of total U.S. banking assets. For a historical comparison on the development of the Canadian and U.S. banking systems, see Bordo et al. (2015).

There are features of the Canadian banks that make the analysis simpler and more accurate. The Big Six Canadian banks account for account for 89.5% to 93% of total chartered bank assets, so narrowing the cross-section to these banks captures the majority of activity. There is no significant entry or exit of firms which simplifies the analysis considerably. The only merger of note occurred in 2000 and that was between a trust company, Canada Trust, and Toronto-Dominion. Due to its modest size and Canadian accounting practices, it caused little disruption in the data. Each of the five biggest banks had a significant presence across all provinces and a significant international operation. National Bank, the smallest of the large Canadian banks, specializes in serving French-speaking Canadians and has a more modest international presence. This homogeneity among banks reduces the potential bias caused by geographic variation which can be significant in jurisdictions such as the United States where there are thousands of banks of varying size and national coverage. The largest U.S. bank, J.P. Morgan, operated in only fourteen states in 2004 and the second largest, Wells Fargo, operated in just twenty-nine that same year[4]. While the Big

[4]In 2016, Wells Fargo was operating in 39 states.

Six Canadian banks are not global systemically important financial institutions (G-SIFIs), they are large banks by most standards and international institutions recognize them as domestic systemically important financial institutions (D-SIFIs). Five of these Canadian banks would rank as high as the fifth largest U.S. bank by total assets. Stated slightly differently, if the five largest Canadian banks were U.S. banks, then they would rank five through nine by assets in the United States. The smallest bank, National Bank, would be in the top 15. Fifth and finally, the methods used in this paper provide a useful example of estimating RTS in a panel with a short cross-section and long time series.

I test for robustness in a thorough and rigourous fashion. First, a number of alternative model specifications confirm the major result that constant RTS defines the average Canadian bank over the sample. Second, I test whether the estimated cost function satisfies basic microeconomic properties. This analysis is not common in the banking literature, but it is a standard feature of stochastic frontier analysis in the airline and insurance industry. Hughes and Mester (2008) observe that there are two ways to measure bank performance: nonstructural and structural which includes the translog cost function. There is an argument to be made that for it to be an actual structural cost function, it should satisfy the necessary microeconomic requirements: monotonicity in outputs and prices, and price convexity. Otherwise, it could be more appropriately regarded as a reduced form model. If output and prices are increasing, then total cost should increase, and if all prices double, then total cost should double. The estimated short-run cost function satisfies monotonicity for more than 90% of observations. If the first step of the Ryan and Wales (2000) method is employed, then price concavity is satisfied in over 85% of observations. Given data limitations, defining the price of physical capital is problematic. Consequently, I estimate two cost functions: a short-run cost function that omits physical capital expenses and a long-run cost function includes the cost of physical capital. Both give similar results however the long-run cost function fails to satisfy price concavity.

The remainder of this paper is organized as follows. Section 2 explains how this paper fits into the literature on banking returns to scale and Canadian banking. A number of recent papers find increasing returns to scale while others find constant returns to scale. Section 3 explains why I believe a trans-log cost function is the most appropriate model for studying

returns to scale among Canada's largest banks. Section 4 outlines the statistical methods including why this paper uses maximum loglikelihood and a panel dynamic ordinary least squares estimator. Section 5 explains features of the data including why 2011 is chosen as an end date to the analysis and how transitioning from Canadian GAAP to IFRS accounting standards change the balance sheet. This paper estimates a short-run cost function that excludes the cost of physical capital and a long-run cost function that includes it. This is defined and explained in section 6. In section 6.2, I address issues such as identification and omitted variables and how these affect my results. Section 7 shows the results. It ends with a discussion on cost efficiency as measured by noninterest expenses per asset. Section 8 concludes with a brief summary of the results and a discussion on some of the implications.

2 Literature review

This paper contributes to the literature in a number of ways. First, it provides a study of returns to scale and technical efficiency in Canadian banking that is thorough and recent. Second, it contributes to the ongoing debate regarding increasing returns to scale among large or international banks. Third, this paper uses a new data set with a short cross-section (N) and a long time dimension (T). This is a relatively rare form of panel data and this paper uses the most recent techniques from the literature to estimate returns to scale, returns to scope and technical efficiency. See section 4 for further details. Lastly, the Canadian banking system is dominated by six large firms which inherently limits the number of observations available for inference. I posit a parsimonious selection of outputs and inputs to capture the curvature of the cost function. I justify this specification through robustness tests and analyzing the microeconomic properties of the cost function.

Looking at relevant studies on Canadian banking, Beyhaghi et al. (2014) estimate the size of the implicit subsidy in the Canadian banking system. This term stems from the concept of a financial institution being 'Too Big to Fail' or, more accurately, the government views them as being too important to the financial system. One can think of these institutions as receiving an insurance policy with no premium. These unpaid payments are the implicit subsidy. In prosperous periods, one expects the benefit from this insurance

policy to be quite small while in more volatile periods, the benefit is quite significant[5]. Using hand collected information from 1990 to 2010 on bond issues, and after controlling for idiosyncratic and market risk factors, Beyhaghi et al. (2014) find that the Big Six Canadian banks enjoy a 70-80 basis point spread over smaller rivals.

Allen et al. (2014a,b) observe that there was a high amount of price dispersion in the Canadian mortgage market that cannot be attributed to borrower and lender characteristics – search costs and bargaining power play significant factors in setting loan prices. Allen et al. (2013) study horizontal mergers in the Canadian mortgage market from 1993 to 2001 and, focusing on geographic areas with an acquiring branch, they find that the loss of a competitor led to a 6 bps increase in local mortgage rates and there existed a positive relationship between price dispersion and the number of competing firms. This raises an interesting question about the market structure of the Canadian banking system: was the market perfectly competitive, oligopolistic or perhaps a monopolistic competition based on branch-networks? Allen et al. (2013) note that the Big Six banks, Desjardins (a cooperative), and ATB Financial (owned by the province of Alberta) held over 90 percent of mortgages. This was a high concentration among the largest firms however many other competitors – loan companies, small banks, trusts and credit unions – continued to compete with these institutions. Recently, the secondary market for mortgage-backed securities grew significantly. For example, in 2013, the Canadian Mortgage and Housing Corporation (CMHC) purchased and held nearly $270 billion dollars worth of mortgages compared to $950 billion dollars worth of mortgages held by all domestic and foreign chartered banks operating in Canada. This could have given smaller institutions more opportunity to issue and sell, rather than hold, mortgages with an end result of fewer assets on the balance sheet, less risk, and lower costs. Since 2001, there were no major mergers among the large Canadian banks to act as a natural experiment. So while the number of large Canadian banks, six, suggests oligopoly, the number of smaller financial intermediaries suggest that monopolistic competition might be more accurate.

Perez-Saiz and Xiao (2014) study competition and market entry using a simulated method of moments estimator and data on 1,447 rural Canadian areas but excludes any

[5]For more on the implicit subsidy, see Ueda and Di Mauro (2013)

area within 50 kilometres of an urban centre of 100,000 or more people. They compare the 'Big 8'[6] financial institutions against local credit unions. Credit unions were more competitive than the Big 8, and they conclude this was consistent with an oligopoly structure – larger oligopolists were less efficient than smaller local operators. This is an interesting result although the focus on rural banks may be misleading. Their model relied on an assumption of perfect information among firms, yet it would not be unreasonable to think a local operator has superior knowledge of local clients. Their study excluded urban areas where most business activity occurs and, presumably, where the Big 8 would have found the highest returns.

Previous studies on RTS in the Canadian banking system include McIntosh (2002) and Allen and Liu (2007). McIntosh (2002) estimates a profit function using data from 1976-1996 on the 'big five'[7] Canadian banks and he finds increasing RTS of 22 percent. This suggested that a 10 percent increase in assets would lead to a 8.2 percent increase in total costs. His estimates are robust to both a model of Cournot or monopolistic competition. Using a counter-factual simulation, he considers two mergers: Royal Bank of Canada with Bank of Montreal and Toronto Dominion Bank with Canadian Imperial Bank of Commerce. He finds that mergers in Canada were welfare improving: the gains to society from cheaper financial services outweighed the loss to consumers from increased market concentration. Allen and Liu (2007) measures RTS by taking advantage of a cointegrated time series. They use a panel dynamic ordinary least squares estimator (PDOLS) from Kao and Chiang (2001) to obtain more accurate estimates. Using proprietary data from 1982 to 2002, they find RTS of 6 percent. Xiang et al. (2015) study efficiency among a number of Australian, Canadian and UK banks. They believe that a negative relationship between assets and efficiency suggests diseconomies of scale. This reduced form methodology differs considerably from other studies on RTS. They include foreign subsidiaries such as HSBC Canada and a regional operator, Laurentian Bank of Canada, that may have differing cost structures. HSBC Canada may have head office operations located overseas and off its income statement[8]

[6] these include the Big Six Canadian banks, Desjardins (a cooperative) and ATB Financial (a provincially owned crown corporation).

[7] this excludes the smallest of the Big Six, the National Bank of Canada

[8] I exclude HSBC Canada for this reason and because the number of workers employed at HSBC Canada is not available – many operations are likely conducted by employees overseas and it is not possible to

while Laurentian Bank of Canada was previously owned by Desjardins Group.

Outside of Canada, there is ambiguity in the literature about RTS among U.S. and international banks. Hughes and Mester (2013), Anderson and Joeveer (2012) and Davies and Tracey (2014) study RTS among U.S. and international banks respectively. Davies and Tracey (2014) estimates a trans-log cost function and the associated cost shares covering more than one hundred banks and ten years (2001-2010) of annual data. They find that once the implicit subsidy protecting large financial institutions is accounted for, increasing RTS disappeared and constant or small diseconomies of scale were present.

Anderson and Joeveer (2012) consider the rents earned by shareholders and bank employees separately. Using this approach, they find large RTS and the effect was strongest for the largest U.S. banks. However, as Admati and Hellwig (2014) point out, Anderson and Joeveer (2012) fail to account for the implicit subsidy and, uniquely, they include rents captured by employees which is a departure from the standard definition of RTS. Hughes and Mester (2013) estimate two models: an almost ideal demand system (which has a trans-log function nested within) and a manager risk-preference model that accounted for endogenous risk-taking. They use the cost share equations from the ideal demand system to estimate a static system of equations for the years 2003, 2007 and 2010. Significant IRTS were present when risk-taking was endogenous and none otherwise. In those years, the implicit subsidy made little impact on RTS. However Admati and Hellwig (2014) argue that Hughes and Mester (2013) fail to explicitly define the implicit subsidy in the financing of large banks. Furthermore, they note that the periods used by Hughes and Mester (2013) were associated with years of high bank profits, such as 2007, and wonder whether these are representative of a larger trend.[9]. A further critique that was previously mentioned, Kumar (2013) show that when market power is unaccounted for RTS estimates were biased upward that. This might explain why Hughes and Mester (2013) finds such high values for increasing RTS.

There is also a body of work using Bayesian and non-parametric techniques. Feng and Serletis (2010) estimate a trans-log distance function using a Bayesian technique on a sample of 292 American banks from 2000-2005. They find large increasing RTS. Feng and

disentangle the two.

[9] Admati and Hellwig (2014) reference an earlier working paper of the Hughes and Mester (2013) paper. In the published version of the paper, they also estimate their model using 2003 and 2011 data. However the model remains static and the analysis omits troubled periods such as 2001 and 2008.

Zhang (2012) estimate a random effects stochastic distance frontier model using a Bayesian procedure which is related to the Greene (2005) method employed later. RTS was higher among the largest U.S. banks than the smaller U.S. banks. However in their most recent study, Feng and Zhang (2014) consider technological heterogeneity and find no correlation between cost and asset size. Berger and Mester (2003) estimate a bank cost function and two profit functions and determine that a study that fails to account for noninterest income will produce an estimate of RTS that is biased downwards. Wheelock and Wilson (2012) estimate a fully non-parametric model using U.S. bank data from 1984-2006. This has the benefit of being far more flexible than a fully parametric model. They test and reject the trans-log model likely due to the extreme variation in bank size. Using the non-parametric approach, they find significant increasing RTS for nearly all banks at any point in time. Depending on bank size, the increasing RTS is roughly 4 to 7 percent. Higher values are associated with earlier time periods and smaller banks. Restrepo-Tobón and Kumbhakar (2015) identify some serious flaws in their methodology. Using an input-oriented distance function and a more accurate non-parametric measure of RTS, it became smaller on average and some banks operated with constant or even decreasing RTS. Almanidis et al. (2015) estimate a trans-log input-distance function using a semi-parametric spline function and U.S. quarterly data covering 1984 to 2010. Inputs are the level of deposits, capital, and the number of employees while output includes assets. They discover significant increasing RTS the 1980's that slowed over time until it became constant and then mildly decreasing at the end. Restrepo et al. (2013) also use a nonparametric technique and compare this model to other parametric techniques. They find that the estimates of RTS are biased upward when the excess capacity of capital is unaccounted for. RTS are frequently present but not across all banks at all times. For more on the RTS literature and how it relates to the U.S. banking system, see McKeown (2017a).

3 Model

3.1 Intermediary model of banking

In measuring returns to scale and technical efficiency, a necessary first step is to choose an appropriate framework for how to think about the operations of a bank. Specifically, how do we define a cost and an output? I adopt the intermediation approach to banking where deposits are inputs, and assets and fee income are outputs. An alternative approach, put forward by Berger and Humphrey (1992) and known as the value-added approach, considers demand deposits as outputs. However the intermediary approach is preferred by most researchers as it is intuitive: banks match lenders with savers while charging a fee (interest rate) for the service. This also has a conceptual advantage as it is how most people understand the operations of a bank. Furthermore, interest paid is often the largest single expense, and it seems counter-intuitive to consider it an output. Lastly, the data available from OSFI sometimes fails to clearly distinguish between demand and notice deposits, and longer term deposits such as guaranteed-investment certificates (GICs). Consequently, I adopt the intermediation approach. Other possible inputs include equity, borrowing (repurchase agreements and subordinated debt), labour and physical capital. Banks use these inputs to create outputs, namely: loans, marketable securities and fee income from investment banking, wealth management and retail operations.

A cost function captures many of the key features that drive bank profitability. Dietrich and Wanzenried (2011), using data from 372 commercial banks in Switzerland from 1999 to 2009, find that bank profits are driven by operational efficiency (ratio of revenue to cost), the growth of total loans, funding costs, share of net interest to noninterest income, and the effective tax rate. Clearly, managing costs is an extremely important component of bank profitability. To measure RTS, I adopt the technique first proposed by Christensen et al. (1973) and later developed by Kopp and Diewert (1982) that is known as the translog cost function. In short, it incorporates much of the available information but also requires that markets are perfectly competitive. If they are, then duality applies. Namely, that maximizing the profit function or minimizing the cost function would result in the same solution to the firm profit maximization problem. Shaffer (1993) finds that banking

in Canada is consistent with perfect competition, and in a study of 50 banking systems, Claessens and Laeven (2004) find that bank competition and market concentration were not negatively correlated. Bikker et al. (2012) survey the literature and conclude that market concentration was a very poor measure of competition. In 2014, there were 25 domestic banks, 24 foreign subsidiaries and 27 foreign bank branches operating in Canada. This is similar to the number operating in 1996. There are many monoline lenders and credit unions competing with the banks for deposits and loans. While the dominance of the Big Six may lead one to conclude that the market is an oligopoly, as previously stated in section 2, the large number of smaller banks make this conclusion less certain. Goddard and Wilson (2009) find that previous fixed-effect estimates of bank competition were biased toward concluding a banking sector was monopolistic. They use an Arellano and Bond (1991) GMM estimator and the H-statistic of Panzar and Rosse (1987) to test for bank competition in G7 economies. It measures the elasticity of revenue income to a change in the cost of inputs. If a market is perfectly competitive and input prices increase, then revenues would increase proportionately and the H-statistic would equal 1. Under a monopoly and in response to an input price increase, output would decrease and the H-stat would be less than zero. Goddard and Wilson (2009) reject the null hypothesis that the Canadian banks are perfectly competitive and that they are monopolists. This leads them to conclude that the Canadian banks had a H-stat between zero and 1 – that is the definition of monopolistic competition or oligopoly[10]. Bikker et al. (2012) criticize this result. They show that using a price-equation or scaled revenue function will generate an invalid measure of competition. Shaffer and Spierdijk (2015) argue that considerable market power may exist even if the H-stat is greater than zero. This is robust to the timing of bank actions in response to a change in input prices, relative costs, degree of product differentiation, and a number of other factors. If true, this raises serious concerns about a large body of literature.

A degree of market power in the Canadian banking system is likely. The Big Six banks in particular emphasize relationship banking where banks collect information on customers and use that information to sell them a wide-range of products. For more on relationship

[10] According to Shaffer and Spierdijk (2015), researchers remain uncertain about how to interpret a H-stat between 0 and 1.

banking[11] and bank decision-making processes, see Berger et al. (2014) and Greenbaum et al. (2015). Allen (2011) finds that consumers have different preferences for financial services and this allows the banks market power over mortgage rates. He finds that high-income households and loyal customers pay higher mortgage rates while search costs and bargaining play a role in the final transaction price. If a translog cost function is estimated and the market is not perfectly competitive, then I expect to observe banks operating away from the point of minimal cost. There should be observations with significantly increasing or decreasing RTS. There is also a probability that the model itself is misspecified, and market power is an omitted variable.

Considering the drawbacks of a translog cost function, a common alternative is the input-oriented distance function. The setup is similar in that both require output decisions be exogenous however its main advantage is that it does not require any assumption on the competitive market structure of the industry. This is advantageous however the drawback of this model is that it fails to make full use of all available information. The input-oriented distance function uses output and input quantities while the trans-log cost function also uses total costs and expenses. In theory, if the market is perfectly competitive, then the input-oriented distance and trans-log cost function should produce identical results. For more detail on the input-oriented distance function, see Kumbhakar et al. (2015). In McKeown (2017a), I estimate an input-oriented distance function and find similar estimates of RTS that may suggest that market power is not significantly affecting the translog RTS estimates.

3.2 Trans-log cost function and returns to scale

Consider the firm's cost minimization problem in equation (1).

$$\min \sum_{j=1}^{k} W_j X_j(Y, W) \tag{1}$$

[11]For example, a self-employed borrower is considered much riskier than a full-time employee, all else being equal. However if a borrower and her business use the same bank to process transactions than this provides the bank with exclusive and private information about the borrower's income and viability of her business. This allows the bank to undercut its competitors' offers while capturing risk-adjusted rents.

subject to a production constraint:

$$F(Y, X) = 0 \tag{2}$$

where Y is a vector of m output quantities, W is a vector of j prices and X is a vector of j inputs. To derive the trans-log cost function, substitute the production constraint into (1), take the logarithm of each side of the equaiton, and apply a second order Taylor Series expansion. The result is equation (3).

$$c = \alpha_0 + \sum_{q=1}^{m} \alpha_q y_q + \sum_{j=1}^{k} \beta_j w_j + \left(\frac{1}{2}\right) \sum_{q=1}^{m} \sum_{w=1}^{m} \sigma_{qw} y_q y_w + \sum_{w=1}^{m} \sum_{j=1}^{k} \gamma_{wj} y_w w_j + \left(\frac{1}{2}\right) \sum_{p=1}^{k} \sum_{j=1}^{k} \delta_{pj} w_p w_j + \epsilon_{it} \tag{3}$$

where all lower case variables are in natural logarithms, α_0 is a constant and ϵ is the error term. There are m outputs and k prices and the time subscripts are suppressed for simplicity. In order for the trans-log cost function to be well-behaved, it is necessary to impose some restrictions. First, input demand should be downward sloping such that an increase in input prices reduces the use of that input when possible. Second, cross-price effects are symmetric ($w_l w_j = w_j w_l$). Third, the sum of own and cross-price elasticities must equal zero. And fourth, if output is held constant, a proportional increase in all input prices shifts cost similarly. For example, a 5 percent increase in each input price would increase total costs by 5 percent.

