VAT Treatment of Financial Institutions: Implications for the Real Economy

Fatih Yilmaz*

Department of Economics
University of Calgary
(Job Market Paper)

October 29, 2013

Abstract

For various technical reasons, most countries have subjected financial institutions to exempt treatment under the Value Added Tax (VAT), whereby they pay non-recoverable VAT on their inputs but do not charge VAT on their outputs. Recent studies claim that such treatment leads to several distortions and inefficiencies in the economy. This paper theoretically identifies these distortions and provides numerical analysis from policy simulations under different scenarios in a Melitz type general equilibrium model. The numerical simulations show that moving from exempt treatment to hypothetical full taxation, holding tax revenue constant, reduces the cost of financial intermediation, increases the number of the firms in equilibrium, which increases competition leading to lower market prices and higher output. Overall welfare is higher. Finally, the impact of exempt VAT on the cost of borrowing (in the context of marginal effective tax rate on capital) is also examined.

Key Words: VAT, Financial Services, Exempt Treatment, General Equilibrium, Heterogeneous Firms
JEL Codes: H20 H22 H24 H25 H30

*Email: fyilmaz@ucalgary.ca I am very grateful to Kenneth J. McKenzie for his great supervision. This paper has benefited from comments of the seminar participants at 2011 CEA Meeting in Calgary and at the Brown Bag Seminar Series, Department of Economics in University of Calgary in Summer 2013. I like to particularly thank my colleagues Jevan Cherniwchan, Irving Rosales, Ali Shajarizadeh and Trevor Tombe for their very valuable feedbacks. All remaining errors are my own.

1 Introduction

Exempt treatment has been the foremost practice in taxing financial services under the value added tax (VAT) all over the world.[1] Under this treatment financial institutions do not charge VAT on their sales, but likewise do not receive input credits for VAT paid on their inputs. As such, the VAT is embedded in the price of exempt financial services. The resulting increase in the price of financial services creates several direct and indirect distortions in the economy. This paper theoretically identifies these distortions and provides numerical analysis from policy simulations under different scenarios.

Three key distortions are focused on here, as they have captured the most attention in recent policy studies:[2]

1. Consumption distortions on the part of final consumers;
2. Input distortions and self-supply bias in the financial sector;
3. Input distortions and tax cascading in the business sector.

Given the role of financial institutions as intermediaries, these distortions can best be captured in a general equilibrium setting, and that is the case here.[3] The model is then simulated under reasonable assumptions regarding parameter values.[4] A numerical comparative steady state analysis is performed in the form of a policy exercise: shifting the VAT policy from exempt treatment to full taxation, holding tax revenue constant. The results show that such a policy shift leads to a decrease in the loan interest rate, which lowers entry barriers, especially for micro and small firms. The entry of new firms increases competition in the industry, which decreases the market price and increases total output. As a result, welfare is higher under full taxation. The policy exercise is also repeated for intermediate cases with partial recovery instead of full taxation. Results are qualitatively the same however the magnitude is lower.

In contrast to the existing literature, this paper considers both the intensive and extensive margin movements of firms in response to changes in the tax regime.[5] This is important because firms with different characteristics (e.g., size, profit margin, productivity,

[1] For a cross-country policy discussion, see Gendron (2007) and Schenk (2009).

[2] The VAT literature, much of which is non-technical and informal, identifies four key distortions that result from the exempt treatment of the financial sector (McKenzie and Firth (2011)). The fourth one, "Import bias and an impediment to international competitiveness", requires an open economy model that is not case in here.

[3] Caminal (2002) discusses the impact of financial tax policies on the economy in a "partial equilibrium" model, where she also recognizes the short comings of such approach and concludes that results are incomplete in a partial equilibrium analysis.

[4] While every effort is made to employ parameter values from the literature, it would be misleading to refer to the exercise as a full calibration because of uncertainty regarding some of the parameters. This is a general problem of the literature, see Kerrigan (2010).

[5] For instance, both Büttner and Erbe (2012) and Chisari et al. (2013) attempt to quantify the welfare and tax revenue consequences of repealing exempt treatment only along intensive margin in a static general equilibrium setup.

etc.,) operating in different industries (e.g., capital intensive, high or low financial service intensity, etc.) are affected differently by changes in the price of financial intermediation. For instance, small and medium sized enterprises (SMEs) and micro firms finance their investments largely through mainstream financial intermediaries, i.e., banks.[6] Larger firms usually have better access to equity and other means of finance (e.g. money markets[7]). Therefore, SMEs may be affected more than larger firms in the event of an increase in the cost of bank loans due to the exempt treatment of financial services. In that regard, SMEs may be forced to exit the market (i.e. the extensive margin) while relatively larger firms are only forced to cut back from production (i.e. the intensive margin). The firm and industry level responses on both the extensive and intensive margins are modelled here in a general equilibrium frame work with heterogenous firms, using the approach developed in the seminal paper by Melitz (2003). This is entirely new in the literature on financial intermediation in public finance.

Most of the papers in the literature tend to focus on one role of financial institutions. This makes the results peculiar to a certain market, or type of distortion. As the goal here is to model several distortions and analyze the general equilibrium affects numerically, the paper combines loan services in the spirit of Russ and Valderrama (2012) and payment services motivated by Lockwood (2010). Thus, banks play two roles in the model: the provision of payment and depository services to consumers (credit card services, online payments, cheque clearance, checking and saving account services, etc.), and the intermediation and administration of loans between savers and firms.

It is clear from the results that business-to-business (B2B) transactions (i.e., transactions between financial institutions and VAT registered businesses - e.g. loan services) are over-taxed under exempt treatment and business-to-consumers (B2C) transactions (i.e., transactions between financial institutions and final consumers - e.g. payment services) are under-taxed. However, tax revenues from both exempt loan services and payment services are small as a percentage of overall VAT revenues. While moving from exempt treatment to full taxation (or partial recovery) reduces the share of unrecovered VAT, it increases the revenue from payment services even more. Overall, the share of tax revenues coming from payment services (with no exemption) becomes larger. On the other hand, repeating the aforementioned policy shift but now holding welfare constant shows that tax revenue is higher under full taxation and/or partial recovery than it is under exempt treatment. This implies that the distortions due to exempt treatment are larger than the distortions created by an increase in VAT applied to B2C transactions under full taxation. Finally, the results also indicate that increasing the VAT rate, holding everything else constant, does not always increase total revenues, as type of Laffer Curve

[6]The term bank is used as a surrogate for all suppliers of financial services through out the paper.

