# Who said large banks experience scale economies? A critical view on the usage of a risk-return driven cost function

Andrei Dubovik[1,2]

[1]CPB The Netherlands Bureau for Economic Policy Analysis
[2]Erasmus University Rotterdam

February 10, 2017

**Abstract**

I revisit the definition of scale economies when various input vectors are associated with various levels of risk taking. This issue is most relevant for the empirical research on scale economies in the banking industry. I show how the commonly used definition only partially accounts for the price of risk-taking. Adopting a definition more aligned with social welfare might change the conclusions on whether banks are characterized by increasing economies of scale. Therefore, current empirical literature on banks' scale economies should be taken with a grain of salt when used in policy discussions.

## 1 Introduction

Big banks are potentially more valuable than small banks for two reasons: big banks can underwrite big loans and big banks might benefit from scale economies. The question of scale economies is being discussed extensively in the literature and no stylized consensus has been reached so far, see, e.g., Beccalli et al. (2015). One of the complications in estimating scale economies in the banking industry is caused by the endogenous risk taking: bigger banks take bigger risks. In Hughes et al. (1996), Hughes et al. (2001), and Hughes and Mester (2013) the authors argue that bigger risks result in higher costs. Therefore, ignoring the effect that endogenous risk has on costs yields underestimated scale economies. In particular, Hughes and Mester (2013) implicitly model managerial choice and find substantial scale economies in the banking industry, more so for larger banks.

As is explained in more detail in Hughes et al. (2001), a way to estimate scale economies in the presence of endogenous risk-taking is to compute

scale economies along the value-maximization path, where the value of a bank captures not only the bank's expected profits but also its risks. However, this approach raises an important policy dilemma. If we compute scale economies along the value-maximization path, thus admitting that risk-taking plays a significant role, should we adjust the definition of scale economies to account for the price of this risk-taking, or should we continue to define scale economies solely through expected profits, as is done in the aforementioned papers?

Suppose a manager of a bigger bank chooses more risky liabilities. Riskier liabilities are likely to be cheaper for the bank if we ignore the price of risk, therefore we obtain lower expected costs. Are these scale economies? Arguably, the search for scale economies receives its importance from the policy discussions about making big banks smaller. Therefore let us take the society's perspective on what constitutes a proper definition of scale economies.

If the extra risks that banks are taking are idiosyncratic, and thus do not depress the social welfare, then defining scale economies based solely on expected profits is correct. However, if the managerial behaviour leads to an increase in the systemic risk, then this risk should be included in the definition of scale economies. As I formally show in this note, including the price of risk-taking in the definition of scale economies can, in principle, reverse the positive conclusions in Hughes and Mester (2013).

The rest of this note is split into two sections. In Section 2 I restate the aforementioned logic formally. In Section 3 I give a simple example of a bank production function and illustrate how the approach adopted in Hughes and Mester (2013), etc. can yield overestimates of scale economies.

## 2 General Exposition

A bank uses a vector of inputs $\mathbf{x}$ to produce a vector of outputs $\mathbf{y}$. Admissible production plans are given by $T(\mathbf{x}, \mathbf{y}) \leq 0$. The output vector includes outputs with various credit risk profiles, the input vector includes inputs with various liquidity risk profiles. The prices are given by $\mathbf{p} = \{\mathbf{p}_x, \mathbf{p}_y\}$. I model both credit and liquidity risks by assuming that the prices are stochastic. This approach is straightforward when we speak of credit risks. If we speak of liquidity risks, we might consider that the bank needs to resort to more expensive liabilities if a liquidity risk is realized. Therefore liquidity risk can be indirectly modelled with stochastic input prices.

The profits are given by

$$\pi(\mathbf{x}, \mathbf{y}, \mathbf{p}) = \mathbf{p}_y \cdot \mathbf{y} - \mathbf{p}_x \cdot \mathbf{x}. \tag{1}$$

and the value of the bank is

$$V(\mathbf{x}, \mathbf{y}) = \mu(\pi(\mathbf{x}, \mathbf{y}, \mathbf{p})) - \lambda\sigma^2(\pi(\mathbf{x}, \mathbf{y}, \mathbf{p})), \tag{2}$$

where $\lambda$ is the price of risk. A more general specification could be given within the expected utility framework but I choose the mean-variance framework for the ease of exposition. The manager of the bank maximizes $V$ in $\mathbf{x}$ and $\mathbf{y}$.