The following explanation follows Kumbhakar et al (2015) closely. In order to ensure price homogeneity, the following restrictions are imposed on equation (3):

$$\sum_{j}^{k} \beta_j = 1 \qquad \sum \gamma_{wj} = 0 \qquad \sum \delta_{lj} = 0 \tag{4}$$

and with symmetry imposed:

$$\delta_{lj} = \delta_{jl} \qquad \sigma_{qw} = \sigma_{wq}$$

Applying the price homogeneity and symmetry restrictions above directly to (3) produces the following equation for estimation:

$$log\left(\frac{C}{W_1}\right) = \alpha_0 + \sum_{q=1}^{m} \alpha_q y_q + \sum_{j=2}^{k} \beta_j \left(w_j - w_1\right) + \frac{1}{2}\sum_{q=1}^{m}\sum_{w=1}^{m} \sigma_{qw} y_q y_w$$

$$+ \sum_{q=1}^{m}\sum_{j=1}^{k}\sum_{j\neq q} \gamma_{qj}\left(y_q w_j - y_1 w_1\right) + \frac{1}{2}\sum_{p=1}^{k}\sum_{j=2}^{k} \delta_{pj}\left(w_p w_j - w_1 w_1\right) + \epsilon_{it} \tag{5}$$

With the restrictions applied, some coefficients need not be estimated directly because their values can be inferred. Following any estimation, the price restriction is tested and the coefficients are used to determine if the cost function is concave in prices and monotonic in output and prices. Price monotonicity requires that:

$$\frac{\partial\ ln\ C}{\partial\ ln\ W_j} = \left[\beta_j + \sum_q \gamma_{qj} y_q + \sum \delta_{jp} w_p\right] > 0 \quad \forall\ W_j \tag{6}$$

and similarly output monotonicity requires that:

$$\frac{\partial\ ln\ C}{\partial\ ln\ Y_q} = \left[\alpha_q + \sum_j \gamma_{qj} w_j + \sum_w \sigma_{qw} y_w\right] > 0 \quad \forall\ Y_q \tag{7}$$

Diewert and Wales (1989) show that a sufficient condition for price concavity is if the Hessian Matrix of input prices is negative semidefinite. Each element of this Hessian matrix is:

$$\frac{\partial^2 C^*}{\partial\ \mathbf{w}_i\ \partial\ \mathbf{w}'_j} = \frac{\partial^2 ln\ C^*}{\partial\ ln\ \mathbf{w}_i\ \partial\ ln\ \mathbf{w}'_j} - diag(\mathbf{S}_i) + \mathbf{S}_i\mathbf{S}'_j \tag{8}$$

where any price share is equal to:

$$\mathbf{S}_j = \frac{\partial\ ln\ C*}{\partial\ ln\ \mathbf{w}_j} = \beta_j + \sum_{q=1}^{m} \gamma_{qj} y_q + \sum_{p=1}^{k} \delta_{jp} w_p \tag{9}$$

To measure RTS from the above, I calculate the cost elasticity with respect to output which is also used in Davies and Tracey (2014), Allen and Liu (2007), and others. Holding input prices constant, if each output is increased by 1 percent and total costs increase by more than 1 percent, then there are decreasing RTS. Similarly, if total costs increase by less than 1 percent, then there are increasing RTS, and if the proportionate changes are equal, then there are constant RTS. The elasticity is calculated as the partial derivative of equation 5 with respect to each output (y_l) and summed across each output. This generates the elasticity measure of RTS in equation 10. Although the time subscripts are suppressed for simplicity of reading; it should be noted that RTS are measured for each cross-section in each time period. To generate a comparable measure of RTS and to summarize the results, I state RTS as an average of all observations. Alternatively, I also calculate RTS at the average of outputs and input prices but this could give a misleading result if there are no actual observations at this data point.

$$\left(\sum_{q=1}^{m} \frac{\partial\ log(C)}{\partial\ log(Y_q)} \right)^{-1} \tag{10}$$

Economies of scope are defined as the relative cost savings associated with complementary products (for example, consumer loans and mutuals fund fees). The expression is nested within the definition of RTS. If all outputs increase by 1%, the corresponding cost savings through offering multiple products is:

$$\frac{\partial\ ln(C)}{\partial\ ln(Y_q) ln(Y_w)} = \frac{\partial\ log(C)}{\partial\ log(ln(Y_q) ln(Y_w))} \tag{11}$$

If the above was less than one then there are economies of scope present.

3.3 Expansion-path returns to scale (ERTS)

Using the estimated coefficients, expansion-path returns to scale (ERTS) illustrates how RTS varies with output and input prices. This allows us to observe the curvature of the cost function and separately illustrate the impact of prices and output on RTS. Furthermore, it allows us to observe whether RTS is increasing, decreasing or constant as output increases at fixed prices. This will allow comparison to Wheelock and Wilson (2012) who find RTS

that are increasing with bank size. Consider equation 12:

$$ERTS(\gamma y, w) = \left(\sum_{q=1}^{m} \alpha_q + \sum_{q=1}^{m} \sum_{j=1}^{k-1} \delta_{qj} w_{j,o} + \sum_{q=1}^{m} \sum_{q=1}^{m} \sigma_{qg} \gamma y_{g,o} \right)^{-1} \tag{12}$$

If ERTS is greater than 1, then there will be decreasing RTS. Similarly, if ERTS is less than 1, then there will be increasing RTS. The subscript o denotes the median observation. Emanating from the median observation, γ takes on a range of values that generate a series of outputs that correspond to the 5th and 95th percentile of the sample. For example, if the media observation of y is 10, the 5th percentile is 9, and the 95th percentile is 11, then γ would take on a range between 0.9 and 1.1. This corresponds to $(\gamma_{min})(y_o) = 0.9$ and $(\gamma_{max})(y_o) = 1.1$, and RTS can be computed within this range and graphed. Of course, it is entirely possible that no bank in the sample actually produces these combinations of output. As previously discussed in section 2, Wheelock and Wilson (2012), using a nonparametric equivalent of the ERTS in equation 12 that they call ray-scale economies, find extremely high RTS for banks of every size. They attribute this to the hypothetical nature of the product 'mix' which they believe might generate erroneous results. However using a similar definition, I do not observe extreme estimates of RTS. Rather, they appear to be relatively flat which implies a similarly flat average cost curve.

3.4 Cost of Equity

Canadian banks, like their international counterparts, set return on equity and dividend goals every year. Payout ratios in Canada are often above 40 percent of after-tax earnings. If a bank uses more equity in its funding mix, then more profit is necessary to achieve its desired ratio. To minimize funding activities through equity, banks undertake expensive risk management projects to minimize risk weighted assets (RWA) and convince regulators that they are safe institutions. The Big Six operated with a leverage ratio between 12 and 23 percent over the sample period.[12] The literature on dividend policy is vast, see

[12]The Canadian office of the superintendent of financial services set the maximum ratio at 20, but they reserve the right to reward 'good' banks by increasing this ratio and punish 'bad' banks by lowering it. Prior to the financial crisis, leverage ratios were close to the ceiling of 20 however following the crisis some banks decreased the ratio to as low as 12. Guidara et al. (2013) argue that this suggests banks are responding to market-influenced discipline, they want to show investors they are low-risk. Alternatively when banks had these low leverage ratios, it was just prior to the transfer to IFRS which required additional equity, so the

Bhattacharyya (2007) for a relatively recent review. Canadian banks have been paying dividends for a long time, in some cases for nearly 200 years, and it is a reasonable assumption that they plan to continue to do so. Hughes and Mester (2013) rightly observe that failing to account for the cost of equity leaves any cost function mis-specified and implies that a theoretical 100 percent equity firm would pay zero dollars for funds. It is not the purpose of this paper to explain dividend policy however I include a measure of the cost of equity into the cost function.

Given that each bank in the sample is publicly traded, a natural model to estimate the required return on equity is the capital asset pricing model (CAPM) using monthly stock returns. CAPM suggests that the return on equity for any share is a function of its exposure to a common systemic risk factor and that any non-systemic (idiosyncratic) risk goes unrewarded. This relationship is shown in equation 13 where the expected market premium ($E\left(r^m - r^f\right)$) represents the common systemic risk factor. Alternatively, the Fama and French (1993) factor model uses three common factors: the expected market premium, a small capitalized firm premium and a book-to-market value premium. However CAPM has some advantages: (i) it is theoretically founded and can be simply derived; (ii) the risk-free rate and the expected market premium are more intuitive than the Fama-French factors. Regardless, either method will generate similar estimates of the required return.

$$r_i^e - r^f = \beta_i E\left(r^m - r^f\right) + \epsilon_i \tag{13}$$

Taking equation 13 to the data: r_i^e is the market return on shares of bank i, β_i is the CAPM measure of risk for bank i and r^m is the return on the S&P TSX Toronto Stock Market Index or the U.S. S&P500. ϵ_i is a random error term representing idiosyncratic risk. The Canadian banks are some of the largest companies on the Toronto Stock Exchange. As of writing and by market capitalization, four of the top five largest Canadian companies are Big Six banks. To avoid possible endogeneity, I use the U.S. S&P500 as an instrument for the TSX in a two-stage least squares procedure, but the results are similar whether the Toronto or New York stock index returns are used. Observation frequency is monthly.

banks might have been preparing for this transition.

This is ideal for estimating the cost of equity: higher frequency observations (weekly, daily) may be biased by the liquidity of the trading book and large individual transactions. See Da et al. (2012) and Bruner et al. (1998) for further details. The CAPM $\beta_{i,t}$ is estimated for each bank separately using OLS and data from the previous 60-months worth of three-month t-bills, the index return and the stock return for bank i. A rolling average of 5 years, or 60 months, is estimated for each bank month from November 1995 to October 2011 and the quarterly average is calculated. The estimated equation is:

$$\hat{r}^{e}_{i,t} - r_{f,t} = \hat{\beta}_{i,t} E\left(r_m - r_{f,t}\right) \tag{14}$$

Given the estimates of β_i, two calibrations are required. First, what is the appropriate risk-free rate and second, what value should the expected market premium take? There are many available choices for choosing this value. Bruner et al. (1998) have a survey asking financial firms, corporations and academics how they choose. Common methods include: historical means, spreads above t-bills, or a fixed amount. I choose to set the expected market premium to 4% (or 1% per quarter). This is a relatively modest premium however it is quite a bit higher than the actual premium received. Over the sample, market returns are low and this implies that the cost of equity should also be low - expectations likely reflect recent history. Another consideration is that firms pay tax on equity while they do not pay tax on interest expenses. This would presumably inflate the cost of equity relative to debt. To maintain simplicity, I assume a 4 percent expected annual market premium and no tax implication against a model with the cost of equity omitted. Given these conservative assumptions, there is little evidence that excluding the cost of equity meaningfully changes the result. Compared to the required return on book equity, the CAPM estimates are much smaller. For example in 2006, RBC had a beta coefficient of 1, the required return was 4 p.a. plus the 3-month t-bill rate of 4 p.a. led to a required return of 8.17 p.a. For the year following, 2007, RBC management set a target book ROE of more than 20 percent and a diluted earnings per share of more than 10 percent.

4 Statistical methods

Three estimators are used to measure returns to scale, returns to scope and technical inefficiency. The first model is the fixed effect OLS (abbreviated as FE) using dummy variables to capture the time invariant heterogeneity of each cross-section. The fixed effect model is a common technique with panel data however given the long 'T' dimension of the pane, I follow Allen and Liu (2007) who use the panel dynamic ordinary least squares estimator (PDOLS). In order to test for inefficiency, this paper uses the Greene (2005) 'true' fixed effect model (TFE) to identify and test for technical inefficiency. The following sub-sections explain.

4.1 Panel data with large 'T' and short 'N'

If the series are co-integrated then the simulation tests performed by Kao and Chiang (2001) apply. They find that the OLS estimates are biased downwards in finite samples which, given our definition of RTS, implies it would be biased upwards. This allows us to define the fixed effect OLS estimates as a ceiling on actual RTS. In order to improve the finite sample estimates, the panel dynamic ordinary least squares estimator (PDOLS) of Kao and Chiang (2001) and Mark and Sul (2003) is applied. PDOLS uses leads and lags of changes in the independent variables to more accurately estimate coefficients. These leads and lags reduce bias from any possible endogeneity between total costs, output and prices. To test whether the trans-log cost function defined in equation 5 is cointegrated, this paper applies the modified augmented Dickey-Fuller (MADF) test developed by Sarno and Taylor (1998) and the Levin-Lin-Chu test Levin et al. (2002) to test whether the residuals are stationary. As the results in section 7 show, the null hypothesis that none of the series is cointegrated is rejected and the null hypothesis that the residuals are non-stationary is rejected.

Given that the panel is cointegrated, this paper estimates the trans-log cost function using the panel dynamic ordinary least squares (PDOLS) estimator. The Mark and Sul (2003) PDOLS estimating equation is:

$$y_{it} - \bar{y}_i = (X_{it} - \bar{X}_i)'\beta + \sum_{j=-q_i}^{q_i} b_j \Delta X_{it+j} \tag{15}$$

where i represents each bank. For each observation, the dependent and exogenous variables are de-meaned. Lead and lag changes in the exogenous prices and outputs are represented by the ΔX terms. One may interpret the β coefficients as being the long-run coefficients while short-run fluctuations are absorbed by the b coefficients. Given that the sample period has a length of 64 observations, the recommendation from Mark and Sul (2003) is to include one lead and one lag in the estimation so that $q_i = 1$. This of course means that the first and last observation in each time series are fixed.

4.2 Technical inefficiency

A common measure of technical inefficiency is called the distribution-free approach. This technique exploits the residuals from estimation to rank observations according to how effectively a firm is able to produce outputs. In a cost function, the residual of each observation is measured against the observation with the smallest residual and it is known as the 'best practices' observation. The absolute or relative distance (depending on the function in question) from each observation to this observation is measured and summarized. Sometimes time trends and fixed effects are also included. The distribution-free approach to measuring technical efficiency is defined as:

$$\begin{aligned} u^* &= u_{t,i} - min(\hat{u}) \\ efficiency_{i,t} &= exp(u^*) \end{aligned} \tag{16}$$

where u is the estimated residual, u^* is the relative distance from best practices. This approach informs the practitioner about any unexplained variance in the residuals, but it offers no indication whether the errors are systematic or predictable. The measure is heavily influenced by the 'best practices' observation which is, by definition, an extreme value. This can make inference difficult. For example, consider a situation where the average relative distance from any residual to the 'best practices' residual is 80%. By this method, one would conclude that the average technical inefficiency in the industry is 80% even if there are no other residuals better than 81% distant from the smallest residual. Finally, while

the measure might be interesting it has no testable implications.

In order to test for technical inefficiency, this paper makes an explicit assumption on the error terms and estimates the transcendental log cost function using a maximum log-likelihood estimator. This method is attributed to Greene (2005). A standard assumption on the error term is that it can be broken into two parts: a random, white noise component and a predictable systematic component that has a half-normal distribution. This captures the unconditional mean of technical inefficiency.

$$ln\ C^a = ln\ C^*(\mathbf{w}, \mathbf{y}) + \eta + \nu \tag{17}$$

$$\eta \sim N^+(0, \sigma_u^2) \tag{18}$$

$$\nu \sim N(0, \sigma_\nu^2) \tag{19}$$

and the log-likelihood function for any observation becomes:

$$L = -ln\left(\frac{1}{2}\right) - \frac{1}{2}ln\left(\sigma_\nu^2 + \sigma_u^2\right) + ln\phi\left(\frac{-(\nu+\eta)}{\sqrt{\sigma_\nu^2 + \sigma_u^2}}\right) + ln\Phi\left(\frac{\mu}{\sigma}\right) \tag{20}$$

with

$$\mu = \frac{\sigma_u^2(\eta+\nu)}{\sigma_\nu^2 + \sigma_u^2} \tag{21}$$

$$\sigma = \frac{\sigma_\nu \sigma_u}{\sqrt{\sigma_\nu^2 + \sigma_u^2}} \tag{22}$$

This half-normal error term accounts for inefficiency while a normal distribution error term accounts for random noise. I estimate a fixed effect model by OLS and summarize the residuals. If the residuals display skewness, then a half-normal or exponential inefficiency term is appropriate. An additional advantage of the ML estimator is it can estimate error terms as functions of inefficiency-explaining variables (for example impaired assets or physical capital) without introducing these variables – and possible multicollinearity – into the cost function. However as the results show, there is neither much evidence for technical

inefficiency nor did I find any inefficiency-explaining variables.

Greene (2005) explains that when using the fixed effects model, it is common practice to consider fixed effects as time-invariant and firm specific inefficiencies. If for a period of time, one cross-section has a cost advantage over another and a time invariant dummy variable is present, then Greene (2005) argues this may be misinterpreted as inefficiency. In order to overcome this problem, he suggests using a maximum loglikelihood estimator with time invariant firm-specific dummy variables to capture heterogeneity. In order to test for technical inefficiency, a loglikehood ratio test (LR-Test) between the TFE and FE models is performed. This generates a test for technical inefficiency that is presented and discussed in section 7.4.

4.3 Monotonicity and price concavity

A well-defined cost function should exhibit good microeconomic properties such as price concavity and monotonicity. Chua et al. (2005), writing about returns to scale in the airline industry, note that inferences regarding cost functions that fail to satisfy the proper curvature requirements may be misleading. They follow the procedure from Ryan and Wales (2000) to impose local concavity which greatly improves the number of observations that satisfy concavity in input prices. This technique is designed to apply to a seemingly unrelated regression (SUR) with cross-equation constraints. Due to the large 'T' and small 'N' panel, I choose instead to use the PDOLS estimator. However the first step in this procedure is to normalize all observations by one observation. Any one of them is a viable candidate to be normalized, so N x T cost functions are estimated, and each has a different observation as the normalizer. Then the results are compared to the unmodified data. Section 7.7 shows how applying this first step of Ryan and Wales (2000) significantly improves the microeconomic properties of the estimated cost function.

5 Data

Accounting data are from the Office of the Superintendent of Financial Services (OSFI), bank financial statements, Compustat and CANSIM. For a detailed summary of the OSFI

data, see McKeown (2017b). Income statement items are available quarterly while balance sheet data is monthly which is then averaged into quarterly values. Balance sheet and income statement items are converted into real 2002 Canadian dollars using the Canadian CPI index available from CANSIM. The sample period covers the first fiscal quarter of 1996 and ends in the last quarter of 2011.[13] This creates a time series of 64 quarters, six banks and a maximum of 384 observations. In order to estimate the return on equity, monthly stock market index returns on the TSX and the S&P500 and monthly stock returns for the six banks are taken from Datastream covering January 1990 to December 2011.

There are five main advantages to focusing this study on the largest Canadian banks. *First*, these banks offered the same broad product lines (mortgages, loans, wealth management services, capital markets) during the entire sample period. Smaller Canadian banks typically did not offer a full suite of services, so I assume that the cost structure of these banks differ. Smaller Canadian banks predominantly specialize in retail banking, mortgage loans or investment banking but do not offer a full suite of services. Many of them offer discount services or online banking. Furthermore, there was more merger activity among smaller banks with many acquired by a Big Six bank. *Second*, the Big Six were of similar size: the National Bank of Canada being the smallest ($163 billion total assets in 2011) while the Royal Bank of Canada was the largest (with over $780 billion total assets in 2011). The ratio of largest to smallest was approximately $4:1$ over the entire sample period. The Big Six banks represented between 85% and 93% of total bank assets over the sample. *Third*, all these banks operated across Canada. If they had similar branch networks, this limits some of the concerns over market power. None of the small Canadian banks could offer the kind of branch and bank machine network that the largest banks did[14]. *Fourth*, if there is extreme variation in output among the cross-sections, then a parametric estimation would have difficulty capturing a u-shaped long-run average cost curve. This is especially true if these firms have a different mix of services and input costs. A more complex function, such as the Fourier approximation, or a non-parametric estimation technique may

[13] In Canada, fiscal years end on October 31st with the first fiscal quarter of the following year ending on January 31st.