[7]There is a large literature on financing choice, such as pecking order models, and firm size. See Denis and Mihov (2003) on bond issue and firms size. Also, see Houston and James (1996) and Johnson (1997) for evidence on bank loans concentrating among small US firms.

where aggregate tax revenues decline with increases in the VAT at a certain point. The simulations suggest that the peak of the aggregate Laffer Curve is in the range of a VAT rate from 17% to 20%, which intersects with the area of the rate in many countries, particularly EU countries.

Marginal effective tax rates (METR) on capital in the real sector are also computed to capture the interaction of the treatment of financial services under the VAT with overall capital tax system (see McKenzie (2000)). Specifically, an increase in the loan interest rate due to embedded VAT under the exempt approach acts as an implicit tax on real capital. The results suggest that increasing the VAT rate under exempt treatment, holding everything else constant, increases the cost of capital and the METR on capital. In contrast, holding the VAT rate constant but now moving from exempt treatment to full taxation decreases the cost of capital and the METR. Finally, a case study from Ontario is considered. It is found that moving from a 5% Goods and Services Tax (GST) rate coupled with an 8% provincial sales tax (PST) to a 13% harmonized sales tax (HST) under exempt treatment increases the METR on capital. This will primarily affect small firms that rely on bank finance. Overall, the price level increases, and aggregate output and welfare fall, as explained above.

In the following section, a review of the literature on taxing financial institutions under VAT is provided. This is followed by a discussion of the theoretical model. Section 3 and 4 presents aggregation and market clearing conditions. A discussion of the equilibrium is presented in Section 5. Section 6 discusses technical details about parameter values used in the simulations. Sections 7 and 8 summarize the research findings and provides a case study. Section 9 concludes.

1.1 Taxing Financial Services Under VAT

Taxation of financial services under value added taxes (VAT) has been one of the most vexing issues in the public finance literature. The literature can be summarized by considering two main questions:[8] "should we tax financial services?" And if so, "how can we tax them?" Despite the intensive research, there has not been a clear concesus on either question. In many ways, the decision of most countries to adopt exempt treatment for financial services can be viewed as the compromise of choice along several dimensions. Although exempt treatment is practically simple to implement, it creates different problems for the real economy. In this section, the related literature focusing on both questions are discussed along with a brief presentation of the current policy debate.

[8]See McKenzie and Firth (2011) and Poddar (2003).

1.1.1 Should we tax finaiclal services?

In principle, there appears to be no controversy regarding the taxation of B2B transactions; the clear consensus is that they should not be taxed. In principle this can be accomplished in two ways: via full taxation or by zero-rating B2B transactions, which involves providing full input credits for VAT paid on inputs into financial intermediation and a zero tax rate on outputs. However, B2C transactions are not as straight forward and economists have different views on how to treat them. McKenzie and Firth (2011) summarize four different views:

1. No taxation.
2. Full taxation of fee based services, no taxation of margin based services.
3. Full taxation of both fee and margin based services.
4. Taxation of both fee and margin based services but at lower rates.

The first approach "no taxation" is based on the observation that financial services are not final consumption goods and therefore do they show up in consumers' utility functions. They are mostly used to facilitate consumption and therefore should not be taxed.[9] Boadway and Keen (2003) oppose this view, calling it a "fallacy", indicating that there are many other goods, which do not directly increase consumption nor show up in utility functions are taxed under VAT.

Jack (2000) recognizes the practicaly difficulty of implementing a VAT on margin based transactions and suggests fee based services should be taxed but implicit, margin or spread based transactions should be zero-rated. Auerbach and Gordon (2002) find no distinction between margin based and fee based transactions in terms of tax treatment and claim that it is equivalent to a labor income tax. This result is further advocated by Roussiang (2002) with different a set of assumptions. Finally, the last camp suggests that all financial services should be taxed under VAT but perhaps at a lower rate (Lockwood (2010) and Kleven et al. (2000)).

The different views have different points of focus. Those in support of"no taxation" are in general concerned about distorting consumption by increasing the price of services that are in principle used to facilitate it. However, the second camp (in favour of taxing only explicit fee based financial services) claim that while, for example, pizza delivery is a service that does not directly enter into utility, it saves time and resources which can be utilized to do other things that may enter utility. This suggests that they should be taxed (and indeed are in most systems). Yet, there is some hesitation here by those who argue for zero-rating margin based transactions about increasing the cost of capital and thus

[9]For further discussion, see Piggott and Whalley (2001), Grubert and Mackie III (2000) and Chia and Whalley (1999).

distorting production. In contrast, the third and fourth groups stress the importance of neutrality and the uniformity of the VAT system. Poddar (2003) summarizes this view with the statement: "VAT is designed to be a tax consistently applied to all the inputs that contribute to value added". All in all, it is fair to say that the most prevalent view is that in principle financial services should be taxed. While there is some disagreement, the literature seems to have converged on position that in principle financial transactions should be taxed (or zero-rated). The question that has vexed policy makers is precisely how to go about this in a practical way.

1.1.2 How can we tax financial services?

The taxation of financial institutions has perhaps resulted in the most drastic split between the practice and theory of VAT implementation. While methods of bridging this gap have been proposed (see Bird et al. (2005) for a detailed list of proposals), they are typically viewed as being either too complex to implement in practice or do not fully achieve their intended purpose (Kerrigan (2010)). The complications generally arise from identifying and measuring the value added of financial institutions on a transaction-by-transaction basis. Despite intensive efforts, there has not been a sound practical solution proposed for the full taxation of financial institutions, and thus exempt treatment has been seen historically as the only viable option and is the approach taken in virtually all countries.[10]

1.1.3 Current Policy Debate

Aside from the academic and technical discussion, there is also an intensive policy debate regarding the taxation of financial institutions, especially following the recent financial crises. Recent reports, IMF (2010) and EU Commision (2010), claim that financial services are under-taxed and there is a large revenue loss due to exempt treatment. The EU Commission in fact has taken legislative action and is introducing a new financial transaction tax effective January, 2014 to help remedy this. With the support of the biggest banks in the EU, PWC (2011)[11] issued a report (with Ben Lockwood) arguing that exempt treatment already creates several distortions in the economy and that the further taxation of financial institutions may very well exacerbate these distortions for the real economy, including a reduction in tax revenues.