Define

$$c(\mathbf{x}, \mathbf{p}_x) = \mathbf{p}_x \cdot \mathbf{x}, \tag{3}$$

$$vc(\mathbf{x}) = \mu(c(\mathbf{x}, \mathbf{p}_x)) + \sigma^2(c(\mathbf{x}, \mathbf{p}_x)). \tag{4}$$

Suppose further that $\text{cov}(\mathbf{p}_x, \mathbf{p}_y) = 0$ and define

$$\tilde{\mathbf{x}}(\mathbf{y}) = \underset{\mathbf{x}:T(\mathbf{x},\mathbf{y})\leq 0}{\arg\max}\ \mathbb{E}\pi(\mathbf{x}, \mathbf{y}, \mathbf{p}) = \underset{\mathbf{x}:T(\mathbf{x},\mathbf{y})\leq 0}{\arg\min}\ \mathbb{E}c(\mathbf{x}, \mathbf{p}_x), \tag{5}$$

$$\hat{\mathbf{x}}(\mathbf{y}) = \underset{\mathbf{x}:T(\mathbf{x},\mathbf{y})\leq 0}{\arg\max}\ V(\mathbf{x}, \mathbf{y}) = \underset{\mathbf{x}:T(\mathbf{x},\mathbf{y})\leq 0}{\arg\min}\ vc(\mathbf{x}). \tag{6}$$

Then $\tilde{\mathbf{x}}(\mathbf{y})$ is the choice of inputs that minimizes expected costs given the desired output $\mathbf{y}$, and $\hat{\mathbf{x}}(\mathbf{y})$ is the choice of inputs that minimizes risk-adjusted costs.

There are three possibilities to define scale economies $\varepsilon$:

$$\tilde{\varepsilon} = 1/\left(\frac{\partial \ln \mathbb{E}c(\tilde{\mathbf{x}}(\mathbf{y}), \mathbf{p}_x)}{\partial \mathbf{y}} \cdot \mathbf{y}\right), \tag{7}$$

$$\varepsilon_{hm} = 1/\left(\frac{\partial \ln \mathbb{E}c(\hat{\mathbf{x}}(\mathbf{y}), \mathbf{p}_x)}{\partial \mathbf{y}} \cdot \mathbf{y}\right), \tag{8}$$

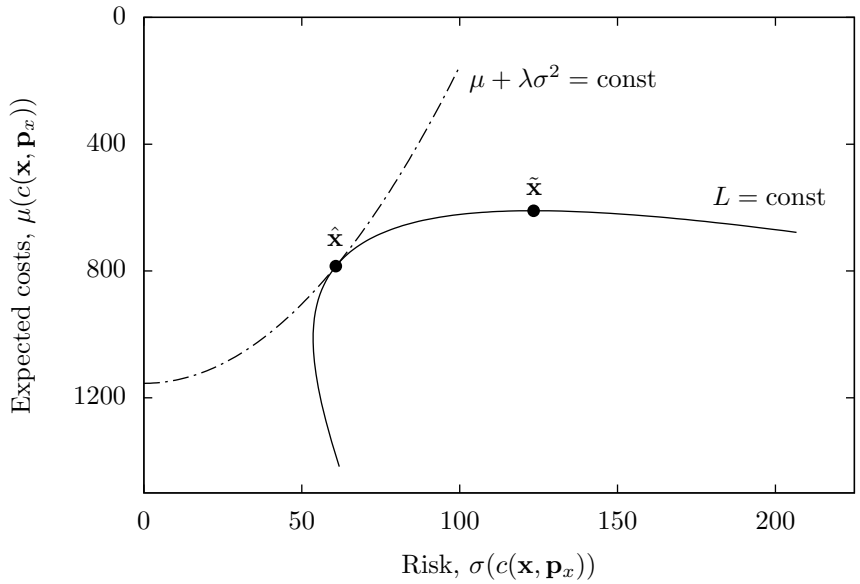
$$\hat{\varepsilon} = 1/\left(\frac{\partial \ln vc(\hat{\mathbf{x}}(\mathbf{y}))}{\partial \mathbf{y}} \cdot \mathbf{y}\right). \tag{9}$$

In either case $\varepsilon > 1$ means increasing economies of scale and $\varepsilon < 1$ means decreasing economies of scale.

The first definition is based on the cost function and is often used in the literature, with Davies and Tracey (2014) being a recent example. However, this definition completely disregards the price of risk and is therefore subject to the critique by Hughes et al. (1996), Hughes et al. (2001), and Hughes and Mester (2013). These papers, in turn, adopt the second definition, which mixes two different approaches. On one hand, this definition disregards the price of risk in the computation of scale economies. On the other hand, it computes scale economies along the value-maximization path, i.e. along the path where the price of risk is accounted for by the decision maker. Finally, the third definition computes scale economies both along the value-maximization path and including the price of risk in the definition itself.

Scale economies in the banking industry are used as one of the arguments against the policy suggestions of splitting big banks. If different approaches to defining scale economies bring along different results, these differences cannot be neglected. The next session gives an example where $\tilde{\varepsilon} > \varepsilon_{hm} >$

Figure 1: Cost optimization




The parameters are as follows: $L = 50$, $\alpha = 2$, $\beta = 10$, $\lambda = 0.1$, $\mu(p_e) = 12$, $\mu(p_d) = 8$, $\sigma(p_e) = 0.5$, $\sigma(p_d) = 4$.

$1 > \hat{\varepsilon}$. Depending on the chosen perspective, Hughes and Mester (2013) have either done it correctly, underestimated, or overestimated scale economies of banks.