[14] Laurentian Bank of Canada had a branch in most provinces but only in Quebec did they have multiple branches, particular focus in Montreal. HSBC Canada is in a similar position as they focus on the major population centres such as Vancouver. ATB Financial is a significant mortgage lender and operated in Alberta. Desjardins is a large credit union with a broad branch-network in Quebec.

be appropriate however both require a higher number of parameters to estimate and the number of observations available is limited. *Lastly*, restricting the sample to the largest banks guarantees that all the banks are Too Big to Fail (TBTF). Beyhaghi et al. (2014) finds that in Canada, the largest banks enjoy a 70-80 basis point funding advantage over smaller banks. If each bank in the sample benefits from an implicit guarantee, this makes the cost of funding readily comparable.

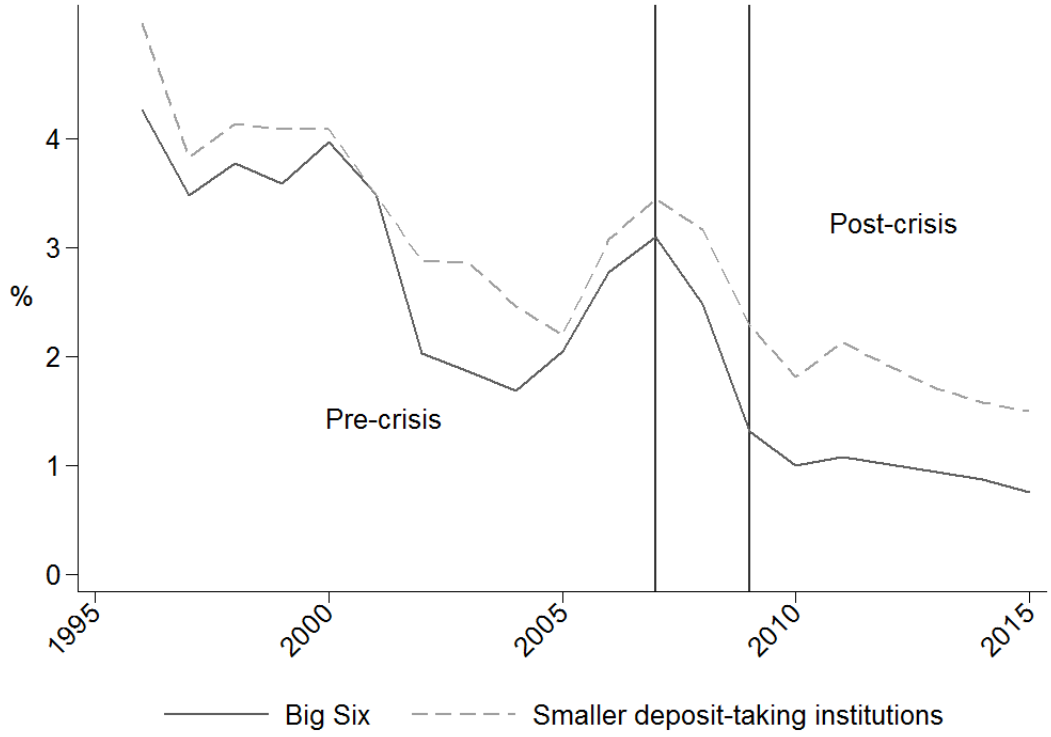


Figure 3: Average interest expense per asset

Annual interest expense divided by average annual assets. The smaller banks include a selection of seventeen chartered banks, and 2 trust companies. The list is not exhaustive. Source: OSFI.

A single observation is identified as an outlier. In the third fiscal quarter of 2006, TD bank sold a business for the sum of $1.41 billion dollars. This is attributed to noninterest income which is 150% greater than in any previous quarter. Consequently, I remove this one-time sale from noninterest income so that the original value of $2.68 billion is replaced with $1.27.

5.1 Regime change and a structural break

As stated in the introduction, I choose to end the sample period in 2011 for two overarching reasons: first, there appears to be a structural break beginning in 2011 and second, there is an accounting regime change from 2011 to 2012. The regime change significantly increases the amount of assets on bank balance sheets[15]. Adjusting the assets and equity solves some,

[15]Under Canadian GAAP, any mortgages sold to the Canadian Mortgage and Housing Corporation is removed from the balance sheet despite the banks retaining some liability if the sold mortgages default. IFRS

but not all, of the comparability issues. As securitization income is included in noninterest income, failing to properly account for this change could result in double-counting outputs.

The majority of the data is calculated under Canadian GAAP (1996-2011), so transforming the IFRS accounting data (2012-2015) into a Canadian GAAP comparable would extend the time series. Kelly and Janssens (2012) calculate the change in equity and assets resulting from this transition which can be attributed to three main causes. *First*, banks often retain a fraction of the risk when they sell securitized mortgages and credit cards. Consequently, even if they have been sold, IFRS requires banks keep these assets on the balance sheet. The largest purchaser of securitized mortgages is the Canadian Mortgage and Housing Corporation (CMHC)[16] which purchases these mortgages under the National Housing Act (NHA). *Second*, IFRS uses a more stringent qualitative assessment of whether a special purpose entity (SPE) may be held off balance sheet. The result is more assets and liabilities on the balance sheet. *Third*, there are material changes in how minority equity interests are reported that decrease balance sheet equity. The totality of these changes are summarized in table 2. To reconcile data after the transition to IFRS, the asset adjustment could be multiplied by total bank assets, then this number could be subtracted from the combined total of consumer and mortgage loans. Unfortunately, over time, this adjustment is bound to become more and more inaccurate as banks change the composition of their asset portfolios.

Additionally in 2012, a period of rapid balance sheet expansion began at the same time that the bank rate and the cost of deposits fell to a historical low. This contradicted previous lending patterns associated with the business cycle. In the 1996-2011 period and earlier, banks typically increase balance sheet assets when deposit rates and total costs are increasing (the expansionary phase of the business cycle), then assets level-off or slightly decrease when deposit rates and total costs are decreasing (the contractionary phase of the business cycle). The change in behaviour, namely an increase in lending commensurate with a decrease in rates, creates a structural break in the estimated parameters. The relationship between total cost and the weighted average cost of capital before 2012 differs

rules are stricter and these mortgages must remain on the balance sheet. For a more detailed explanation, see Kelly and Janssens (2012)

[16]CMHC is a Crown corporation and as such any profits or losses belong to the federal government of Canada.

afterward. From 2011 to 2015, the weighted average cost of capital, excluding equity, declines by more than 20% while total loans, leases and securities increase 38%. See figure 4. For this study, I choose to end the sample after the fourth quarter of 2011. In McKeown (2017a), I use dummy variables to extend the analysis through 2015 for comparability with U.S. commercial banks. Average RTS over the entire sample is unchanged. However when average RTS per fiscal year is calculated, RTS is increasing at an increasing rate post-2012.

6 Model specification

In the intermediation approach to banking, inputs include funding (deposits, bonds, equity), labour and physical capital. Correspondingly, the trans-log cost function requires input prices be exogenous. The price of labour is defined as total compensation expense per full-time equivalent employee which is standard in the literature. With equity, I consider three cases: i) equity is grouped with other liabilities to create the weighted average cost of capital, ii) equity is considered as a separate input which increases the number of right-hand side coefficients to be estimated, and iii) equity is removed from the estimation altogether. One reason to include equity with other forms of financing is that equity, once acquired, is a perfect substitute for deposits, so this approach forces banks to maintain their leverage ratio – a behaviour observed in a number of empirical studies. For example, see Berger et al. (2008) and Gropp and Heider (2010). In the model with equity separate from other funds, a bank can simply choose the cheapest source of financing available; this favours banks that operate with more leverage. Illes et al. (2015) argue that the bank rate, being less costly than the weighted average cost of capital, is a poor approximation for the marginal cost of bank funding hence they focus on the weighted average cost of liabilities. I take a similar approach.

In order to study how physical capital affects RTS and efficiency, two models are estimated: a short-run cost function that includes labour expense, interest expense and the implied cost of equity, and a long-run cost function that includes these variables plus physical capital. Theoretically, the short-run cost function should be nested in the long-run cost function. However, since the dependent variable for each model differs, the estimated

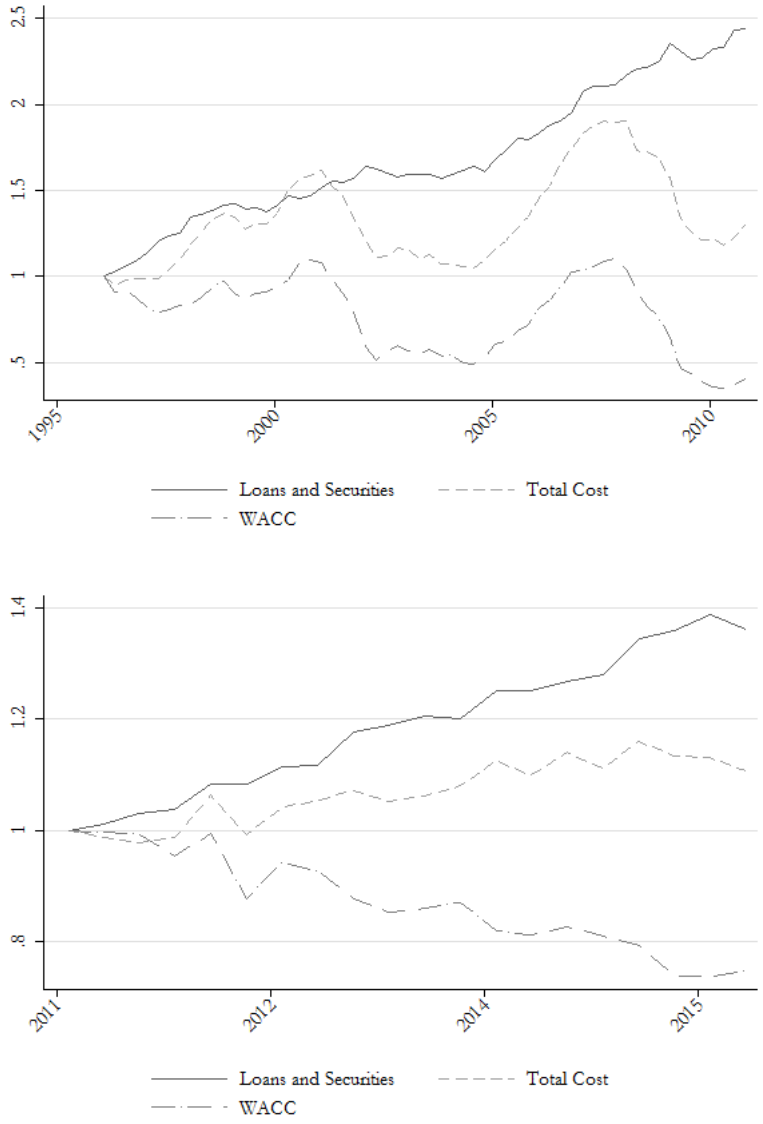


Figure 4: The relative relationship among key variables

The variables are in real 2002 Canadian dollars and have been normalized by the first observation in each sample period. The figure on the left covers the 1996-2011 period. Total cost includes labour, physical capital, and interest expenses. The cost of funds is total interest expense divided by the sum of deposits, repurchases agreements and subordinated bonds. Notice that WACC and total cost moved in opposite directions from 2011 to 2015 while prior to 2011 they moved together.

parameters will likely differ as well. Importantly, a further reason to estimate both of these cost functions is that previous studies sometimes use a mismeasured price of physical capital that may confound the estimates. Partly for this reason, Restrepo-Tobón and Kumbhakar (2015) prefer the input-oriented distance functions to the trans-log cost functions. See Restrepo-Tobón and Kumbhakar (2015) and section 6.1 for an explanation. It is primarily of interest to researchers interested in estimating production and cost functions with physical capital and under similar data constraints. The general reader may skip this without loss and continue from section 6.2.

Outputs are collected into three variables: government and business assets, loans to household and fee income less account fees.[17] This parsimonious specification has the benefit of limiting the number of parameters required for estimation, and it avoids potential multicollinearity among the square and cross-product output terms. To facilitate this, loans and securities from government and business are grouped together and remain separate from loans to households that include both residential mortgages and consumer loans. The remaining output term represents noninterest income which is defined in 3 but includes all noninterest income less gains and losses from trading assets, gains and losses from non-trading assets, and retail account fees. Clark and Siems (2002) find that retail account fees distort noninterest income upward because bank managers have already accounted for these revenues when the deposit price is determined; to include them would be a form of double-counting. This paper follows Clark and Siems (2002) and Davies and Tracey (2014) to exclude these fees although robustness tests do not suggest that it has a particularly strong influence on the results. Clark and Siems (2002) also finds that using noninterest revenue is a superior measure to both credit equivalent assets and the Boyd-Gertler asset from Boyd et al. (1994). The Boyd-Gertler Asset transforms revenue into a theoretical asset and it is defined in section 7.1.

For two reasons, I choose to remove trading gains and losses as a measure of output. *First*, the trans-log cost function is not able to capture the relationship between these and total costs through any available independent variables. Trading revenue is among other things a function of the risk-appetite and ability of bank employees, and neither of these

[17] See table 3 for a complete list and detailed description.

are directly observable. If two banks have identical balance sheets and similar expenses, then it remains entirely possible that one hedges to nearly eliminate exposure while the other is taking a position to double its exposure. Calculating correlations, I find that when trading revenue is positive, the relationship between trading revenue and total cost is weak; when trading revenue is negative, then it becomes non-existent. It is unlikely that a cost function can capture this behaviour. Hughes and Mester (2013) attempt to solve this problem by modelling the risk appetite of bank managers however it is not certain that they have done so. See Kumar (2013) and Admati and Hellwig (2014) and the discussion in section 2. Furthermore, this technique is not feasible in a time-series, and there are too few observations for a cross-sectional analysis. *Second*, in the sample of the 384 observations from 1996 to 2011, the sum of trading revenues and gains/losses on securities not held for trading had 61 observations with negative values. The Canadian Imperial Bank of Commerce suffers such high losses in 2008 that total noninterest income is negative! Only considering positive trading income or setting losses equal to zero eliminates this problem, and this may create upward bias in the RTS estimates. Robustness tests suggest that even if the positive values are added to noninterest income and negative values are discarded, then RTS estimate do not change significantly. Trading revenues are a relatively small component of total income; they infrequently account for more than 10 percent of the sum of net interest and noninterest income. This is illustrated in figure 5.

6.1 The price of physical capital

In the literature, a frequent measure of the price of capital has land, building and equipment expenses in the numerator and total physical capital in the denominator. However this is a problem for a number of reasons. OSFI provides the net capital asset value (book price less accumulated depreciation) rather than the gross asset position of each bank. This implies that over time the price of capital, depreciation[18] plus rent divided by net capital assets, is constantly rising: if the numerator remains constant while the denominator is decreasing, then the price is increasing as depreciation accumulates. Even if the bank purchases no new capital, then this measure will have an upward drift over time. The high price of physical

[18]Over the sample, banks use either the straight-line or both straight-line and double-declining method.

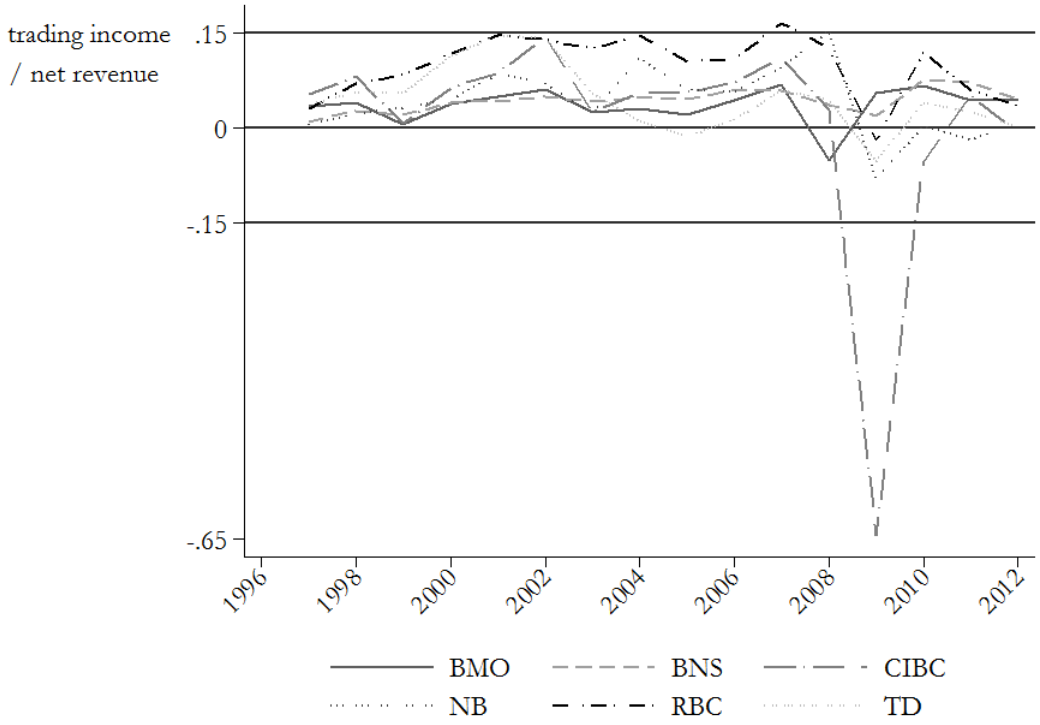


Figure 5: Trading income ratio

The ratio is calculated as realized and unrealized trading gains and losses divided by the sum of net interest and noninterest income excluding trading income and gains/losses on non-trading securities. Prior to the financial crisis of 2007-'09, trading revenue stayed within 0% and 15% of net revenues. CIBC suffered severe losses during the crisis related to their U.S. operations. Prior to the crisis, RBC had the largest share of income from trading operations. However, since the crisis, the net effect has been for trading revenue to decrease as a share of income.

capital makes the firm appear to be more efficient than it actually is. Similarly, each new purchase causes the price of physical capital to decrease. This has an unintended consequence – a bank appears less efficient since its cost is unchanged but the exogenous price is lower. Among the Big Six banks, physical capital purchases are large, infrequent and irregular. Physical capital predominantly includes land, buildings, computer equipment, furniture and leasehold improvements including works-in-progress. As an example, the balance sheet for Royal Bank of Canada on October 31, 2011 showed computer equipment with a book value of $1.494 billion, accumulated depreciation of $1.092 billion, and a net carrying value of approximately $402 million. Meanwhile, the value of buildings was $1.275 billion with accumulated depreciation of $0.456 billion and a net carrying value of $819 billion. In this example, net physical capital overweights land and buildings while under-weighing

computer equipment,[19] the latter of which had a shorter useful life and depreciated more quickly. A possible solution is to replace net capital assets with book value. If banks apply the straight-line method of depreciation and all expenses are capitalized then this would be an improvement. However a closer look at the data reveals another problem: expenses in the numerator include not just depreciation but also rent and many costs associated with maintaining the premises and equipment such as insurance. From the annual report in 2006, the BMO amortizes premises and equipment by $0.36 billion. However the amount reported to OSFI was $1.2 billion because rent and other costs are included. Unfortunately, I can find no way to separate rent expense from amortization for all observations in the sample and not all lease agreements are capitalized. In a preliminary version of this paper, capital expenses divided by total assets was the measure of the average cost of physical capital. This caused a potential problem that is identified by Hughes and Mester (2013): increasing assets causes the price of physical capital to decrease which may confound the parameter estimates.