The current policy debate is centred around identifying and quantifying the economic distortions created by exempt treatment. Although, there has been some proposals (see

[10]There are some countries which have attempted to go beyond simple exempt treatment. For instance, New Zealand, and Quebec (of Canada) zero-rate most of the margin based transactions. Belgium, France and Germany allow financial institutions the option of being fully taxed under VAT. South Africa uses "reduced exemption method", where all the fee based services are fully taxed and only some of the margin based transactions are exempt. For further discussion, see Bird et al. (2005).

[11]For a similar but earlier work, see Huizinga (2002)

Büttner and Erbe (2012) and Chisari et al. (2013)), it is still not clear whether repealing exempt treatment and moving to full taxation (even after solving the technical issues) would result in a welfare improvement for the society and/or increase tax revenue. The purpose of this paper is to contribute to the current debate on the matter.

2 The Model

In this section, a simple general equilibrium model with consumers, final good producing sector, banks, bank input sector, intermediate good sector and government is presented. There are simply three markets in the model: (1) labor market, where consumers supply their labor to firms in intermediate good sector, (2) loan/deposit market, where the bank collects all the deposits from consumers and issues loans to firms in the intermediate good sector to finance their investment and finally, (3) final (consumption) goods market, where payment services (e.g. credit/debit card services) - produced and supplied by the bank - are used to facilitate the transactions. Government policy is imposed exogenously and all the tax revenue is returned back to consumers.

2.1 Consumers

A representative consumer maximizes discounted expected lifetime utility

$$V = \sum_{t=0}^{\infty} \beta^t U(C_t, N_t),$$

subject to the budget constraint

$$\begin{aligned} P_t C_t + P_t(1 - t_k\kappa)K_{t+1} = (1 - t_w)w_t N_t + r_t(1 - t_k)P_t K_t \\ + P_t(1 - t_k\kappa)(1 - \phi)K_t + LST_t. \end{aligned} \tag{1}$$

He allocates the final (aggregate) good, Y_t, between consumption (C_t) and savings (K_t), and also determines labor supply (N_t). Consumption and savings are measured in terms of the final good and thus, priced at P_t. All savings are deposited at the bank for a risk free rate of return, r_t, which are then loaned to firms through hybrid leasing type of bank intermediation. Savings are effectively converted to physical capital (e.g. used to finance physical capital) through intermediation. The return to capital is subject to a tax, t_k, along with an investment allowance (e.g. investment subsidy), κ, which proxies the present value of tax depreciation deduction on \$1 spent on capital, net of physical

capital depreciation, ϕ.[12] Labor income, w_t, is also subject to payroll tax, t_w. Finally, LST are lump-sum transfers that consist of firms' profits (Π_t), revenues from the bank input sector (Ex_t) and total tax revenue (TR_t). These will be explained in more detail in the subsequent sections.

Consumer preferences are represented by a contemporaneous utility function suggested by Greenwood et al. (1988):

$$U(C_t, N_t) = \frac{1}{1-\eta}\left[C_t - \frac{N_t^{\psi}}{\psi}\right]^{(1-\eta)}.$$

This formulation eliminates the effect of wealth on labor supply and thus, labor supply depends solely on real wage rate, net of the payroll tax. The first-order condition with respect to labor yields labor supply:

$$N_t^s = \left(\frac{(1-t_w)w_t}{P_t}\right)^{\frac{1}{\psi-1}}. \tag{2}$$

Solving the consumers' maximization problem with respect to consumption and capital gives the inter-temporal Euler Equation:

$$\left[C_t - \frac{N_t^{\psi}}{\psi}\right]^{-\eta} = \beta\left[C_{t+1} - \frac{N_{t+1}^{\psi}}{\psi}\right]^{-\eta}\left[r_{t+1}(1-t_k) + (1-t_k\kappa)(1-\phi))\right]. \tag{3}$$

2.2 Final Goods Producers

Final good producers assemble a continuum of differentiated goods in a perfectly competitive market via a constant elasticity of substitution (CES) production process:[13]

$$Y = \left(\int_{i\in\Omega} y(i)^{(\sigma-1)/\sigma} di\right)^{\sigma/1-\sigma}. \tag{4}$$

Differentiated goods are defined on the interval $(0, G] \in \Omega$, where each variety i is associated with a demand $y(i)$ and a price, $p(i)$. The representative final good producer maximizes profits:

$$\Pi = PY - \int_{i\in\Omega} \bar{p}(i)y(i)di,$$

by choosing $y(i)$ for each variety, subject to its production technology, (4) given the aggregate price index,

[12] These are summarized in the consumers' problem for technical convenience as they are the owners of the firms in the model and it would be equivalent to doing the same in the firms' problem.

[13] Hence forth, time subscripts are dropped.

$$P = \left(\int_{i \in \Omega} \bar{p}(i)^{1-\sigma} di \right)^{1/1-\sigma}.$$

$\bar{p}(i)$ is the consumer price of intermediate goods, which includes the price of payment services, $\bar{c}_s$, and associated the VAT, τ:

$$\bar{p}(i) = \left[(1 + a\bar{c}_s) \right] \left[(1 + \tau) p(i) \right].$$

where a is an exogenous adjustment parameter to account for the payment services intensity of the sector.[14] This incorporates the fact that some sectors may require more payment services than others (e.g. peanuts versus automobiles). Finally, solving the maximization problem renders a standard Dixit-Stiglitz demand function for each variety:

$$y(i) = \left(\frac{\bar{p}(i)}{P} \right)^{-\sigma} Y. \tag{5}$$

2.3 Banks

The representative bank produces two stream of services in a perfectly competitive environment: loans (L) for firms (in the spirit of Russ and Valderrama (2012)), and payment services (F) for purchases of intermediate goods (motivated by Lockwood (2010)). It is assumed that there are no economies of scope in the production of loans and payment services, and each stream is effectively separate.