# 3 An Example

Consider a simple bank specification with cash ($C$) and loans ($L$) on the assets side and equity ($E$) and debt ($D$) on the liabilities side. Let the production technology be given by

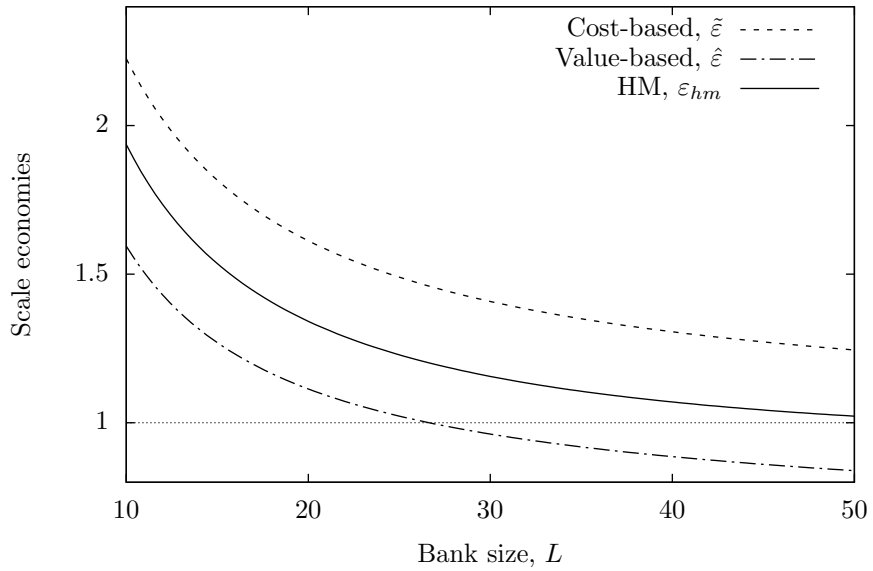
$$L = \alpha\sqrt{(E - \beta)D}. \tag{10}$$

The cash holding then follow from the balance equation: $C = E + D - L$. I will consider economies of scale with respect to $L$ alone as opposed to $\{L, C\}$ as otherwise the optimization problem is degenerate and more variables need to be added to the model.

With this technology function we can solve explicitly for $\tilde{\varepsilon}$, $\hat{\varepsilon}$ and $\varepsilon_{hm}$. The related equations involve polynomials of $4^{\text{th}}$ degree and while the explicit solutions are straightforward they are also long, and are therefore not shown here. Instead I present two figures of interest.

Fig. 3 illustrates the optimization problems (5) and (6). A given amount of loans can be achieved with various combinations of equity and debt.

Figure 2: Scale economies




The parameters are as follows: $\alpha = 2$, $\beta = 10$, $\lambda = 0.1$, $\mu(p_e) = 12$, $\mu(p_d) = 8$, $\sigma(p_e) = 0.5$, $\sigma(p_d) = 4$.

Each combination has specific expected costs and risks associated with it. The corresponding isoline is denoted $L = \text{const}$. If the manager minimizes expected costs, then the optimal solution is given by $\tilde{\mathbf{x}}$. If the manager assigns value to the risk and minimizes $vc(E, D)$, then the optimal solution is given by $\hat{\mathbf{x}}$ and entails higher costs but lower risks than $\tilde{\mathbf{x}}$.

Fig. 3 plots $\tilde{\varepsilon}$, $\hat{\varepsilon}$ and $\varepsilon_{hm}$ with respect to $L$. The scale economies are decreasing with bank size but that directly follows from the chosen production function. What is more noteworthy is that different definitions of scale economies yield quantitatively and qualitatively different results. In this particular example, $\tilde{\varepsilon} > \varepsilon_{hm} > \hat{\varepsilon}$. Moreover, $\varepsilon_{hm} > 1 > \hat{\varepsilon}$ for $L > 26.66$.

Endogenous risk taking needs to be accounted for when estimating scale economies in the banking industry as otherwise the estimates are inconsistent. However, acknowledging this extra factor presents new challenges when defining scale economies themselves as different definitions can yield very different results. Consequently, previous empirical estimates need to be considered with a grain of salt as they are subject to this critique.

# References

Beccalli, E., Anolli, M., and Borello, G. (2015). Are european banks too big? evidence on economies of scale. *Journal of Banking & Finance*,

58:232–246.

Davies, R. and Tracey, B. (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(1):219–253.

Hughes, J. P., Lang, W., Mester, L. J., and Moon, C.-G. (1996). Efficient banking under interstate branching. *Journal of Money, Credit and Banking*, 28(4):1045–1071.

Hughes, J. P. and Mester, L. J. (2013). Who said large banks dont experience scale economies? evidence from a risk-return-driven cost function. *Journal of Financial Intermediation*, 22:559–585.

Hughes, J. P., Mester, L. J., and Moon, C.-G. (2001). Are scale economies in banking elusive or illusive? evidence obtained by incorporating capital and risk-taking into models of bank production. *Journal of Banking & Finance*, 25:2169–2208.