6.2 Identification and omitted variables

The following subsection details some of the potential identification issues this paper encounters. The cost function I estimate has total expenses (cost) as the dependent variable and the individual expenses divided by input units are how input prices are defined on the right-hand side. Given that the input prices are created from the same components of total cost, an exact solution is possible. This reduces but does not eliminate the possibility for omitted variable bias. There is concern that some outputs have been omitted. For example, if trading income is increasing significantly and labour costs rise in response, this would bias RTS estimates downward. Similarly, not all expenses have been accounted for. Physical capital and labour expense account for 80 percent of all noninterest expenses over most of the sample – Labour accounts for 60 percent. The other costs include advertising, legal fees, theft insurance, and a sundry list.

Section 6 identifies the price of physical capital as problematic – the value of owned and rented physical capital is not available or incorrectly measured. The problem is first

[19] National Bank of Canada is a particularly strong and consistent outlier. Using the mismeasured method of calculating the price of capital, it is significantly higher than the other five banks.

circumvented by omitting physical capital to estimate a short-run cost function. The long-run cost function includes physical capital expense and defines its input price as physical capital expense per full-time equivalent worker. This avoids possible multicollinearity with outputs however it introduces collinearity between the price of physical capital and the price of labour – both have the same denominator. This is of limited concern because the coefficients on these input-prices and squared-prices do not appear in the calculation for RTS. Consequently, the results for both the short-run and long-run cost functions are presented for comparison.

Omitted variables[20] can be a problem as it relates to noninterest income, a measure of service output. There is no general consensus in the literature on how to measure it. I follow the recommendation of Clark and Siems (2002) and use noninterest income as a proxy for service output, but this is not ideal. The amount of services produced might be the same in two periods despite differences in revenue. This is easy to imagine at a trading desk. On one day, effort is rewarded with gains while similar effort on the next goes unrewarded with losses. Another example is investment banking. Banks may produce as many initial public offerings in a downturn as in the previous time period however the total value of the IPO (and revenue) is less. If revenue is high, it appears more services are produced. Similarly in wealth management, fund managers may perform an equal amount of work but if the fund performs poorly then revenues are likely to fall. In our estimation, this appears as a decrease in output while total cost remains unchanged or, if employees receive significant bonuses, decrease slightly. Given our estimates in section 7, an increase in noninterest income leads to a small or modest increase in total costs. This suggests that noninterest income is capturing at least some of the relationship between service output and total costs. If total cost and revenue income are both positively correlated to the omitted variable, services produced, then the coefficients could be biased upwards and RTS could be biased downwards.

[20] See Cameron and Trivedi (2005)

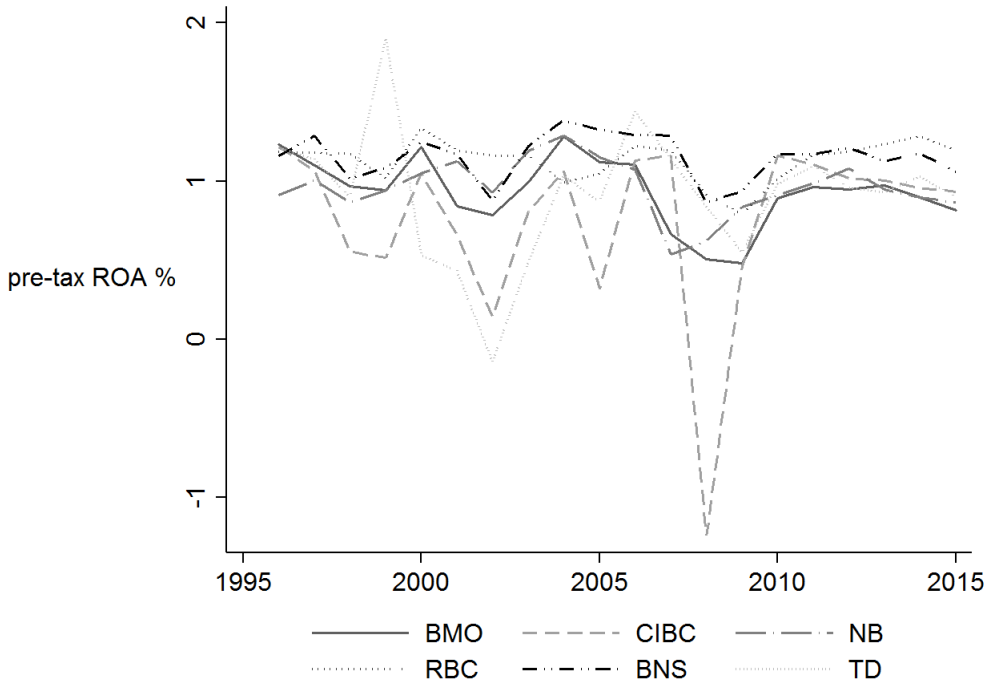


Figure 6: Pre-tax ROA
Source: OSFI.

A second omitted variable has to do with unobservable risk. If WACC is a function of risk and total cost is a function of risk, then there is omitted variable bias of a similar form as that described above: the coefficients are biased upwards. Since the input-price variables only appear in the RTS estimate through cross-product terms with output, any bias is likely to have only a small effect on the RTS estimates. Including the cost of equity from market prices could capture some of this risk. Additionally, all the banks in the sample are large and enjoy an implicit government guarantee that should mitigate the consequences of unobserved risk. If market participants believe this guarantee will be honoured in the event of a default, then it is even possible that there will be no relationship between funding costs, total costs and risk. Figure 6 illustrates pre-tax ROA among the Big Six banks. Section 4.1 explains how the panel dynamic ordinary least squares estimator (PDOLS) can provide more accurate in-sample estimates than a FE OLS estimator when 'T' becomes large. This limits endogeneity by introducing pre-determined variables to the estimation.

7 Results

7.1 Allen and Liu (2007) alternative model

To evaluate RTS and the methodology described, the Allen and Liu (2007) model is estimated using more recent data. They estimate a trans-log cost function with three inputs: physical capital, labour, and deposits. They include five outputs: consumer loans, non-mortgage loans, mortgage loans, other financial assets, and noninterest income (also known as non-traditional activities). Data was from the first fiscal quarter of 1983 to the third fiscal quarter of 2003. Table 4 shows a description of output and prices. Categories reflect bank business activities and the data availability. At this time, banks were entering new fields of financial services including investment banking, securitization, and wealth management. The time period includes as many as six regulatory regime changes. In order to make assets and revenue comparable, the Boyd-Gertler asset (BG asset) transforms noninterest income into a proxy asset using the Boyd et al. (1994) method. It makes a reasonable assumption that the return to on-balance sheet assets is equal to the return to off-balance sheet assets. Using this identity as defined in equation 23 and rearranging generates the theoretical value for off-balance sheet assets stated in equation 24.

$$\frac{Interest\ income}{Total\ Loans} = \frac{Non\ interest\ income}{BG\ asset} \tag{23}$$

$$BG\ asset = \frac{(Total\ Loans)(Non\ interest\ income)}{Interest\ income} \tag{24}$$

A point of difference in the estimation is that Allen and Liu (2007) consider only the price of deposits. They study a sample period from 1983 to 2003, and in the pre-1997 years, the Big Six are either not making use of alternative sources of funds such as repos or there is insufficient data available. However since 1996, Canadian banks rely on funding such as the repo market and subordinated debt to finance operations, so these are included in the average cost of funds.

Allen and Liu (2007) find that the higher order output terms in the trans-log function exhibit multicollinearity thus they constrain these to be zero. They estimate the trans-log

cost function in equation 25 with five outputs and three prices. Lower case letters denote the natural logarithm of the variable in question, Y_q represents the q^{th} output, and w_j represents the j^{th} price. t is a time dummy variable that represents technological change over the sample period. Equation 25 is estimated. In keeping with the previous notation, this equation has the time subscripts suppressed, and lower case variables are in natural logarithms. To test whether the series is stationary, the MADF test determines whether the series is cointegrated and uses residuals from an OLS fixed effects model. These are summarized in table 6.

$$log\left(\frac{C}{W_1}\right) = \alpha_0 + \sum_{q=1}^{5} \alpha_q \left(y_p\right) + \sum_{j=2}^{3} \beta_j \left(w_j - w_1\right) + \sum_{q=1}^{5} \sum_{j=1}^{3} \sum_{j} \gamma_{pj} \left(y_q w_j - y_1 w_1\right)$$

$$+ \left(\frac{1}{2}\right) \sum_{p=1}^{3} \sum_{j=2}^{3} \delta_{pj} \left(w_k w_j - w_1 w_1\right) + t + \epsilon_{it} \tag{25}$$

Equation 26 measures RTS at the average price $\bar{w}_j$. RTS by bank and across time are presented in figure 16.

$$\bar{RTS} = \left(\sum_{p=1}^{5} \frac{\partial\ log(C/W)}{\partial\ log(Y_p)}\right)^{-1} = \sum_{q=1}^{5} \alpha_q + \sum_{q=1}^{5} \sum_{j=1}^{3} \gamma_{qj} \bar{w}_j \tag{26}$$

The coefficients $\gamma_{1,1}$ and β_1 are not estimated directly, but the value is implied using the constraints in equation 4.

Table 5 summarizes the results. The time trend squared coefficient is no longer statistically significant in either the FE or PDOLS estimation which differs from the original study of Allen and Liu (2007). A LR-test using the FE model and the TFE model with a half-normal error term fails to reject the null of no expected inefficiency. The PDOLS estimator generates RTS of 3% however these are not significant at the 5% level. This can be interpreted as a 1% increase in all outputs generates additional total costs of approximately 0.971%, and the average cost curve is above the marginal cost curve. The fixed effect model has higher estimates than the PDOLS that is to be expected given the simulations of Kao and Chiang (2001), and the MADF test for cointegrated. These RTS are smaller than those

found by Allen and Liu (2007) in the 1983-2003 time period[21]. The RTS from the PDOLS estimator are not significantly different from constant RTS.

7.2 Short-run and long-run cost function estimation

A short-coming of the Allen and Liu (2007) specification of inputs and outputs is that, by excluding the higher order output terms, the model does not capture the curvature, or u-shape, of an average cost function. The following short-run cost model, described in section 6 and shown in equation 27, is able to estimate these higher order terms without introducing multicollinearity or otherwise confusing the results. The Boyd-Gertler asset has been replaced with the level of noninterest income less retail and commercial bank account fees following the recommendations of Clark and Siems (2002) and Davies and Tracey (2014). Trading gains and losses, and gains and losses from securities held for non-trading purposes have also been removed from noninterest income, a departure from Allen and Liu (2007). The cost of physical capital is difficult to measure, so it is omitted in the short-run cost function, and the cost of deposits is replaced with WACC. Outputs have been grouped together in three categories: loans to households, securities and loans to government and business, and noninterest income. Alternative specifications are presented in section 7.6. A summary of these variables is available in table 3. The trans-log short-run cost function estimating equation becomes:

$$log\left(\frac{C_{SR}}{W_1}\right) = \alpha_0 + \sum_{p=1}^{3} \alpha_p \left(y_p\right) + \beta_2 \left(w_2 - w_1\right) + \sum_{p=1}^{3} \sum_{j=1}^{2} \sum_{j \neq l} \gamma_{pj} \left(y_p w_j - y_p w_1\right)$$

$$+ \left(\frac{1}{2}\right) \sum_{j=1}^{2} \delta_{j2} \left(w_j w_2 - w_1 w_1\right) + \left(\frac{1}{2}\right) \sum_{p=1}^{3} \sum_{g=1}^{3} \sigma_{pg} \left(y_p y_g\right) + t + \epsilon_{it} \qquad (27)$$

where w represents the price of inputs, y represents outputs, C represents total costs and t is a time trend. Lower case variables are in logarithms. The long-run cost function, despite the difficulties identified in section 6.1, includes the cost of physical capital in the dependent variable, total cost. The price of physical capital, defined in table 3, is an

[21] Allen and Liu (2007) find economies of scale of 12.6% with the fixed effect model and 6.1% with PDOLS

additional independent variable along with its cross-product and square terms:

$$log\left(\frac{C_{LR}}{W_1}\right) = \alpha_0 + \sum_{p=1}^{3}\alpha_p\left(y_p\right) + \sum_{j=2}^{3}\beta_j\left(w_j - w_1\right) + \sum_{p=1}^{3}\sum_{j=1}^{2}\sum_{j\neq l}\gamma_{pj}\left(y_p w_j - y_p w_1\right)$$

$$+ \left(\frac{1}{2}\right)\sum_{j=1}^{3}\delta_{jq}\sum_{q=2}^{3}\left(w_j w_q - w_1 w_1\right) + \left(\frac{1}{2}\right)\sum_{p=1}^{3}\sum_{g=1}^{3}\sigma_{pg}\left(y_p y_g\right) + t + \epsilon_{it} \tag{28}$$

Table 6 shows the result from the unit root tests. The null hypothesis of non-stationary residuals in the fixed effect model is rejected while table 7 summarizes the MADF test for cointegration. There are 64 time observations for 6 cross-sectional units. The null hypothesis that none of the series is cointegrated is rejected at the 5% percent level of significance.

7.3 Returns to scale

Table 8 displays coefficients and standard errors for the short-run and long-run cost functions. The time trend is often interpreted as technological change in the stochastic frontier literature[22]. If the coefficient is negative, technological efficiency improved over the sample period – the cost per asset and services declined. In both the short and long-term cost function, the time trend is of modest size, negative, and statistically significant for both the fixed effect (FE) and PDOLS estimator. However it should be noted that its magnitude is small and smaller than that found in the original Allen and Liu (2007) results. In earlier estimations, a squared time trend was included and found to be statistically insignificant, so it has been excluded here.

Short-run RTS is calculated as:

$$R\bar{T}S = \left(\sum_{q=1}^{3}\frac{\partial\ log(C/W)}{\partial\ log(Y_q)}\right)^{-1} = \left(\sum\alpha_q + \sum_{q=1}^{3}\sum_{j=1}^{2}\gamma_{qj}\bar{w}_j + \sum_{q=1}^{3}\sum_{q=1}^{3}\sigma_{qg}\bar{y}_g\right)^{-1} \tag{29}$$

[22] see Kumbhakar et al. (2015)

and the equation has one additional cross-product term when physical capital is included as an input in the long-run cost function. The RTS calculation is the inverse of the marginal cost of producing an additional 1% of each output. For example, a RTS of 1.1 implies a 10% increase in each output will add 9.091% to total cost, so there are increasing RTS. The results for the short-run cost function are summarized in table 9. RTS is evaluated at the mean value of input prices and output quantities. This value is consistently close to unity, the definition of constant RTS. The economies of scope in the short and long-run are small and statistically insignificant.

If WACC is driving the RTS estimates, RTS would be low when interest expense per asset is high, such as the years 2006 to 2008 and it would be high in years when rates were low such as the years 2002 to 2004. Looking at table 10, average RTS seems to be slightly higher in years with lower WACC. Figures 17 and 19 are scatterplots of RTS against bank size. Over the entire sample, the scatterplots show RTS as a constant 'cloud' without a determinable trend. Banks adjusted total financial assets on the balance in response to changes in the cost of funding, and other costs and business conditions. This is supported by figures 8 and 18. An increase in the WACC shifts the RTS curve down and to the left while RTS is decreasing or constant as bank size increases. If interest rates are low, then banks benefit from a lower cost of funding. If rates are high, then banks are more likely to face constant or decreasing RTS. Regardless, if the return on assets and noninterest revenue were high enough, banks could have been more profitable. Table 11 illustrates that in some periods, some banks operated with decreasing RTS and others operated with increasing RTS. While the price of physical capital may have been imperfectly defined, including the physical capital in the analysis increased the statistically significant increasing RTS from 11 observations to 99.

7.4 Technical efficiency

Practitioners often study the residual from a cost function to determine technical efficiency. This is explained in section 4.2 and defined in equation 16. The results from the short-run and long-run cost function are reported in table 12 and figure 20. Looking at the short-run cost function, the best practices observation is the National Bank in the third fiscal quarter

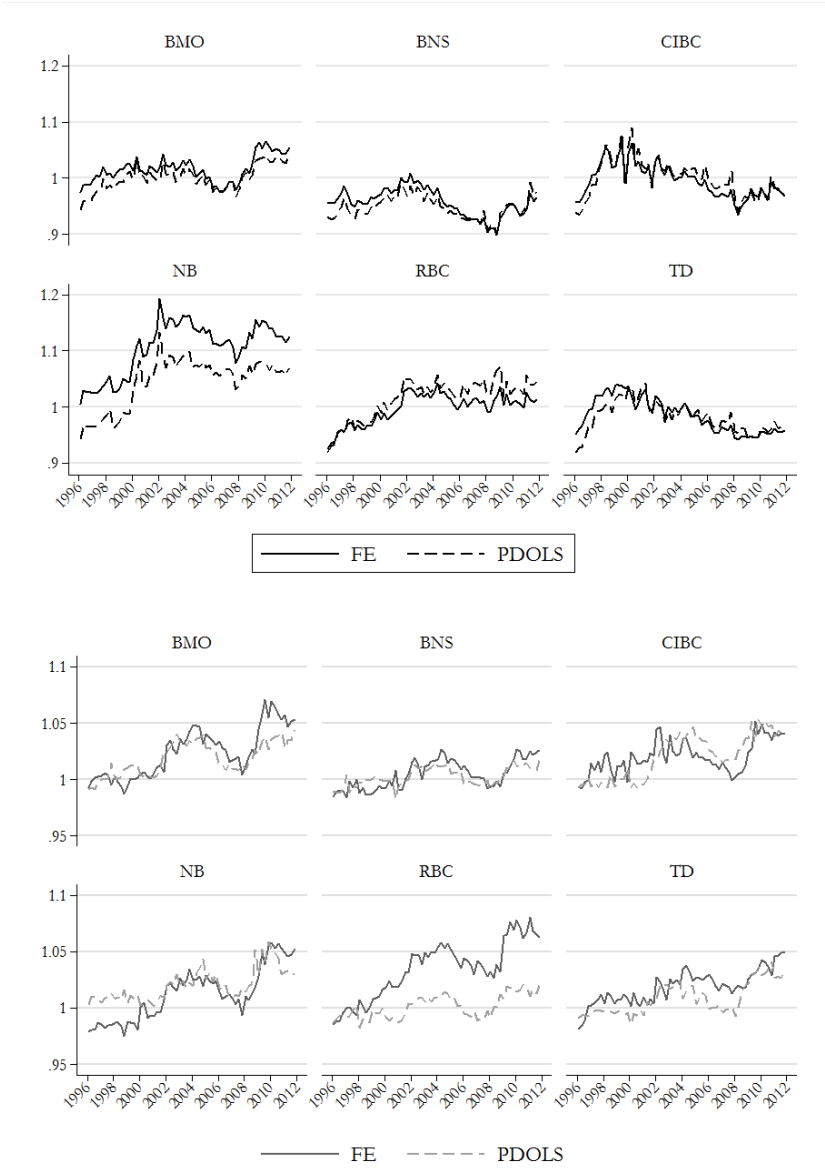


Figure 7: Returns to scale

Note: RTS for the short-run cost function (top) and the long-run cost function with physical capital (bottom).