Moreover, in the production of loans and payment services, banks use inputs from internal sources and external sources. Writing legal contacts and performing due diligence on behalf of banks, collections and liquidation services, many different types of IT services (e.g. computer systems, software packages and their maintenance, online secure payment systems etc.) are some of the services that banks may outsource or purchase from third party suppliers.[15] Making this distinction between internal and external inputs enables one to incorporate the exempt treatment of financial services under VAT that is followed by most countries. Thus, banks pay VAT on purchases of inputs from external suppliers, as financial services (e.g. loan and payment services) are exempt from VAT, banks do not charge VAT on these services and thus, cannot collect VAT credits on their inputs. This implies financial institutions effectively pay an additional cost (other than the actual cost

[14] The constant payment service intensity assumption for each differentiated good is to account for payment service intensity differences across industries (i.e. high versus low a). In a more appealing framework, $a(i)$ may be assumed to follow a distribution to capture differences in the level of payment services intensity across differentiated goods and sectors. This would create a new source of distortion between differentiated goods that is left for future research.

[15] In a simpler and more realistic world, this could be thought as internal labor versus external labor. The bank can simply either use its internal labor or hire contractors to do the work. I am currently working on incorporating this into the model.

of the inputs) on their outsourced services as a result of the exempt treatment.

Such treatment, in principle, creates an incentive to produce these services in-house, resulting in a "self-supply bias" (e.g. see McKenzie and Firth (2011), IMF (2010) and PWC (2011)). This means that the non-recovered VAT due to exempt treatment is embedded in the price of financial services, which creates several distortions in the real economy.

2.3.1 Loan Services

The bank produces loans via a perfectly complementary production process where K^S is the stock of savings supplied by consumers and M represents administrative and monitoring activities carried out by the bank in order to issue loans:

$$L = \min\left\{K^S, bM\right\}$$

Thus, each dollar loan requires one dollar of savings from consumers and b units of intermediation (administrative and monitoring) activities. Loan administration and monitoring requires internal inputs I_m and external inputs E_m, provided at prices μ_I and μ_e, and produced according to a constant return to scale Cobb-Douglas production function:

$$M = I_m^{1-z} E_m^z.$$

Minimizing the cost of administrating and monitoring loans subject to the production function gives conditional demands for E_m,

$$E_m^* = \left[\frac{z}{1-z}\frac{\mu_I}{(1+\tau)\mu_e}\right]^{1-z} M \tag{6}$$

and for I_m,

$$I_m^* = \left[\frac{z}{1-z}\frac{\mu_I}{(1+\tau)\mu_e}\right]^{-z} M. \tag{7}$$

Note that the cost of external bank inputs are $(1+\tau)\mu_e$ which reflects the fact that external inputs are subject to a VAT under exempt treatment, while this is not the case for internal inputs, I_m. Finally, one can compute the marginal cost of administrating and monitoring loans

$$c_m = \frac{{\mu_I}^{1-z}[(1+\tau)\mu_e]^z}{Z} \qquad where \qquad Z = (1-z)^{(1-z)} z^z$$

For convenience, it is further assumed that the before tax price of these inputs are the same, $\mu_e = \mu_I = \mu$, which renders the marginal cost for administrating and monitoring loans, $c_m = \frac{(1+\tau)^z \mu}{Z}$, under exempt treatment. By way of contrast, under the full recovery of the VAT on financial inputs, which is the hypothetical case of full taxation, the marginal

cost of providing these services would be $c_m = \frac{\mu}{Z}$. More generally, one can write the intermediate case with partial recovery as

$$\bar{c}_m = \frac{\mu}{Z}\left[\rho + (1-\rho)(1+\tau)^z\right], \tag{8}$$

where ρ is a recovery rate parameter takes values between 0 and 1 (i.e. $\rho = 0$ is exempt treatment and $\rho = 1$ is full taxation).

All the borrowers are subject to a uniform exogenous default rate, δ. Defaulting firms exit the market and return the loan principle (i.e., the loans are fully collateralized) but are unable to pay the loan interest, r_l. Nevertheless, the bank is still liable to its depositors (r) and have to cover its marginal cost of intermediation, $\bar{c}_m$. The representative bank makes zero expected profits given the arbitrage condition for each \$1 loan at the equilibrium:

$$(1-\delta)(r_l - r - \frac{\bar{c}_m}{b}) = \delta(r + \frac{\bar{c}_m}{b}).$$

The loan interest rate is therefore:

$$r_l = \frac{r}{(1-\delta)} + \frac{\bar{c}_m}{b(1-\delta)}. \tag{9}$$

The equilibrium loan rate is the sum of the deposit interest rate, plus the marginal cost of intermediating a \$1 loan, all adjusted for default risk. The latter is referred as the "spread", which is the difference between the risk adjusted interest rate paid on bank deposits and the loan rate. Importantly for our purposes, due to exempt treatment the spread reflects the VAT (embedded in $\bar{c}_m$; see (8)), which means that the unrecovered tax is indirectly reflected in the cost of borrowing to firms. As a result, the cost of borrowing is higher under exempt treatment than it is under full taxation. This contradicts the basic principle of VAT, as the effective VAT rate for business to business (B2B) transactions should be zero.

2.3.2 Payment Services

Payment services (e.g. credit and debit cards, check clearing, online payment services, etc.) are produced by the bank for purchases of differentiated goods. As before, internal inputs (I_s) and external inputs (E_s) are used in the production of these services (F) via

$$F = I_s^{1-s} E_s^s.$$

As above, conditional input demands are:

$$E_s^* = \left[\frac{s}{1-s}\frac{\mu_I}{(1+\tau)\mu_E}\right]^{1-s} F \tag{10}$$

and

$$I_s^* = \left[\frac{s}{1-s}\frac{\mu_I}{(1+\tau)\mu_E}\right]^{-s} F. \tag{11}$$

Thus the marginal cost of payment services is:

$$c = \frac{{\mu_I}^{1-s}[(1+\tau)\mu_E]^s}{S} \qquad where \qquad S = (1-s)^{(1-s)}s^s$$

Incorporating the earlier assumption about the before tax price of these inputs, $\mu_E = \mu_I = \mu$ gives $c = \frac{(1+\tau)^s\mu}{S}$. As before, one can incorporate partial recovery of VAT for payment services, which then gives the weighted marginal cost of $c_s = \frac{\mu}{S}\left[\rho + (1-\rho)(1+\tau)^s\right]$. As the recovered payment services are charged full VAT but not the unrecovered financial services, the consumer price of payment services with partial recovery is:

$$\bar{c}_s = \frac{\mu}{S}\left[\rho(1+\tau) + (1-\rho)(1+\tau)^s\right]. \tag{12}$$

As payment services are solely used in the purchase of differentiated goods, they are classified as business to consumers transactions (B2C). The result is that the price of payment services under exempt treatment is lower than would it be under full taxation, $\bar{c}_s^{Exempt} = \frac{(1+\tau)^s\mu}{S} < \bar{c}_s^{full} = \frac{(1+\tau)\mu}{S}$ (since $s < 1$). This implies that B2C transactions are under-taxed with respect to VAT under exempt treatment, which creates distortions for the real economy.

2.4 Intermediate Goods Sector

Firms in this sector operate in a monopolistically competitive environment and are heterogenous with respect to productivity as in Melitz (2003). Each firm employs labor $n(i)$ and borrows capital from the bank $k(i)$ to produce intermediate consumption goods according to a crs production function:

$$y(i) = \varphi(i)n(i)^{\alpha}k(i)^{1-\alpha},$$

where $\varphi(i)$ is the productivity parameter for firm i. All the firms pay a fixed cost of f for loans.[16] The fixed cost of loans is modelled as a monetary cost though it could also be thought as non-monetary cost of the nature of reputation building, credit background, references etc., for loan applications. It is also measured in terms of final good, P, and borrowed from the bank.

As the first step, each firm minimizes its cost of production, which gives conditional labor and capital demand functions:

[16]The loan fixed cost is paid by borrowers every period to obtain loans from the bank.

$$n^*(i) = \left[\frac{\alpha}{1-\alpha}\frac{r_l}{w}\right]^{1-\alpha}\frac{y(i)}{\varphi(i)} \tag{13}$$

and

$$k^*(i) = \left[\frac{\alpha}{1-\alpha}\frac{r_l}{w}\right]^{-\alpha}\frac{y(i)}{\varphi(i)}. \tag{14}$$

These give rise to the marginal cost function $\frac{W}{\varphi(i)}$

$$W = \frac{w^{\alpha}r_l^{1-\alpha}}{B} \qquad where \qquad B = (1-\alpha)^{(1-\alpha)}\alpha^{\alpha}.$$

Each firm then chooses its price to maximize profits,

$$\max_{p(i)}\left[p(i)y(i) - \frac{W}{\varphi(i)}y(i) - (1+r_l)Pf\right]$$

subject to the demand for its product, (5). This generates a profit maximizing producer price for firm i which is a mark-up over its marginal cost:

$$p(i) = \left[\frac{\sigma}{\sigma-1}\right]\frac{W}{\varphi(i)}. \tag{15}$$

As in Melitz (2003), more productive firms charge lower prices since their marginal cost is lower Note that because the marginal cost of production is function of the loan rate r_l, (8), which in turn is a function of the VAT levied on the bank inputs due to exemption treatment, it is evident that both the cost of capital and the price of consumption goods are in turn a function of the VAT. Firm profits are:[17]

$$\pi(\varphi) = \left[\frac{(\sigma-1)^{(\sigma-1)}}{\sigma^{\sigma}}\left(\frac{\varphi}{W}\right)^{(\sigma-1)}\frac{P^{\sigma}Y}{[(1+a\bar{c}_s)(1+\tau)]^{\sigma}} - (1+r_l)Pf\right].$$

Again following Melitz (2003), the marginal firm, which makes zero profits, will be indifferent between producing or not. This allows us to determine the cut-off productivity parameter, φ^*, for the marginal firm, by setting (2.4) equal to zero. That is:

$$\varphi^* = \left[\frac{\sigma(1+r_l)f}{Y}\right]^{\frac{1}{(\sigma-1)}}\frac{\sigma}{(\sigma-1)}\frac{W}{P}[(1+a\bar{c}_s)(1+\tau)]^{\frac{\sigma}{(\sigma-1)}}. \tag{16}$$

This equation gives rise to several important insights. In particular, it is evident that the VAT will have several general equilibrium effects on the cut-off, since P, r_l and $\bar{c}_s$ are all functions of t. As such, and importantly, the VAT levied on the inputs of banks will lead to movements along both the intensive and extensive margins, which is not appreciated in the literature.

[17]We drop the i subscript from now on and use φ to reference firms in the goods producing sector.

3 Aggregation

We assume that the heterogenous productivity parameter follows a cumulative distribution $H(\varphi)$. Denote the mass of firms that exists at the equilibrium in the market as g. The price index for the composite good can then be written as:

$$P = \left(\int_{\varphi^*}^{\infty} [\bar{p}(\varphi)]^{1-\sigma} g \nu(\varphi) d\varphi \right)^{1/1-\sigma},$$

where $\nu(\varphi)$ is the conditional distribution of firms productivity, that is defined as

$$\nu(\varphi) = \begin{cases} \frac{h(\varphi)}{[1-H(\varphi*)]} & \text{if } \varphi > \varphi^*, \\ 0 & \text{Otherwise.} \end{cases}$$

Rearranging terms:

$$P = [(1 + a\bar{c}_s)(1+\tau)] g^{1/1-\sigma} \left(\frac{1}{[1-H(\varphi^*)]} \int_{\varphi^*}^{\infty} p(\varphi)^{1-\sigma} h(\varphi) d\varphi \right)^{1/1-\sigma},$$

and substituting from (15), leads to:

$$P = [(1 + a\bar{c}_s)(1+\tau)] g^{1/1-\sigma} \frac{\sigma}{(\sigma-1)} \frac{W}{\left(\frac{1}{[1-H(\varphi^*)]} \int_{\varphi^*}^{\infty} \varphi^{\sigma-1} h(\varphi) d\varphi \right)^{1/\sigma-1}},$$

and finally:

$$P = [(1 + a\bar{c}_s)(1+\tau)] g^{1/1-\sigma} \frac{\sigma}{(\sigma-1)} \frac{W}{\bar{\varphi}(\varphi^*)} \tag{17}$$

where,

$$\bar{\varphi}(\varphi^*) = \left(\frac{1}{[1-H(\varphi^*)]} \int_{\varphi^*}^{\infty} \varphi^{\sigma-1} h(\varphi) d\varphi \right)^{1/\sigma-1} \tag{18}$$

is the average productivity prevailing in the sector.

In what follows we assume that the productivity parameter is drawn from a Pareto distribution. That is $H(\varphi)=1-\varphi^{-\theta}$. This allows us to obtain a closed form solution for the average productivity as a function of the cuff-off threshold:[18]

$$\bar{\varphi}(\varphi^*) = \left(\frac{\theta}{\theta - (\sigma-1)} \right)^{1/\sigma-1} \varphi^*. \tag{19}$$

[18] In order for $\bar{\varphi}$ to be finite and positive, we assume that $(\sigma-1) < \gamma$.