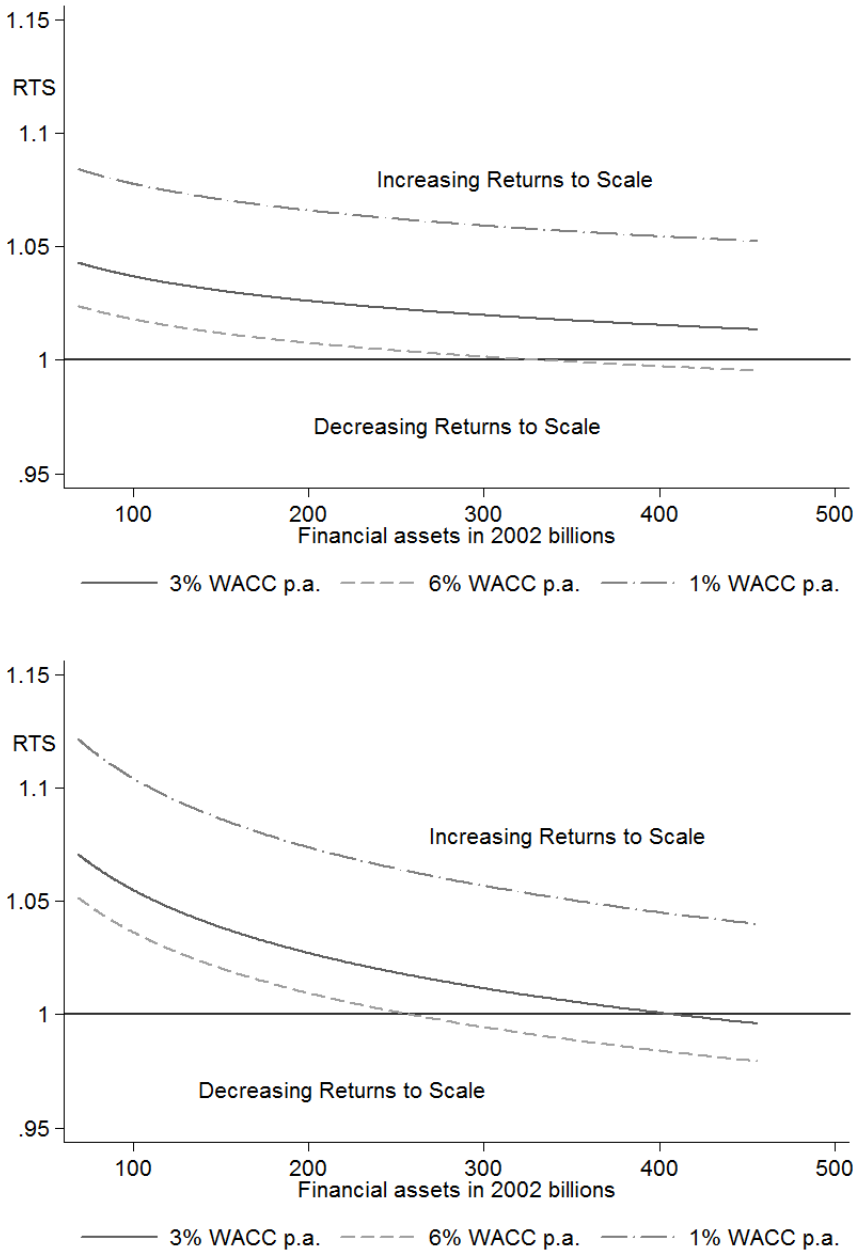


Figure 8: Short-run RTS & financial assets

Note: Ray-scale RTS estimates with the fixed effect estimator (top) and ray-scale RTS estimates with the panel dynamic estimator (bottom). Using the coefficient estimates from table 8, RTS is evaluated at the median cost of labour and by proportionally increasing output for three different weighted average costs of capital (WACC). See section 3.3 for further details on the calculations.

of 2000. Table 12 shows that the Canadian banks are 91.69 percent efficient, but this is an unsatisfactory conclusion. Figure 20 suggests the best practices observation is something of an outlier with just one residual of comparable size. Table 12 shows that over the sample, there is little separating the efficiency of each bank. When physical capital is included, the results are similar. Average bank efficiency is 93.85 percent, and this differs little from bank-to-bank. In order to test for technical inefficiency, this paper estimates the Greene (2005) True Fixed Effect model and uses a LR-ratio test to test for inefficiency as described in section 4.2.

Table 13 shows all the information regarding the test for technical efficiency. Jarque–Bera test for normality on the residuals fails to reject the null hypothesis that these are normally distributed. This is confirmed by the loglikelihood ratio test estimated by ML. Recall that the TFE model assumes that the error term can be divided into a normally distributed term and a half normal or exponential term that captures inefficiency. The results from an exponential error term and half-normal are the same for the short-run cost function. The expected efficiency conditional on the error is 99.9 percent which reflects the failure to reject the null in the LR test.

Consider the long-run cost function, the skewness and kurtosis are both higher when physical capital is included. The null hypothesis of a normal distribution is rejected at the 5 percent level of confidence by the Jarque–Bera test. The TFE model with an exponentially distributed[23] inefficiency term and estimated by ML is tested against the fixed effect OLS estimator. The null hypothesis that these two models are the same is rejected at the 5 percent confidence level. The unconditional expected efficiency of any bank in any time period is 98.9 percent. Alternatively, there is an average expected inefficiency of just 1.1 percent. The result is surprising in two ways: *first*, the total inefficiency differs little if physical capital is included or excluded. *Second*, there is little difference among banks. Given this sample period includes large investments in online banking and ATM machines, and a significant increase in bank size, more variation among banks may have been expected. However this also leads to the conclusion that the Big Six banks operate with similar cost

[23]Both the half-normal and exponential inefficiency term were estimated. The exponential error term estimates produced a higher LR statistic and a lower inefficiency score. Erring on the side of caution, these results are reported and the half-normal error term was omitted. In practice, both distributions give similar results. See Kumbhakar et al. (2015).

structures.

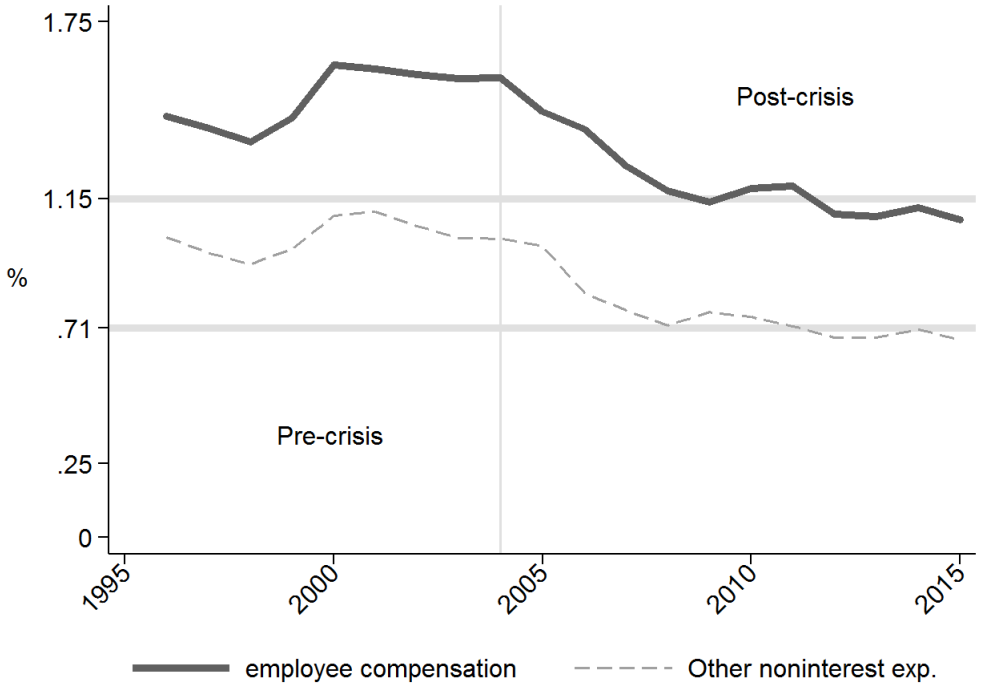


Figure 9: Noninterest expenses per asset

Note: Employee compensation includes salary, benefits, and bonuses. Other noninterest expense includes physical capital expense. The $2.4 billion dollar legal penalty against CIBC has been removed to better illustrate the trend. Source: OSFI.

7.5 Noninterest expense and technological improvement

Where are the returns to scale in Canadian banking? The RTS analysis suggests that increasing bank size will not lead to greatly increased cost efficiency. Furthermore, there is little difference in technical efficiency among the Big Six banks: each manages costs similarly with little or no statistically significant difference. In each estimation, a time trend coefficient was significant and negative. This is usually interpreted as technological improvement in the stochastic frontier analysis literature. Figure 9 illustrates noninterest expenses divided into two categories: (i) employee compensation, and (ii) everything else which includes items such as physical capital expense and legal fees. From 2004 to 2009, labour and other noninterest expenses per asset declined significantly - the average labour expense per asset fell from 1.5 to 1.15 percent and other noninterest expenses fell from 1.1

to 0.71 percent. After 2009, noninterest expense per asset changed little despite a large increase in bank balance sheet assets following 2012. Importantly, this is irrespective of the size of any bank in the sample. Figure 10 illustrates noninterest expense per asset for each of the Big Six banks. Banks are not ordered by size – largest banks were neither the most nor the least efficient. BNS is the third-largest bank in the sample yet maintains the lowest ratio. The smallest and the largest, NB and RBC, have similar noninterest expenses per asset.

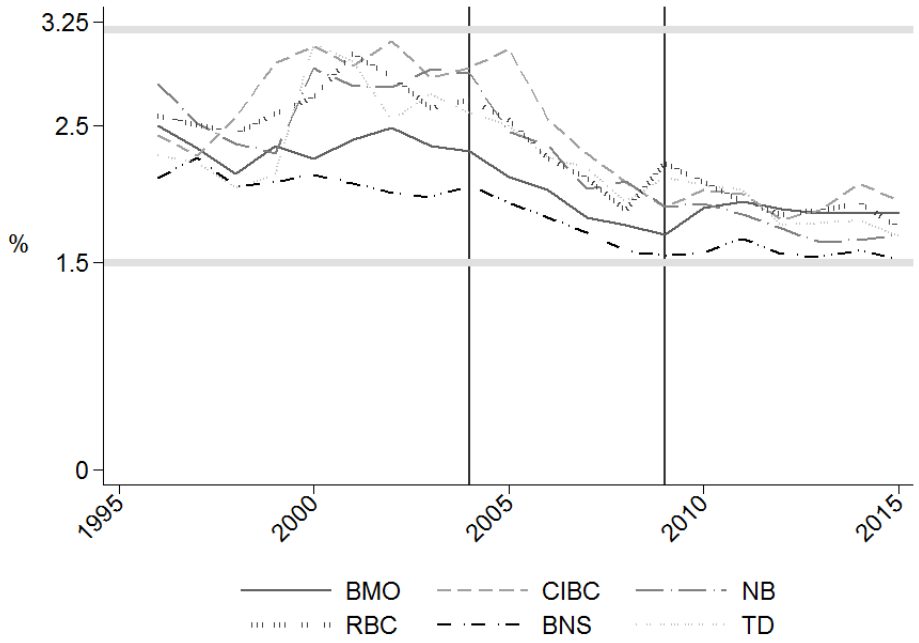


Figure 10: Noninterest expenses per asset

Note: Employee compensation includes salary, benefits, and bonuses. Other noninterest expense includes physical capital expense. The $2.4 billion dollar legal penalty against CIBC has been removed to better illustrate the trend. Source: OSFI.

Calmès et al. (2013) state that the Canadian banks decreased noninterest expenses in response to the subprime crisis, but given subsequent history this claim seems uncertain. The decline in noninterest expense per asset began prior in 2004/2005, continued through the financial crisis, and was maintained afterward. For a comparison on Canadian and US banks, see McKeown (2017a). Calmès et al. (2013) surmise that the product-mix of Canadian banks might be fundamentally different from those of Canadian banks. It is possible that a change in the product-mix of Canadian banks generated the cost-savings

however this would have required the banks to mimic each other's strategy. In RBC's 2007 annual report, they state that in 2006, an appreciating Canadian dollar decreased noninterest expenses denominated in US dollars. However this does not explain how the ratio was maintained in 2015 when the Canadian dollar had decreased in value.

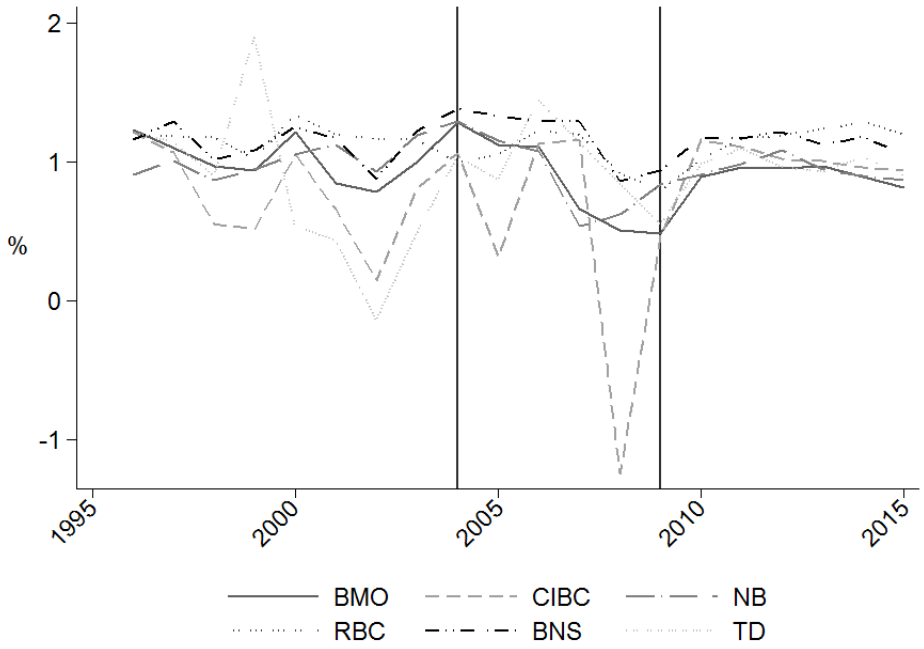


Figure 11: Noninterest expenses per asset

Note: The $2.4 billion dollar legal penalty against CIBC has been removed to better illustrate the trend.
Source: OSFI.

From interviews with bank managers, Allen et al. (2006) offer anecdotal evidence that investment in information and communications technology (ICT) were made largely to improve cost efficiency. Although difficult to quantify, these investments could explain the considerably improvement in cost management. If bank size was not responsible, what caused the improvement noninterest costs per asset? Studying U.S. banks, Berger and Mester (2003) conclude that banks profited considerably from improvements in information and communications technology (ICT – it allowed them to offer more services at higher quality and improve back-office activities such as data collection. Somewhat counter-intuitively, they find that during the 1990's U.S. bank profits and costs increased concurrently. Although the technology was expensive, it allowed greater profits. Did the Canadian banks

experience profits because costs decreased? Figure 11 suggests not. From 1996 to 2015, the trend pre-tax return on assets is around 1 percent. The decrease in noninterest expenses per asset allowed the banks to maintain their ROA rather than increase it.

7.6 RTS robustness tests

Table 14 shows the results from six separate robustness tests. Excluding the implied cost of equity does not materially change the result. This is likely because the implied cost of equity makes up only a small proportion of total costs through most of the 1996-2011 time period. While equity is the most expensive source of funding, it is also quite small relative to deposits and labour. Similarly, estimating the cost of equity as a price separate from other sources of funding produces similar results in the short-run cost function. The transcendental log cost function is suitable when each firm is of similar size. Omitting the National Bank of Canada, a bank that is one-quarter the size of the largest Canadian bank, fails to alter the result of constant RTS. This is true with both the short-run and long-run cost functions. Estimating the long-run cost function and excluding the implied cost of equity similarly fails to alter the main result.

The null hypothesis of constant returns to scale cannot be rejected in most alternative specifications. The exception is the long-run cost function with four output variables – alternative model 6. In this case, there are significant diseconomies of scale at the 10 percent level of significance.

7.7 Microeconomic properties

A well-defined cost function should exhibit good microeconomic properties such as price concavity and monotonicity. It is possible to use the estimated coefficients to test whether these conditions are satisfied. This is explained in section 3. My purpose now is two-fold: (i), to observe whether these conditions are satisfied, and (ii), to determine if the Ryan and Wales (2000) method can improve performance. Using equation 7, Table 15 shows the number of observations that satisfy monotonicity, those that satisfy price concavity in equation 8, and the implied constraints from equation 4 that a doubling of prices, all else being equal, doubles total cost. Selected results are presented in Table 16. Short-run

and long-run model estimates satisfy monotonicity for more than 90 percent of observations. However for any given observation, neither model satisfies price concavity. Applying the first step of the Ryan and Wales (2000) method improves the short-run model results considerably. As many as 89 percent of observations now satisfy price concavity without significantly altering the RTS estimates. The performance of the pricing constraint, that a doubling of input prices also doubles costs, is also greatly improved. The short-run cost function performs extremely well in conjunction with the Ryan and Wales (2000) method; it is capable of satisfying the desired microeconomic conditions for nearly all observations and the estimate of returns to scale is not significantly altered. The long-run cost function performs modestly well. After the normalization, it satisfies monotonicity while the price constraint is close to 1. However it could not satisfy price concavity even after applying the method developed by Ryan and Wales (2000). This is likely because the physical capital price (expenses divided by total assets) is mis-specified or correlates with output.

8 Conclusion

Looking at the Big Six Canadian banks, I find that there is little connection between bank size and cost. On average from 1996 to 2011, the banks exhibited constant returns to scale. Interestingly, average noninterest expenses per asset, a measure of efficiency, improved considerably for all banks from 2004 to 2009. However this technological change was unrelated to the size of a Big Six bank – in fact, each bank enjoyed it regardless of its size relative to another member of the Big Six. In the U.S., there has been similar research questioning the advantages of increasing bank and industry consolidation. Minton et al. (2017) find that smaller banks enjoyed a higher Tobin-q than their larger rivals while Gandhi and Lustig (2015) find that smaller banks enjoyed a higher risk-adjusted return. Trujillo-Ponce (2013) observes that there is little correlation between size and profitability. Haldane (2010) raises concerns about the level at which RTS might be exhausted. He suggests that it could be less than $100 million and that at some point diseconomies of scale will occur. Furthermore, he questions just how much of an efficiency gain occurs from increasing bank size. Theoretically, he notes that if all large banks are fully diversified,

then they all have equal risk-return profiles and their expected return should be equal to the market portfolio – this questions a common claim that large banks are better diversified than smaller banks. Berger et al. (1993) estimates a profit function and finds RTS might be exhausted as low as $1 billion worth of assets. Admati and Hellwig (2014) remain sceptical that increased size leads to increased efficiency through better risk choices, diversification and information processing. Stiroh and Rumble (2006) find that revenue diversification leads to lower risk-adjusted performance and profitability. Amel et al. (2004) study financial institutions following a merger or acquisition. They discover little evidence for economies of scope and none for increased cost efficiency. On average, stock market prices of a combined entity vary little from the pre-existing firm. This suggests few expected synergies from most mergers. Amel et al. (2004) hypothesize that perhaps the benefits from acquisition are not yet revealed and that improvements in risk management are temporarily masked. However the subsequent U.S. banking crisis of 2007-09' casts doubt on this hypothesis.

References

Admati, Anat and Martin Hellwig (2014), *The bankers' new clothes: What's wrong with banking and what to do about it*. Princeton University Press.

Allen, Jason (2011), *Competition in the Canadian mortgage market*. Bank of Canada Review Winter 2010-11.

Allen, Jason, Robert Clark, and Jean-François Houde (2013), "The effect of mergers in search markets: evidence from the Canadian mortgage industry." *The American Economic Review*, 104, 3365–3396.

Allen, Jason, Robert Clark, and Jean-François Houde (2014a), "Price dispersion in mortgage markets." *The Journal of Industrial Economics*, 62, 377–416.

Allen, Jason, Robert Clark, and Jean-François Houde (2014b), *Search frictions and market power in negotiated price markets*. NBER Working Paper No. 19883.

Allen, Jason, Walter Engert, Ying Liu, et al. (2006), *Are Canadian Banks Efficient?: a Canada-US Comparison*. Bank of Canada Staff Working Paper 2006–33.