Substituting this back in to (16) gives

$$\varphi^* = \left[\frac{\sigma(1+r_l)f}{Y}\right]^{\frac{1}{(\sigma-1)}} g^{\frac{1}{(\sigma-1)}}[(1+a\bar{c}_s)(1+\tau)]^{\frac{1}{(\sigma-1)}}\bar{\varphi} \tag{20}$$

as the productivity cut-off. It is important to note here that the average productivity $\bar{\varphi}$ is higher than the zero profit cut-off φ^*. This implies that the firm with the average productivity makes positive profits.

We then follow Melitz (2003) by assuming that the distribution of entering firms is exactly same as the distribution of exiting firms (due to the death shock). This ensures that the average productivity in the industry is preserved. It also ensures that average profits, input demands, price and output decisions of the average firm will be unaffected by the entry and exit of other firms, which is the principle property of Melitz's steady state equilibrium.

4 Market Clearing

We are now in a position to close the model by specifying the market clearing conditions. In what follows, we discuss this for each market separately.

4.1 Loan (Capital) Market

The demand for the average firm's (with average productivity) output, after adjusting for the aggregate price index, is

$$y(\bar{\varphi}) = \frac{Y}{g^{\frac{\sigma}{\sigma-1}}}. \tag{21}$$

Substituting (21) into the conditional capital demand function (14) gives the loan demand for the average firm in the industry:

$$k^*(\bar{\varphi}) = \left[\frac{\alpha}{1-\alpha}\frac{r_l}{w}\right]^{-\alpha}\frac{Y}{\bar{\varphi}g^{\frac{\sigma}{\sigma-1}}}$$

Thus, aggregate loan (capital) demand including the fixed cost of loan finance is:

$$L^* = K^D = g(k^*(\bar{\varphi}) + f) = (1-\alpha)\frac{(\sigma-1)}{\sigma}\frac{PY}{r_l[(1+a\bar{c}_s)(1+\tau)]} + gf \tag{22}$$

Since the exiting firms return the loan principle but not the interest on it, we do not subtract the capital of exiting firms from the total capital demand. As mentioned above, the firms also borrow the fixed cost of loans, which is same across all firms. Capital market equilibrium requires that the capital demanded by firms equal the capital (savings) supplied by the banks, $K^D = L^S$ given that

$$r_l = \frac{r}{(1-\delta)} + \frac{\bar{c}_m}{b(1-\delta)} \qquad where \qquad r = \frac{(1-t_k\kappa)(1-\beta+\beta\phi)}{\beta(1-t_k)}$$

at the steady state equilibrium.

4.1.1 Marginal Effective Tax Rate on Capital

The marginal effective tax rate (METR) summarizes the tax system in a hypothetical rate applied to the marginal unit of capital. It is expressed as the percentage difference between a neutral tax system with respect to capital (meaning zero effective rate on capital) and the existing system. For our purpose, one can define a neutral tax system with respect to capital under the case where κ is equal to 1 (this is a pure cash flow tax), which gives $r_n = \frac{(1-\beta+\beta\phi)}{\beta}$. Recalling that r_l is the before tax cost of capital adjusted for default risk, the METR can be written as

$$METR = \frac{r_l - r_n}{r_l}. \tag{23}$$

Any increase in r_l due to the tax system will increase the METR. This can occur by an increase in t_k, a decrease in κ or an increase in the VAT rate (and therefore $\bar{c}_m$) under exempt treatment.

4.2 Labour Market

As before, substitute (21) in to the conditional labor input demand function (13) to obtain labor demand for the average firm in the industry:

$$n^*(\bar{\varphi}) = \left[\frac{\alpha}{1-\alpha}\frac{r_l}{w}\right]^{1-\alpha} \frac{Y}{\bar{\varphi} g^{\frac{\sigma}{\sigma-1}}}.$$

Aggregate labor demand is therefore:

$$N^* = N^D = (1-\delta)gn^*(\bar{\varphi}) = \alpha(1-\delta)\frac{(\sigma-1)}{\sigma}\frac{PY}{[(1+a\bar{c}_s)(1+\tau)]w}. \tag{24}$$

Considering the labor supply, $N_t^s = \left(\frac{(1-t_w)w_t}{P_t}\right)^{\frac{1}{\psi-1}}$, one can compute the equilibrium wage rate,

$$w^* = [\alpha(1-\delta)Y]^{\frac{1}{1-\alpha}-\frac{1}{\psi(1-\alpha)}} \frac{g^{\frac{1}{(1-\sigma)(1-\alpha)}}}{\alpha^{\frac{\alpha}{1-\alpha}}} \frac{r_l}{1-\alpha} \left(\frac{1}{\bar{\varphi}}\right)^{\frac{1}{(1-\alpha)}} \left(\frac{\sigma}{\sigma-1}\right)^{\frac{1}{\psi(1-\alpha)}} \chi \left[\frac{(1+a\bar{c}_s)(1+\tau)}{(1-t_w)}\right]^{\frac{1}{\psi(1-\alpha)}}. \tag{25}$$

One may wonder how this result would change in the case of an inelastic labor supply. To see this let $\psi \to \infty$, which implies $N^s \to 1$ and w^* is

$$w^* = \left[\frac{\alpha(1-\delta)Y}{\bar{\varphi}}\right]^{\frac{1}{1-\alpha}} \frac{g^{\frac{1}{(1-\sigma)(1-\alpha)}}}{\alpha^{\frac{\alpha}{1-\alpha}}} \frac{r_l}{1-\alpha}. \tag{26}$$

4.3 Input Sector

As indicated earlier, inputs are used in the production of bank intermediation services: administrating and monitoring loans and payment services. Although the supply of these inputs is exogenous, in order to close the model, it is further assumed that all the revenue generated in this sector is collected by consumers as they are the owners of all firms in the model.[19]

Starting with loan production, the equilibrium level of loans, $L^* = K^D$, requires the bank to supply

$$M^* = \frac{K^D}{b}$$

in administrative and monitoring services. This will cost the bank a total of Ex_m^*:

$$Ex_m^* = \frac{\mu}{Z} M^* \left[\rho + (1-\rho)(1+\tau)^z\right], \tag{27}$$

which includes the unrecovered VAT (embedded in the cost) due to exempt treatment of financial services. The total tax revenue collected from inputs that are used in exempt loan services, TR_m^* is

$$TR_m^* = (1-\rho)\tau\mu \left[\frac{z}{1-z}\frac{1}{(1+\tau)}\right]^{1-z} M^*. \tag{28}$$

Similarly, the total demand for payment services is

$$F^* = \left[\frac{a(1-\delta)PY}{(1+a\bar{c}_s)}\right],$$

which is simply the after tax value of aggregate revenues of intermediate goods adjusted by the payment intensity parameter, a, and the default rate. It is important to note that this is not the "value" but the "amount" of demand for payment services that needs to be multiplied by the price of payment services per \$1 in transactions. One then can compute the total expenditure spent on inputs that are used in the production of payment services, Ex_s^*:

$$Ex_s^* = \frac{\mu}{S} F^* \left[\rho + (1-\rho)(1+\tau)^s\right], \tag{29}$$

[19] An alternative approach would be to assume these inputs to be labor, such as inside labor versus outside labor (or contractors). However this approach complicates the labor market equilibrium.

which also includes the unrecovered VAT. Additionally, as payment services are provided to final consumers, the bank will collect VAT from the portion (ρ) of financial services that are not exempt from VAT.[20] We then compute the total weighted tax revenue from this sector, that is TR_s^*:

$$TR_s^* = \tau \frac{\mu}{S} F^* \left[\rho + (1-\rho)(1+\tau)^s \frac{s}{(1+\tau)} \right]. \tag{30}$$

4.4 Value of the Average Firm

The value of the average firm is the discounted value of its expected profits:

$$\bar{v} = \sum_{t=0}^{\infty} [\beta(1-\delta)]^t \bar{\pi}.$$

Using the Euler equation (9), and setting $C_t = C_{t+1}$ in the steady state, one can obtain the discounted value of a firm

$$\bar{v} = \frac{\bar{\pi}}{1-\beta(1-\delta)}.$$

As in Melitz (2003), entering firms pay an entrance fee f_e (measured in final consumption) after which they draw their ex-post productivity parameter φ. Upon drawing their φ, they know whether they can make non-negative profits in the market or not. Firms that make non-negative expected profits ($[1-H(\varphi^*)]$) enter the market and others leave in the first stage. We therefore can write the value of an average firm as $\tilde{v} = [1-H(\varphi^*)]\bar{v} - Pf_e$ where f_e. Free entry continues until the equilibrium is reached in the market, which requires $\tilde{v} = 0$ and therefore profits for the average firm are:

$$\bar{\pi} = \frac{1-\beta(1-\delta)}{[1-H(\varphi^*)]} P f_e. \tag{31}$$

Entry fees are paid by owners of the firms and eventually returned back to consumers. As consumers pay and receive the entry fees every period, the net contribution of these fees to their budget is zero.

4.5 Stability of the Equilibrium

A sustainable equilibrium (i.e., to avoid unstable expansions) requires that the size of the economy remains constant in the steady state. Thus, the number of new successful entrants must equal the number of exiting firms, $[1-H(\bar{\varphi})]g_e = \delta g$. This can be used to determine the aggregate entry fees paid in equilibrium as:

[20] This is not the case for the loan market. As the borrowers are businesses not final consumers in the loan market, they either pay embedded VAT or no VAT, see (28).

$$g_e P f_e = \frac{\delta g}{[1 - H(\varphi^*)]} P f_e.$$

4.6 Resource Constraint

The total value of production (including tax) that is used for consumption and investment at the steady state is:

$$PY = PC + P(1 - t_k\kappa)I = PC + P(1 - t_k\kappa)\phi K,$$

where the net investment is $P\phi K$.[21] The equilibrium level of consumption is

$$C^* = Y^* - (1 - t_k\kappa)\phi K^*,$$

which can also be used to define welfare along with labor:

$$Welfare = \left[C^* - \frac{N^{*\psi}}{\psi}\right].$$

Moreover, the resource constrain at the equilibrium is

$$PY = (1 - t_w)w^* N^* + rP(1 - t_k)K^* + \Pi + Ex + TR. \tag{32}$$

Total profits are the sum of profits from the financial industry and from intermediate goods producing industry, $\Pi = \Pi^b + \Pi^c$,

$$\Pi^b = g(1 - \delta)(1 + r_l)Pf + g_e P f_e$$

and

$$\Pi^c = g(1 - \delta)\left[\frac{PY}{g[(1 + a\bar{c}_s)(1 + \tau)]\sigma} - (1 + r_l)Pf\right].$$

Finally, the net profits are[22]

$$\Pi = (1 - \delta)\frac{PY}{\sigma[(1 + a\bar{c}_s)(1 + \tau)]}. \tag{34}$$

Total input expenditures (excluding tax payments) spent by the bank for loan and payment services are $Ex = Ex^*_m + Ex^*_s$. Both constituents of input expenditures are as

[21] Investment is defined as $I_t = K_t + 1 + (1 - \phi)K_t$ and at the steady state, it will be ϕK.

[22] In general, the net profits would be the difference between firms' profit and entry costs payments:

$$\Pi = (1 - \delta)\frac{PY}{\sigma[(1 + a\bar{c}_s)(1 + \tau)]} - g_e P f_e. \tag{33}$$

However, both firms' profits and entry costs would eventually go to consumers as nothing disappears at the equilibrium in the model.

discussed in the previous section. In addition to this, the total tax revenue (TR) generated in the model comes from the banking sector ($TR_m^* + TR_s^*$), from the intermediate goods sector, capital and payroll taxes:

$$TR = TR_m^* + TR_s^* + \tau \frac{(1-\delta)PY}{[(1+a\bar{c}_s)(1+\tau)]} + t_k(r-\kappa\phi)PK + t_w w N^*. \tag{35}$$

4.7 Number of Firms

Total expenditures (excluding explicit VAT) is equal to the average revenue times the total number of firms at the equilibrium:[23]

$$\frac{PY}{[(1+a\bar{c}_s)(1+\tau)]} = gr(\bar{\varphi}), \tag{36}$$

given that the average revenue can be written as

$$r(\bar{\varphi}) = \sigma\left[\bar{\pi} + (1+r_l)Pf\right]. \tag{37}$$

Plugging (31) and (37) into (36) and considering the distributional assumption we made earlier about the productivity gives the number of firms in equilibrium:

$$g = \frac{Y}{\sigma\left[1-\beta(1-\delta)\right]\varphi *^{\theta} f_e + (1+r_l)f]\left[(1+ac_s)(1+\tau)\right]}. \tag{38}$$