Allen, Jason and Ying Liu (2007), "Efficiency and economies of scale of large Canadian banks." *Canadian Journal of Economics/Revue canadienne d'économique*, 40, 225–244.

Almanidis, Pavlos, Giannis Karagiannis, and Robin C Sickles (2015), "Semi-nonparametric spline modifications to the Cornwell–Schmidt–Sickles estimator: an analysis of US banking productivity." *Empirical Economics*, 48, 169–191.

Amel, Dean, Colleen Barnes, Fabio Panetta, and Carmelo Salleo (2004), "Consolidation and efficiency in the financial sector: a review of the international evidence." *Journal of Banking & Finance*, 28, 2493–2519.

Anderson, Ronald W and Karin Joeveer (2012), *Bankers and bank investors: reconsidering the economies of scale in banking*. CEPR Discussion Paper No. DP9146.

Arellano, Manuel and Stephen Bond (1991), "Some tests of specification for panel data: Monte carlo evidence and an application to employment equations." *The Review of Economic Studies*, 58, 277–297.

Berger, Allen N, Robert DeYoung, Mark J Flannery, David Lee, and Özde Öztekin (2008), "How do large banking organizations manage their capital ratios?" *Journal of Financial Services Research*, 34, 123–149.

Berger, Allen N, Diana Hancock, and David B Humphrey (1993), "Bank efficiency derived from the profit function." *Journal of Banking & Finance*, 17, 317–347.

Berger, Allen N and David B Humphrey (1992), "Measurement and efficiency issues in commercial banking." In *Output measurement in the service sectors*, 245–300, University of Chicago Press.

Berger, Allen N and Loretta J Mester (2003), "Explaining the dramatic changes in performance of US banks: technological change, deregulation, and dynamic changes in competition." *Journal of Financial Intermediation*, 12, 57–95.

Berger, Allen N, Philip Molyneux, and John OS Wilson (2014), *The Oxford handbook of banking*. OUP Oxford.

Beyhaghi, Mehdi, Chris D'Souza, and Gordon S Roberts (2014), "Funding advantage and market discipline in the Canadian banking sector." *Journal of Banking & Finance*, 48, 396–410.

Bhattacharyya, N (2007), "Dividend policy: a review." *Managerial Finance*, 33, 4–13.

Bikker, Jacob A, Sherrill Shaffer, and Laura Spierdijk (2012), "Assessing competition with the panzar-rosse model: the role of scale, costs, and equilibrium." *Review of Economics and Statistics*, 94, 1025–1044.

Bordo, Michael D, Angela Redish, and Hugh Rockoff (2015), "Why didn't Canada have a banking crisis in 2008 (or in 1930, or 1907, or...)?" *The Economic History Review*, 68, 218–243.

Boyd, John H, Mark Gertler, et al. (1994), "Are banks dead? or are the reports greatly exaggerated?" *Federal Reserve Bank of Minneapolis Quarterly Review*, 18, 2–23.

Bruner, Robert F, Kenneth M Eades, Robert S Harris, and Robert C Higgins (1998), "Best practices in estimating the cost of capital: survey and synthesis." *Financial Practice and Education*, 8, 13–28.

Calmès, Christian, Raymond Théoret, et al. (2013), "Is the Canadian banking system really "stronger" than the US one?" *Review of Economics and Finance*, 4, 1–18.

Cameron, A Colin and Pravin K Trivedi (2005), *Microeconometrics: methods and applications*. Cambridge University Press.

Christensen, Laurits R, Dale W Jorgenson, and Lawrence J Lau (1973), "Transcendental logarithmic production frontiers." *The Review of Economics and Statistics*, 28–45.

Chua, Chew Lian, Hsein Kew, and Jongsay Yong (2005), "Airline code-share alliances and costs: imposing concavity on translog cost function estimation." *Review of Industrial Organization*, 26, 461–487.

Claessens, Stijn and Luc Laeven (2004), "What drives bank competition? some international evidence." *Journal of Money, Credit, and Banking*, 36, 563–583.

Clark, Jeffrey Arthur and Tom Siems (2002), "X-efficiency in banking: looking beyond the balance sheet." *Journal of Money, Credit, and Banking*, 34, 987–1013.

Da, Zhi, Re-Jin Guo, and Ravi Jagannathan (2012), "CAPM for estimating the cost of equity capital: interpreting the empirical evidence." *Journal of Financial Economics*, 103, 204–220.

Davies, Richard and Belinda Tracey (2014), "Too big to be efficient? The impact of implicit subsidies on estimates of scale economies for banks." *Journal of Money, Credit and Banking*, 46, 219–253.

Dietrich, Andreas and Gabrielle Wanzenried (2011), "Determinants of bank profitability before and during the crisis: evidence from Switzerland." *Journal of International Financial Markets, Institutions and Money*, 21, 307–327.

Diewert, W Erwin and Terence J Wales (1989), *Flexible functional forms and global curvature conditions*. NBER Technical Working Paper No. 40.

Fama, Eugene F and Kenneth R French (1993), "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics*, 33, 3–56.

Feng, Guohua and Apostolos Serletis (2010), "Efficiency, technical change, and returns to scale in large US banks: panel data evidence from an output distance function satisfying theoretical regularity." *Journal of Banking & Finance*, 34, 127–138.

Feng, Guohua and Xiaohui Zhang (2012), "Productivity and efficiency at large and community banks in the US: a bayesian true random effects stochastic distance frontier analysis." *Journal of Banking & Finance*, 36, 1883–1895.

Feng, Guohua and Xiaohui Zhang (2014), "Returns to scale at large banks in the us: a random coefficient stochastic frontier approach." *Journal of Banking & Finance*, 39, 135–145.

Gandhi, Priyank and Hanno Lustig (2015), "Size anomalies in US bank stock returns." *The Journal of Finance*, 70, 733–768.

Goddard, John and John OS Wilson (2009), "Competition in banking: a disequilibrium approach." *Journal of Banking & Finance*, 33, 2282–2292.

Greenbaum, Stuart I., Anjan V Thakor, and Arnoud Boot (2015), *Contemporary financial intermediation*. Academic Press.

Greene, Willam (2005), "Fixed and random effects in stochastic frontier models." *Journal of Productivity Analysis*, 23, 7–32.

Gropp, Reint and Florian Heider (2010), "The determinants of bank capital structure." *Review of Finance*, 14, 587–622.

Guidara, Alaa, Issouf Soumaré, Fulbert Tchana Tchana, et al. (2013), "Banks' capital buffer, risk and performance in the Canadian banking system: impact of business cycles and regulatory changes." *Journal of Banking & Finance*, 37, 3373–3387.

Haldane, Andrew G (2010), "The $100 billion question." *Revista de Economía Institucional*, 12, 83–110.

Hughes, Joseph P and Loretta J Mester (2008), *Efficiency in banking: theory, practice, and evidence*. FRB of Philadelphia Working Paper No. 08-1.

Hughes, Joseph P and Loretta J Mester (2013), "Who said large banks don't experience scale economies? Evidence from a risk-return-driven cost function." *Journal of Financial Intermediation*, 22, 559–585.

Illes, Annamaria, Marco J Lombardi, and Paul Mizen (2015), *Why did bank lending rates diverge from policy rates after the financial crisis?* BIS working paper 486.

Kao, Chihwa and Min-Hsien Chiang (2001), "On the estimation and inference of a cointegrated regression in panel data." In *Nonstationary panels, panel cointegration, and dynamic panels*, 179–222, Emerald Group Publishing Limited.

Kelly, Gale M. and Francis Janssens (2012), "Transition to IFRS – the impact on the 'Big 6' Canadian banks." *IFRS Technical Report*, 6199.

Kopp, Raymond J and W Erwin Diewert (1982), "The decomposition of frontier cost function deviations into measures of technical and allocative efficiency." *Journal of Econometrics*, 19, 319–331.

Kumar, Pradeep (2013), *Market power and cost efficiencies in banking*. CAPCP Working Paper 2013.

Kumbhakar, Subal C, Hung-Jen Wang, and Alan Horncastle (2015), *A practitioner's guide to stochastic frontier analysis using Stata*. Cambridge University Press.

Levin, Andrew, Chien-Fu Lin, and Chia-Shang James Chu (2002), "Unit root tests in panel data: asymptotic and finite-sample properties." *Journal of Econometrics*, 108, 1–24.

Mark, Nelson C and Donggyu Sul (2003), "Cointegration vector estimation by panel DOLS and long-run money demand." *Oxford Bulletin of Economics and Statistics*, 65, 655–680.

McIntosh, James (2002), "A welfare analysis of Canadian chartered bank mergers." *Canadian Journal of Economics/Revue canadienne d'économique*, 35, 457–475.

McKeown, Robert (2017a), *Costs, size and returns to scale among Canadian and U.S. commercial banks*. Queen's University Working Paper.

McKeown, Robert (2017b), *An overview of the Canadian banking system: 1996 to 2015*. Queen's University Working Paper.

Minton, Bernadette A, René M Stulz, and Alvaro G Taboada (2017), *Are larger banks valued more highly?* NBER Working Paper 23212.

Panzar, John C and James N Rosse (1987), "Testing for 'monopoly' equilibrium." *The Journal of Industrial Economics*, 443–456.

Perez-Saiz, Hector and Hongyu Xiao (2014), *Being Local or going global? Competition and entry barriers in the Canadian banking industry*. http://webmeets.com/files/papers/EARIE/2014/390/PerezSaiz_Xiao_entry_paper_v11.pdf.

Restrepo, Diego A, Subal C Kumbhakar, and Kai Sun (2013), "Are US commercial banks too big." *Universidad EAFIT*, 1.

Restrepo-Tobón, Diego and Subal C Kumbhakar (2015), "Nonparametric estimation of returns to scale using input distance functions: an application to large US banks." *Empirical Economics*, 48, 143–168.

Ryan, David L and Terence J Wales (2000), "Imposing local concavity in the translog and generalized Leontief cost functions." *Economics Letters*, 67, 253–260.

Sarno, Lucio and Mark P Taylor (1998), "Real exchange rates under the recent float: unequivocal evidence of mean reversion." *Economics Letters*, 60, 131–137.

Shaffer, Sherrill (1993), "A test of competition in Canadian banking." *Journal of Money, Credit and Banking*, 49–61.

Shaffer, Sherrill and Laura Spierdijk (2015), "The Panzar–Rosse revenue test and market power in banking." *Journal of Banking & Finance*, 61, 340–347.

Stiroh, Kevin J and Adrienne Rumble (2006), "The dark side of diversification: the case of U.S. financial holding companies." *Journal of Banking & Finance*, 30, 2131–2161.

Trujillo-Ponce, Antonio (2013), "What determines the profitability of banks? Evidence from Spain." *Accounting & Finance*, 53, 561–586.

Ueda, Kenichi and B Weder Di Mauro (2013), "Quantifying structural subsidy values for systemically important financial institutions." *Journal of Banking & Finance*, 37, 3830–3842.

Wheelock, David C and Paul W Wilson (2012), "Do large banks have lower costs? New estimates of returns to scale for US banks." *Journal of Money, Credit and Banking*, 44, 171–199.

Xiang, Dong, Abul Shamsuddin, and Andrew C Worthington (2015), "The differing efficiency experiences of banks leading up to the global financial crisis: a comparative empirical analysis from Australia, Canada and the UK." *Journal of Economics and Finance*, 39, 327–346.

9 Appendix

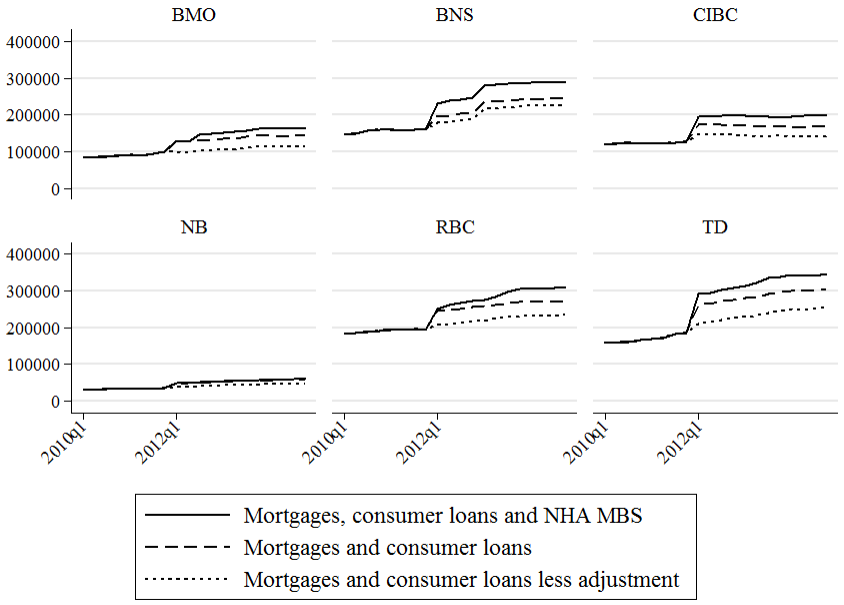


Figure 12: Adjusting for regime change

The solid line is unadjusted assets including NHA MBS (National Housing Act Mortgage Backed Securities), the dark dashed line removes NHA MBS and the dotted line removes NHA MBS and applies the Kelly and Janssens (2012) adjustment.

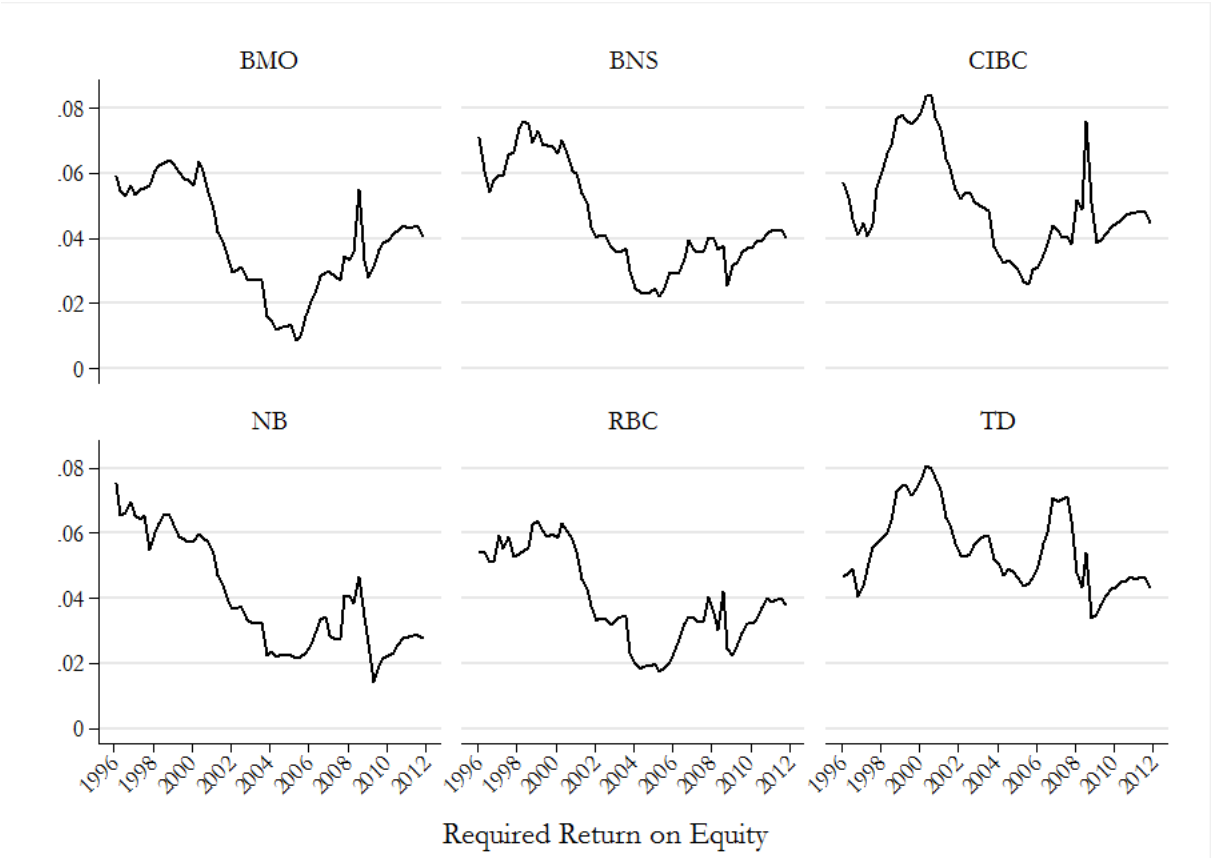
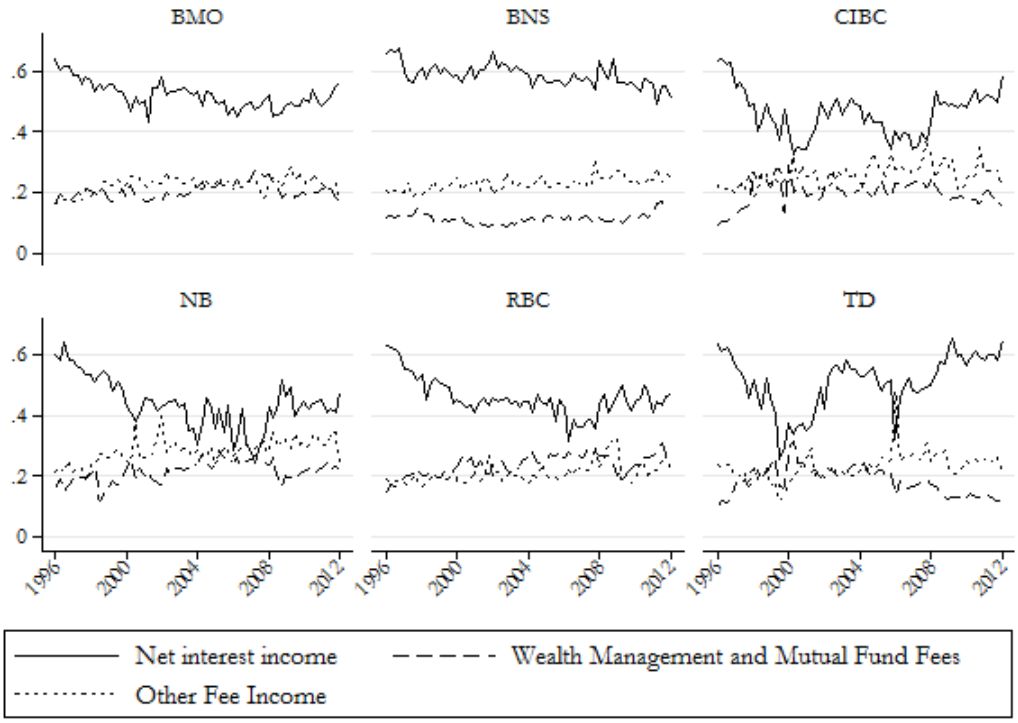


Figure 13: Cost of equity estimated using CAPM

Note: The 3-month Canadian treasury bill market rate is used as the risk-free rate. The expected market premium ($E(r_m - r_f)$) is assumed to be 4% per annum (1% per quarter). Ordinary least squares estimates β_i for each time period and for each bank using data from the previous 60-months. These required costs are then weighted into quarterly averages. Notice that there is only a modest amount of variation among banks because CAPM is predominantly driven by the risk-free rate and expected market premium.



Source: OSFI

Figure 14: Share of net revenue

Other noninterest income includes securitization, investment banking, foreign exchange, retail and card fees. Net interest income holds the highest share of revenue when wealth management is separated from other noninterest income. There appears to be a fair amount of heterogeneity.

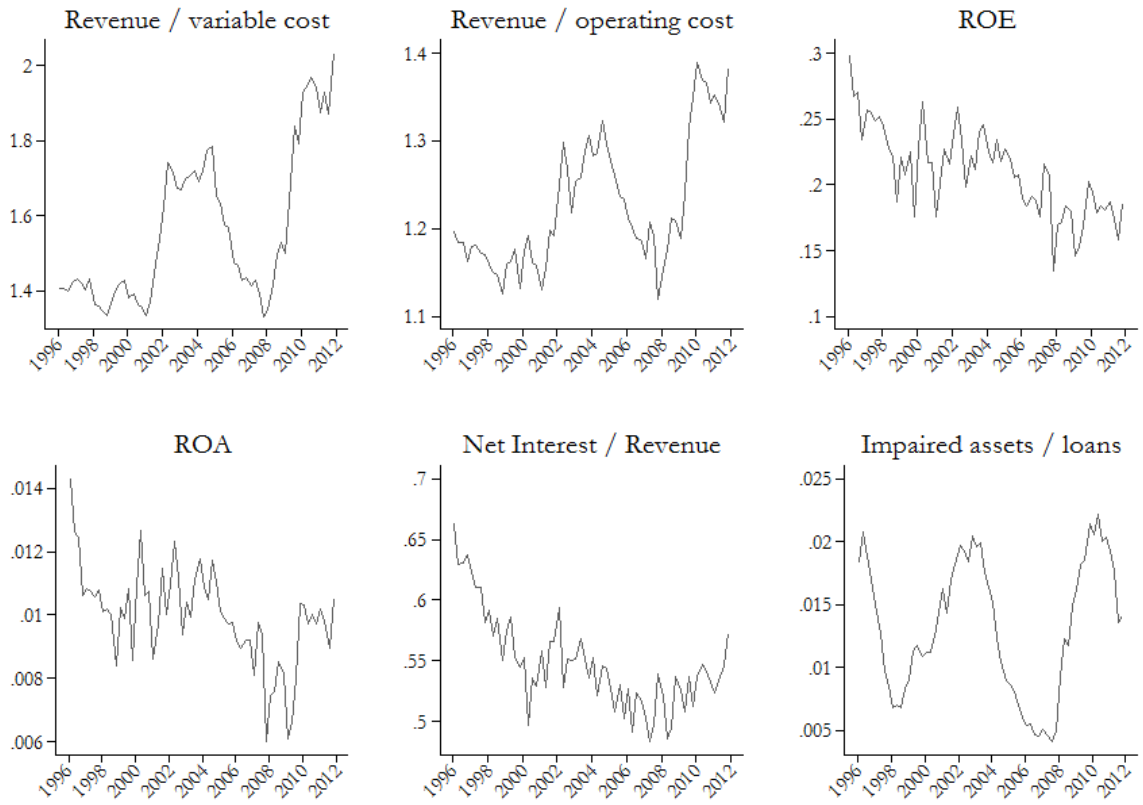


Figure 15: Average characteristics among the Big Six Canadian banks.

Trading and non-trading gains (losses) are excluded from these revenue calculations due to the extreme volatility and occasional negative value. Return on equity and return on assets are calculated using net income that includes trading and non-trading gains (losses). Variable costs are defined as labour expense plus interest expense. Operating costs consist of variable costs plus capital expense. Net interest is interest revenue minus interest expense. Impaired assets comprise mostly of loans that have been in arrears for over 90 days.

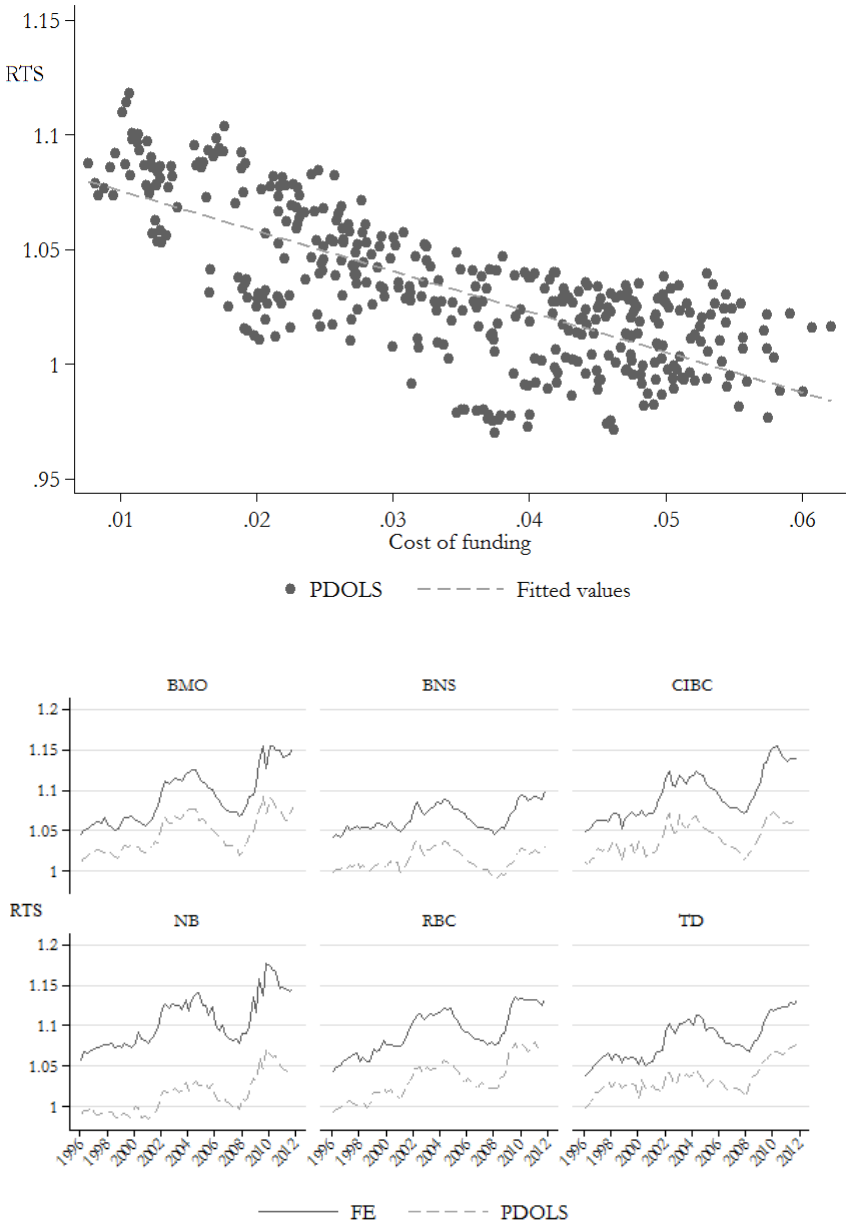


Figure 16: Allen and Liu (2007)model returns to scale

Note: The scatterplot shows returns to scale estimates, using the panel dynamic estimator, on the y-axis graphed against the cost of funding p.a. Periods of low interest rates on deposits, which follow the bank rate, are associated with higher returns to scale (top). It is clear to see that the cost of funding influences the estimates. The next diagram shows RTS by bank over time. Nearly all observations lie above the constant RTS (1.0), and it appears to be increasing over time regardless of the estimator.

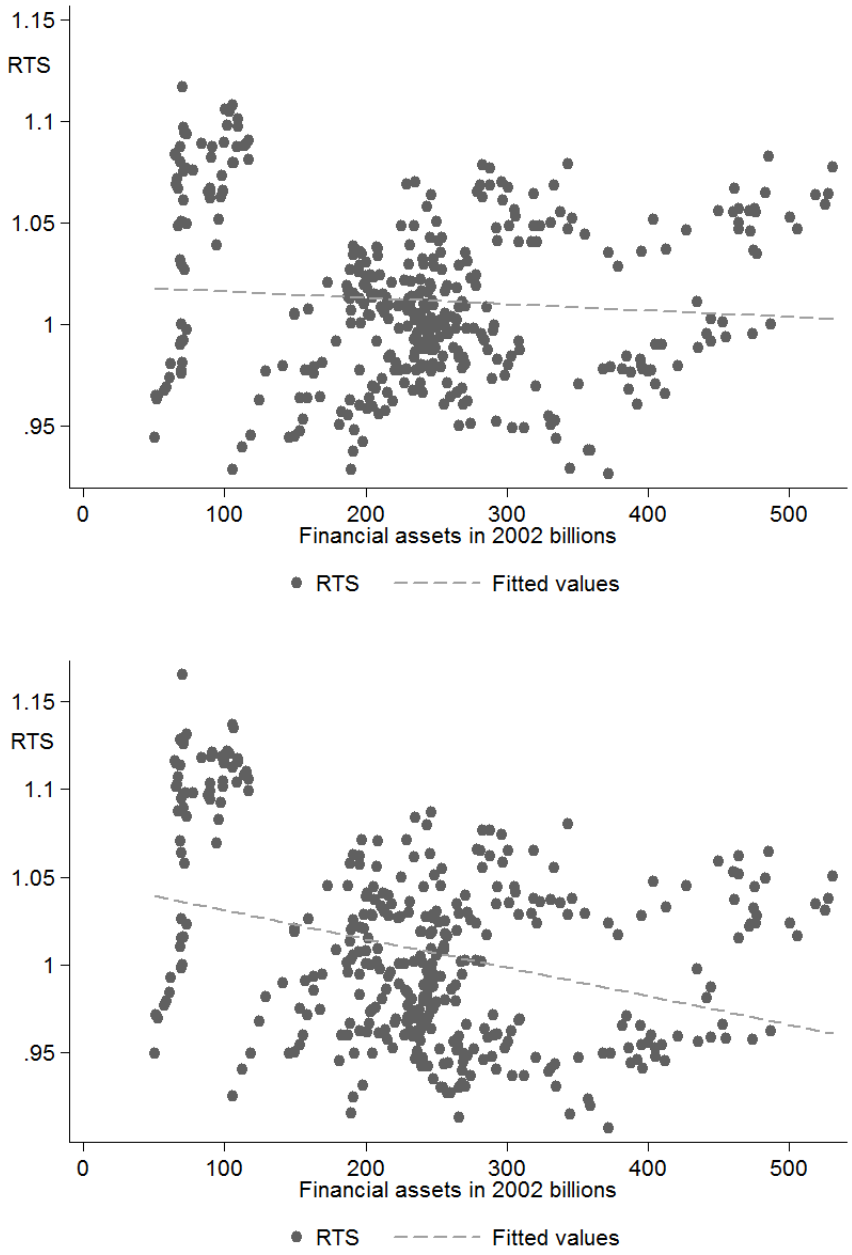


Figure 17: Short-run RTS

Note: A scatterplot of estimated RTS against bank size using the fixed effect estimator (top). Similarly, a scatterplot of estimated RTS against bank size using the panel dynamic estimator (bottom).

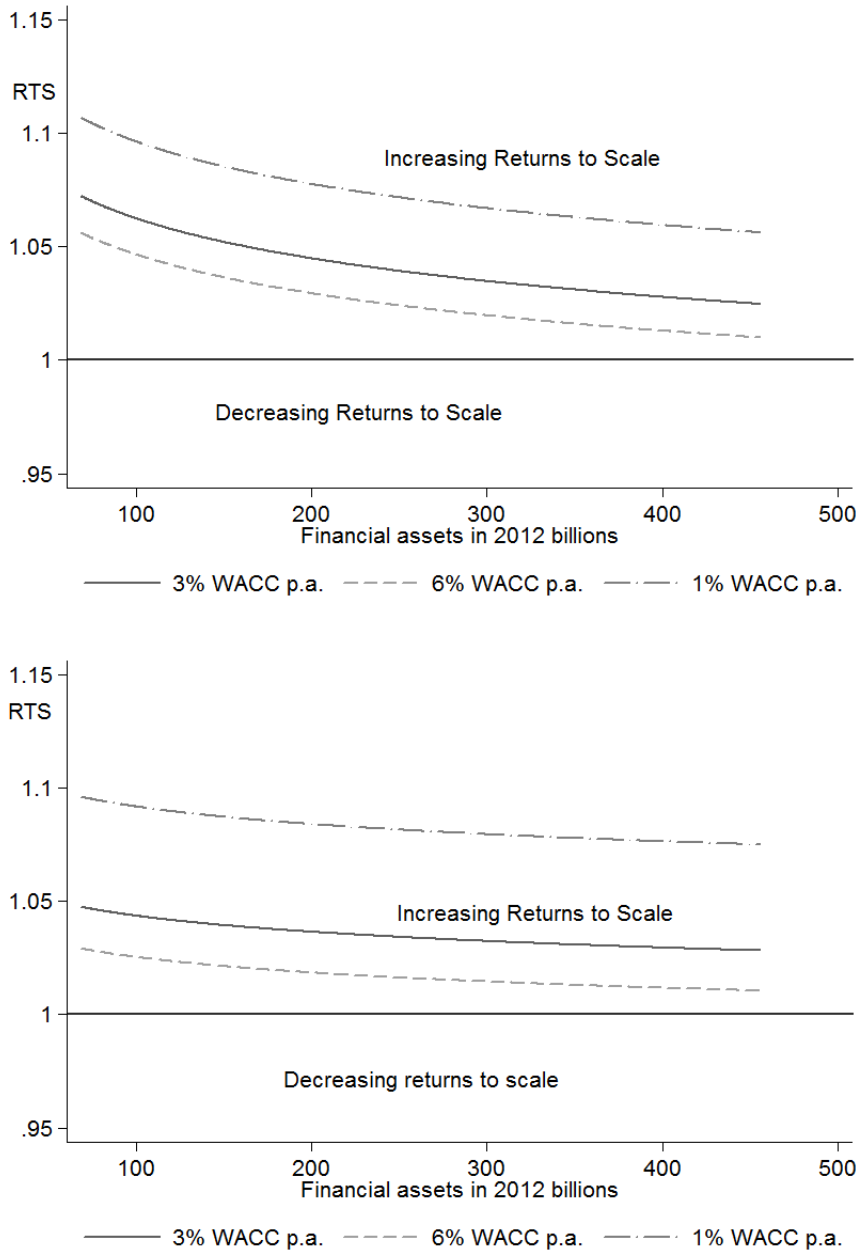


Figure 18: Long-run RTS

Note: Ray-scale RTS estimates with the fixed effect estimator (top) and ray-scale RTS estimates with the panel dynamic estimator (bottom). Using the coefficient estimates from table 8, RTS is evaluated at the median cost of labour and by proportionally increasing output for three different weighted average costs of capital (WACC). See section 3.3 for further details on the calculations.

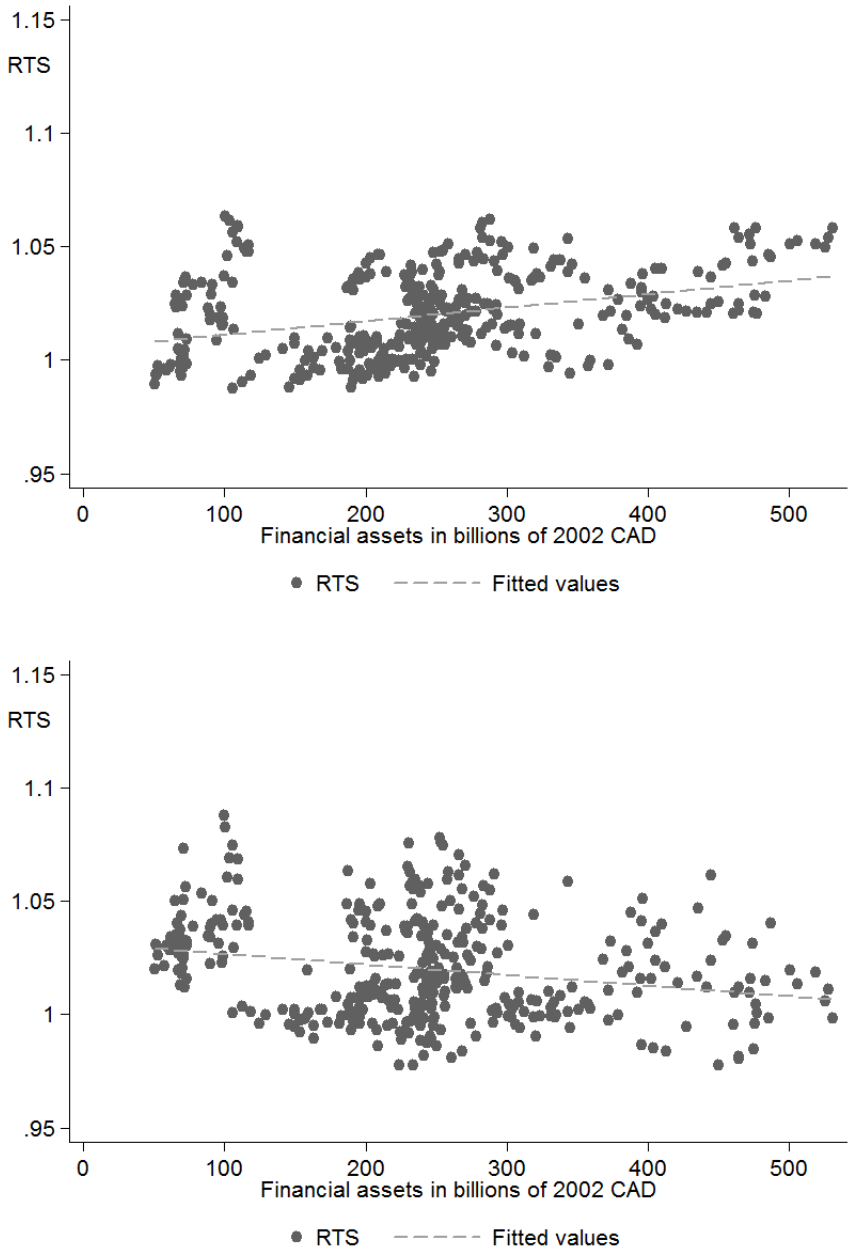


Figure 19: Long-run RTS

Note: A scatterplot of estimated RTS against bank size using the fixed effect estimator (top). Similarly, a scatterplot of estimated RTS against bank size using the panel dynamic estimator (bottom).

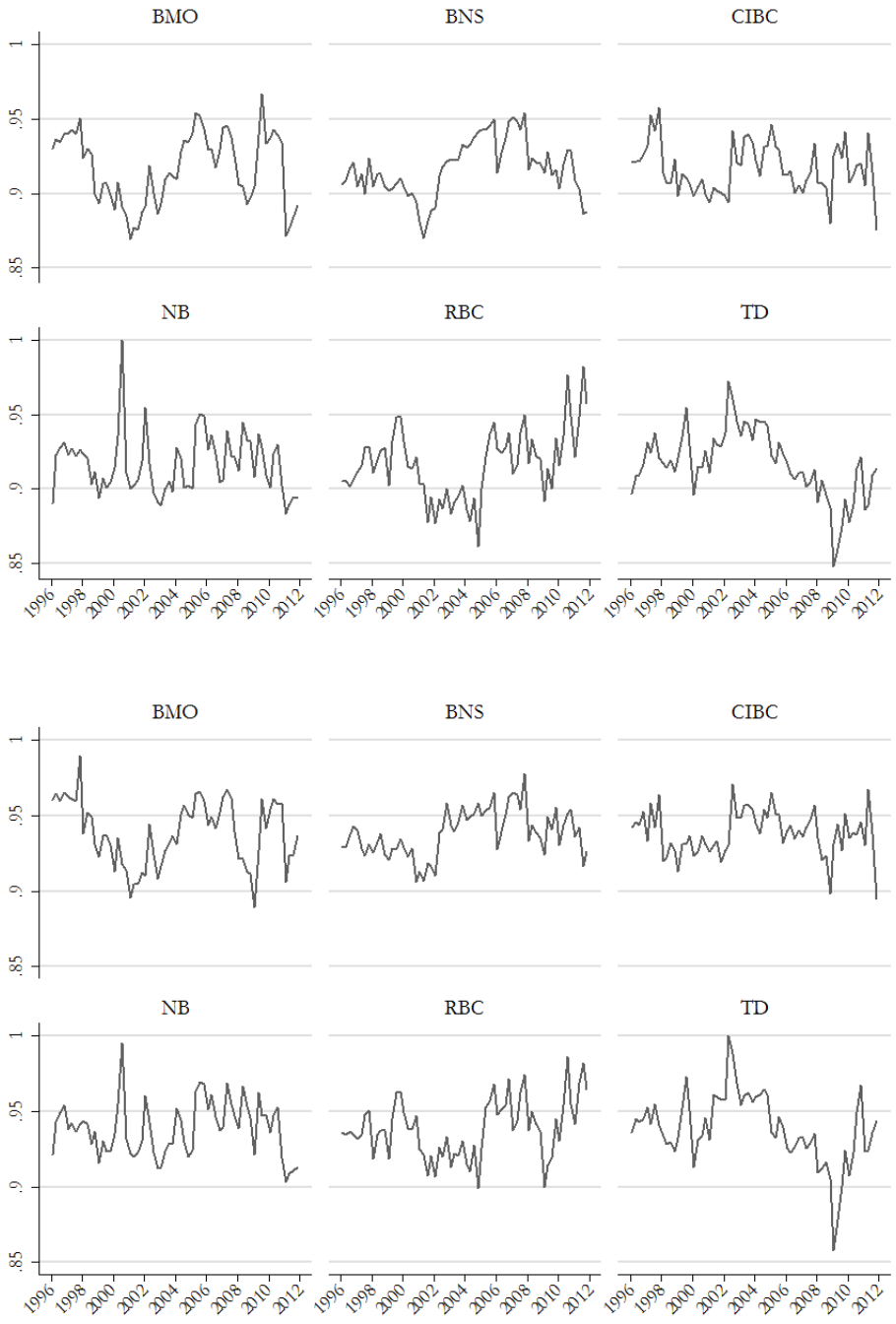


Figure 20: Distribution-free efficiency

Note: Efficiency from the short-run cost function (top) and long-run cost function (bottom) using the fixed effect OLS model.

Table 1: Correlation matrix: loans and net interest rate returns

	Res. Mort. spread	Bus. loan spread	Non-res mort. spread	Consumer loan spread	Net interest spread	Res. Mort.	Bus. Loans	Non-res Mort.	Consumer Loans	Loans, securities
Res. mort. spread	1.00									
Bus. loan spread	-0.56	1.00								
Non-res mort. spread	0.35	-0.29	1.00							
Consumer loans spread	0.14	-0.38	0.28	1.00						
Net interest spread	-0.08	0.05	0.14	-0.07	1.00					
Residential mort.	-0.33	0.31	0.10	0.11	0.11	1.00				
Business loans	-0.17	-0.09	0.16	0.16	0.15	0.56	1.00			
Non-res. mort.	-0.32	0.37	0.07	0.01	0.08	0.47	0.06	1.00		
Consumer loans	-0.29	0.36	0.03	-0.13	0.15	0.76	0.32	0.78	1.00	
Loans, securities	-0.32	0.33	0.07	-0.03	0.19	0.91	0.62	0.58	0.89	1.00

Note: Spreads are calculated as the interest earned from an asset class divided by the dollar amount of that asset outstanding less the weighted average cost of capital. There is little correlation between interest rate spreads and the dollar value of loans outstanding while different asset classes are positively correlated with each other.

Table 2: Adjustments From Canadian GAAP to IFRS

Bank	**Equity**	**Assets**	**NHA-MBS / assets**
BMO	7.3%	-6.2%	3.12%
CIBC	7.3%	-7.7%	6.38%
TD	9.3%	-7.7%	3.97%
BNS	7.0%	-3.3%	5.98%
RBC	9.2%	-5.4%	2.03%
NB	10.7%	-6.4%	0.61%

Equity and assets are the adjustments calculated by Kelly and Janssens (2012). NHA-MBS divided by total assets represent the proportion of assets each bank securitized and sold to CMHC in the third quarter of 2012.

Table 3: Total cost, quantities, and prices

Variable		Description
	Funding	Sum of all demand, notice, chequing, non-chequing and fixed, deposits repurchase agreements and subordinated debt.
	Capital expense	Rental of real estate, premises, furniture and fixtures, computers and equipment.
	Implied equity expense	CAPM estimated cost of equity multiplied by total equity including common shares, contributed surplus and retained earnings
	Net capital assets	Land, buildings, and equipment, less accumulated depreciation
W_1	Price of labour	Labour expense / number of employees
W_2	Pre-tax WACC	(Interest expense + equity expense) / (funding + total equity)
W_3	Price of physical capital	Capital expense / number of employees
C_{SR}	Short-run cost	Labour, interest, and estimated equity expenses
C_{LR}	Long-run cost with physical capital	Labour, deposit, physical capital, and estimated equity expense
Y_1	Government & business securities & loans	Securities issued or guaranteed by Canada, a province or a municipality, deposits with regulated financial institutions less allowance for impairment, deposits with the Bank of Canada, to Canadian federal government, provinces, municipal or school corporations and reverse repurchase agreements, corporate securities corporate securities, business loans, non-residential mortgage loans.
Y_2	Loans to households	Consumer loans and residential mortgages: both insured and uninsured.
Y_3	noninterest income	Credit and debit card service fees, mortgage, standby, commitment and other loan fees, acceptance, guarantees and letter of credit fees, Investment management and custodial services, Mutual(investment) fund, underwriting on new issues and securities commissions and fees, Foreign exchange revenue other than trading and other income (including investment banking fees and securitization income.)

Cost is the dependent variable and varies whether physical capital expense is included (C_{LR}) or excluded (C_{SR}). Similarly, only if the total cost includes physical capital will the price of physical capital be an independent variable. See section 4 for more details. Descriptions coincide with definitions from the Office of Superintendent of Financial Institutions.

Table 4: Allen and Liu (2007) outputs and prices

Y_1	Consumer loans	W_1	Labour expense / employees
Y_2	Non-mortgage loans	W_2	Capital expense / capital
Y_3	Mortgage loans	W_3	Deposit expense / deposits
Y_4	Other	C	sum of labour, capital & deposit expense
Y_5	noninterest income (BG asset)		

Note: Non-mortgage loans include loans to businesses, financial institutions, dealers, brokers, lease receivables, loans to foreign and domestic governments and reverse repurchase agreements. Other assets includes government issued bonds, loans to government, equity shares and corporate bonds.

Table 5: Allen and Liu (2007) model returns to scale

Estimator	**RTS**	**Stat.**	**P-value**	**obs.**	**IRTS**	**CRTS**	**DRTS**
FE	1.069	38.01	0.0000	384	250	134	0
PDOLS	1.032	1.45	0.2287	372	83	289	0

Note: RTS stands for returns to scale. All RTS values are tested using an null hypothesis that RTS = 1. When RTS ¿ 1 then a 1% increase in each output would increase total costs by less than 1% and this is defined as increasing returns to scale over the sample period. All tests are two-tailed. A LR-test using the FE model and the TFE model with a half-normal error term fails to reject the null of no expected inefficiency. IRTS, CRTS, and DRTS represent the number of observations that have increasing, constant, or decreasing statistically significant returns to scale at the 5% level of confidence.

Table 6: Levin-Lin-Chu Modified ADF unit root tests

Model	Statistic	p-value
SR	-2.79	0.0026
LR	-1.47	0.0714

Note: Levin-Lin-Chu unit-root tests for unit root in the residuals from the fixed-effect model. Augmented-Dickey Fuller lags are chosen by the Akaike Information Criterion. Adjusted t-statistics are shown. The null hypothesis is that the panels contain unit roots.

Table 7: Modified Augmented Dickey-Fuller test for cointegration

Short-run cost				**Long-run cost**			
Obs.	Lags	MADF	5% CV	Obs.	Lags	MADF	5% CV
62	2	43.25	19.71	62	2	52.74	19.71
60	4	30.49	19.93	60	4	32.56	19.93
58	6	23.23	20.16	58	6	26.67	20.16
56	8	26.01	20.41	56	8	33.94	20.41

Note: The short-run and long-run fixed effect model estimated by OLS is tested for cointegration. The sample period covered 1996-2011. The null hypothesis that the series is not integrated of order I(1) is rejected at the 5% level of significance for both sample sizes.

Table 8: Short-run and long-run cost function estimates

	SR - FE		SR - PDOLS			LR - FE		LR - PDOLS	
	Coef.	Std. Err	Coef.	Std. Err		Coef.	Std. Err	Coef.	Std. Err
y_1	0.85	0.651	0.13	0.703	y_1	-0.19	0.746	-0.79	0.740
y_2	-2.43	0.479	-2.02	0.650	y_2	-2.30	0.560	-2.05	0.668
y_3	1.82	0.386	2.61	0.417	y_3	2.26	0.455	3.51	0.432
w_2	1.34	0.154	1.42	0.178	w_2	1.30	0.177	1.26	0.191
					w_3	-2.25	0.770	-1.26	0.647
y_1w_2	0.02	0.020	0.04	0.021	y_1w_2	0.02	0.021	0.04	0.021
					y_1w_3	0.21	0.082	0.16	0.070
y_2w_1	-0.12	0.056	-0.11	0.058	y_2w_1	-0.20	0.099	-0.31	0.084
y_2w_2	0.01	0.018	0.01	0.018	y_2w_2	0.02	0.021	0.03	0.018
					y_2w_3	0.08	0.078	0.20	0.066
y_3w_1	-0.02	0.023	-0.01	0.024	y_3w_1	0.24	0.086	0.42	0.074
y_3w_2	0.02	0.022	-0.02	0.021	y_3w_2	-0.01	0.025	-0.03	0.021
					y_3w_3	-0.24	0.087	-0.41	0.074
w_1w_2	-0.15	0.017	-0.15	0.016	w_1w_2	-0.09	0.034	-0.12	0.028
					w_1w_3	-0.14	0.173	0.01	0.144
w_2w_2	0.09	0.006	0.10	0.006	w_2w_2	0.10	0.006	0.10	0.006
					w_2w_3	-0.07	0.027	-0.04	0.023
					w_3w_3	0.10	0.081	0.05	0.066
y_1y_1	-0.09	0.089	-0.18	0.093	y_1y_1	-0.03	0.108	-0.06	0.099
y_2y_2	0.07	0.093	-0.20	0.084	y_2y_2	-0.01	0.102	-0.24	0.085
y_3y_3	-0.02	0.057	0.01	0.055	y_3y_3	-0.10	0.063	-0.07	0.057
y_1y_2	0.11	0.078	0.33	0.076	y_1y_2	0.15	0.092	0.33	0.078
y_1y_3	-0.06	0.058	-0.17	0.053	y_1y_3	-0.11	0.066	-0.26	0.055
y_2y_3	-0.01	0.053	0.02	0.050	y_2y_3	0.04	0.059	0.10	0.053
t	-.0024	.0002	-.0026	.0004	t	-.003	.0005	-.003	.0003
$cons$	12.42	2.651	-	-	$cons$	16.77	3.024	-	-

Note: In the short-run cost model, the dependent variable (total cost) is the sum of interest, labour and the implied equity expense. In the long-run model, the dependent variable is the sum of interest, labour, implied equity, and physical capital expenses.

Table 9: Returns to scale and scope

		Returns to scale			**Returns to scope**		
Sample	**Estimator**	**Scale**	**Stat.**	**P-value**	**Scope**	**Stat.**	**P-value**
SR	FE	1.006	0.021	0.6507	0.077	.19	0.6606
SR	PDOLS	0.996	0.03	0.8701	0.362	4.59	0.0321
LR	FE	0.982	0.27	0.6012	0.073	.1845	0.6012
LR	PDOLS	0.988	0.28	0.5997	0.316	3.78	0.0520

Note: The dependent variable for the short-run cost function (SR) is the sum of interest, labour and estimated equity expense. The long-run cost function estimates include physical capital expense. RTS is an acronym for returns to scale: if RTS ¿ 1 then a 1% increase in all output would increase total costs by less than 1% and there are increasing returns to scale exist. The null hypothesis is that constant returns to scale (RTS = 1) are present. Returns to scale from the Greene (2005) true fixed effects estimator are similar to the fixed effect OLS estimator and consequently omitted from this table. Returns to scale are presented at the mean level of prices and outputs. Economies of scope are present if the scope is less than zero. Testing for statistically significant returns to scope is equivalent to testing that the cross-product output terms are jointly equal to zero. In none of the models presented are returns to scope statistically significant.

Table 10: Returns to scale summarized by year

	Short-run cost function				*Long-run cost function*			
	FE		PDOLS		FE		PDOLS	
Year	RTS	Std. Dev.	RTS	Std. Dev.	RTS	Std. Dev.	RTS	Std. Dev.
1996	0.972	0.028	0.942	0.015	0.991	0.007	0.995	0.007
1997	0.997	0.025	0.972	0.018	0.999	0.010	0.999	0.007
1998	1.005	0.036	0.985	0.034	0.998	0.012	0.998	0.008
1999	1.011	0.034	0.994	0.033	1.001	0.011	1.003	0.007
2000	1.022	0.044	1.017	0.036	1.007	0.010	0.997	0.007
2001	1.023	0.045	1.015	0.030	1.009	0.012	1.002	0.008
2002	1.040	0.059	1.025	0.041	1.026	0.014	1.020	0.010
2003	1.030	0.058	1.014	0.038	1.029	0.012	1.021	0.011
2004	1.028	0.060	1.015	0.042	1.034	0.013	1.025	0.013
2005	1.012	0.061	1.005	0.042	1.028	0.010	1.018	0.013
2006	0.994	0.059	0.992	0.044	1.020	0.012	1.007	0.011
2007	0.994	0.058	0.997	0.045	1.013	0.013	1.004	0.010
2008	0.986	0.067	0.986	0.053	1.014	0.012	1.012	0.015
2009	1.011	0.073	0.999	0.052	1.041	0.021	1.029	0.015
2010	1.012	0.070	1.002	0.047	1.048	0.017	1.032	0.014
2011	1.013	0.060	1.010	0.040	1.047	0.014	1.027	0.012
Total	1.009	0.056	0.998	0.044	1.019	0.021	1.012	0.016

Note: Mean returns to scale were presented by year. Each year contained 24 observations. RTS are relatively flat across the sample period.

Table 11: Fitted and tested RTS estimates

Model / estimator	DRTS	IRTS	CRTS	Obs.
Long-run FE	34	152	198	372
Long-run PD	57	99	216	384
Short-run FE	90	74	220	372
Short-run PD	51	11	310	384

Note: The columns represent decreasing, increasing and constant returns to scale respectively. Each observation is tested for returns to scale at the 5% level of significance.

Table 12: Distribution-free summary of bank efficiency

Half-normal efficiency with MLE

	Short-run		*Long-run*	
	Mean	Std. Dev.	Mean	Std. Dev.
BMO	1	0	0.979	0.013
CIBC	1	0	0.981	0.007
TD	1	0	0.979	0.013
BNS	1	0	0.981	0.009
RBC	1	0	0.978	0.013
NB	1	0	0.980	0.009
Total	1	0	0.980	0.011

Distribution-free approach

	Short-run		*Long-run*	
	Mean	Std. Dev.	Mean	Std. Dev.
BMO	0.927	0.024	0.914	0.024
CIBC	0.927	0.016	0.926	0.019
TD	0.927	0.024	0.929	0.026
BNS	0.927	0.017	0.917	0.019
RBC	0.928	0.026	0.917	0.024
NB	0.927	0.020	0.921	0.023
Total	0.927	0.021	0.921	0.023

Note: Distribution-free approach to efficiency uses the residuals from the fixed effect OLS estimation. The best practices firm has the smallest residual and acts as the benchmark and, by definition, is 100% efficient. The relative distance from this benchmark determines the efficiency of every other observation.

Table 13: Efficiency using Greene's 2005 TFE ML estimator

SR FE residuals		*LR FE residuals*	
Skewness	Kurtosis	Skewness	Kurtosis
0.0567	3.443	0.0431	3.098

Jarque–Bera Skewness and Kurtosis tests for Normality

Model	Pr(Skewness)	Pr(Kurtosis)	χ^2	Joint p-value
SR	0.6442	0.0908	3.08	0.2140
LR	0.6955	0.2590	1.43	0.4881

LR Test for inefficiency

Model	FE LL	TFE LL	LR test	P-value	E(exp(-u))
SR	909.5	909.5	0.000	0.9999	99.9%
LR	906.4	907.1	6.977	0.1082	98.0%

Note: There is little skewness in the residuals although there is more in the long-run than short-run cost function. E(exp(-u) represents the unconditional expected technical efficiency. The Jarque–Bera test for normality shows little evidence for skewness in either model. For the short-run cost model, estimating a model with an exponentially distributed error term is not statistically significant compared to a model without. The long-run model with physical capital is statistically significant however the magnitude is quite small. The average bank is 98.9% efficient.

Table 14: Alternative model robustness testing

Short-run cost function

	Alternate 1		Alternate 2		Alternate 3	
	RTS	P-value	RTS	P-value	RTS	P-value
FE	1.005	0.751	1.012	0.7764	0.986	0.0080
PDOLS	1.014	0.595	1.003	0.8907	1.000	0.9816

Long-run cost function

	Alternate 4		Alternate 5		Alternate 6	
	RTS	P-value	RTS	P-value	RTS	P-value
FE	.999	0.9765	0.976	0.7317	0.980	0.1972
PDOLS	1.006	0.8053	0.981	0.3986	0.978	0.1464

Note: Alternate model 1 excludes the National Bank. Alternate 2 excludes the implied cost of equity from the price of funds. Alternate 3 estimates the short-run cost model with the implied cost of equity as a separate price from the interest and deposit expense. Alternative 4 estimates the long-run cost function without the National Bank. Alternate 5 estimates the long-run cost function without the implied cost of equity. Lastly, alternative 6 estimates separates output into loans to households, securities and loans to businesses (excluding reverse repurchase agreements) and loans to foreign governments, noninterest income less retail fees and other financial assets (cash, government bonds, reverse repurchase agreements).

Table 15: Ryan(2000) short-run cost function

Estimator	Concavity	Monotonicity	Cost % Δ	RTS
		No Normalization		
FE	0%	96.1%	1.75%	1.003
PDOLS	0%	92.4%	1.84%	0.993
		National Bank: 2001 Q2		
FE	88.0%	88.8%	0.895%	1.009
PDOLS	89.1%	82.6%	0.911%	0.999
		Bank of Nova Scotia: 2000 Q1		
FE	88.3%	90.9%	0.925%	1.006
PDOLS	85.7%	85.2%	0.938%	0.989

Note: The short-run cost function is re-estimated with outputs and prices normalized by one observation with Q representing the fiscal quarter. Cost % Δ shows the percentage change in total cost from a 1% increase in all input prices. This should be equal to one in order to satisfy the constraint. RTS represents returns to scale and the percentage represents how many observations satisfy price concavity and price monotonicity respectively. Cost % Δ shows the percentage change in total cost from a 1% increase in all prices. Q represents the fiscal quarter. This table summarizes selected results.

Table 16: Ryan(2000) long-run cost function properties

Estimator	Concavity	Monotonicity	Cost % Δ	RTS
		No Normalization		
FE	0%	100%	2.78%	0.998
PDOLS	0%	100%	3.24%	1.020
		BMO: 1997 Q1		
FE	0%	97.4%	0.969%	0.995
PDOLS	0%	86.7%	0.953%	1.026
		CIBC: 1997 Q1		
FE	0%	99.7%	1.019%	0.994
PDOLS	0%	85.2%	1.015%	1.029

Note: The short-run cost function is re-estimated with outputs and prices normalized by one observation with Q representing the fiscal quarter. Cost % Δ shows the percentage change in total cost from a 1% increase in all input prices. This should be equal to one in order to satisfy the constraint. RTS represents returns to scale and the percentage represents how many observations satisfy price concavity and price monotonicity respectively. Cost % Δ shows the percentage change in total cost from a 1% increase in all prices. This table summarizes selected results.