5 Discussion of Equilibrium

To begin, define the ratio of revenues for the marginal and the average firms with productivities φ^* and $\bar{\varphi}$:

$$\frac{r(\bar{\varphi})}{r(\varphi^*)} = \frac{\bar{\varphi}^{(\sigma-1)}}{\varphi^{*(\sigma-1)}}.$$

Plug this in to the profit function for the average firm in the place of $r(\bar{\varphi})$, giving

$$\bar{\pi} = \frac{\bar{\varphi}^{(\sigma-1)}}{\varphi^{*(\sigma-1)}} \frac{r(\varphi^*)}{\sigma} - (1+r_l)Pf.$$

Given that $r(\varphi^*) = \sigma(1+r_l)Pf$ from $\pi(\varphi^*) = 0$, following Melitz the "Zero Cutoff Profit (ZCP)" is

$$\bar{\pi} = \left[\frac{\bar{\varphi}^{(\sigma-1)}}{\varphi^{*(\sigma-1)}} - 1\right](1+r_l)Pf.$$

[23] One could recover the number of the firms at the equilibrium from the identity for firms profit, $g = \frac{\Pi}{\bar{\pi}}$, which eventually gives exactly the same equation for g as below.

As a result of the distributional assumption, the equation for $\bar{\varphi}$ (19) indicates a linear relationship between $\bar{\varphi}$ and φ^*. Thus the ZCP becomes:

$$\bar{\pi} = \left[\frac{\sigma - 1}{\theta - (\sigma - 1)}\right](1 + r_l)Pf,$$

that is not directly affected by φ^* but indirectly, through P. Secondly, the "Free Entry Condition (FE)" is

$$\bar{\pi} = \frac{1 - \beta(1 - \delta)}{[1 - H(\varphi^*)]} P f_e.$$

FE is a direct and indirect function of φ^*, which implies that an increase in φ^* will lead to a shift as well as a movement along the FE curve.

Finally, using these two equations and the distributional assumption for φ (and thus the solution for $\bar{\varphi}$), one can solve for the unique[24] equilibrium that will render the equation for the productivity cut-off φ^*:

$$\varphi* = \left[\frac{(1 + r_l)f}{[1 - \beta(1 - \delta)]\, f_e}\left(\frac{\sigma - 1}{\theta - \sigma + 1}\right)\right]^{\frac{1}{\theta}}. \tag{39}$$

The stationary equilibrium is defined by φ^* and $\bar{\pi}$ and thus, one can recover the value of all other aggregates at the equilibrium.

Repealing exempt treatment and moving to full taxation (while holding tax revenue constant) decreases the cost of borrowing that affects φ^*, r_l as well as P, which leads to

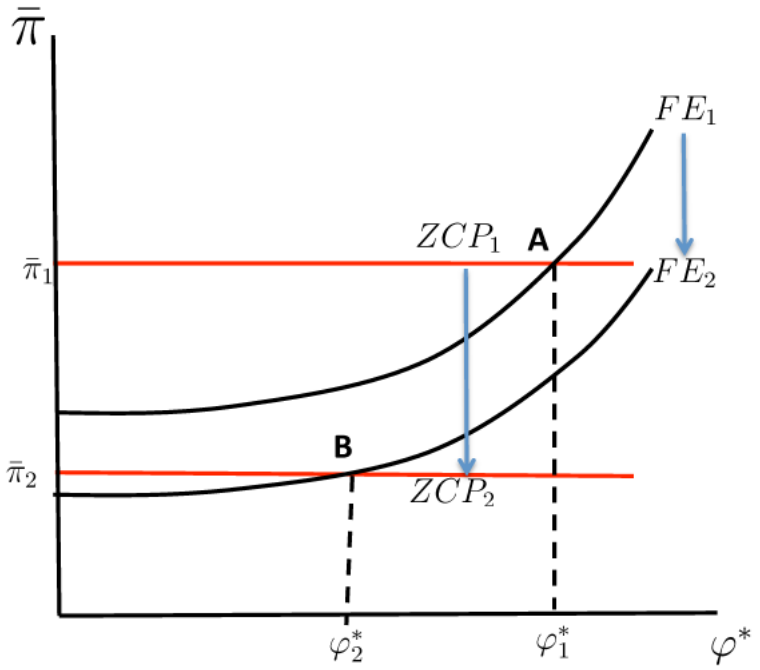

Figure 1: Determination of the equilibrium cutoff φ^* and average profit $\bar{\pi}$

[24]For a formal proof of the existence and uniqueness of the stationary equilibrium, see Melitz (2003).

a downward shift in both curves ZCP and FC. This is illustrated in Figure 1, where the initial equilibrium is assumed to be established at point A, φ_1^* and $\bar{\pi}_1$. As the cost of borrowing is lower now, r_l and P are lower, which then shifts φ^* down. This allows the entry of new firms that drives up the number of the firms in the market. The average productivity is lower at the new equilibrium, however due to the increase in the number of the firms, Y is now higher at a lower P. The new equilibrium is established at B, ZPC_2 and $\bar{\pi}_2$.[25] Many other general equilibrium effects are discussed below.

6 Simulations

The simulations are based on solving three main equations, (20), (32) and (38), simultaneously that define the equilibrium given all other equations defining the endogenous variables, main aggregates and parameters.

Exact figures for most of the parameters are unfortunately not not available due to data limitations. However, "reasonable" approximations can be employed with a direct consultation of the related finance literature. Particularly, Russ and Valderrama (2012) provide reasonable estimates for some of the bank level parameters and confirm them with the literature. Following their work, the fixed loan application costs (f) is assumed to be 0.1 and the loan default rate (δ) is set equal to 5%. Motivated from the estimates provided by Feenstra (1994), σ is set to 8 that is also assumed to be equal to θ, which is in line with Helpman et al. (2004). Other parameters are standard in the literature: labour share in production, α, is 0.65, elasticity of labour supply, ψ, is 2 for the elastic case (and ∞ for the inelastic case), discount factor, β, is 0.96.

Policy parameters are chosen for illustrative purposes. For instance, low payment service intensity is illustrated with a equal to 25% and the high payment intensity is 75%. Exempt treatment is illustrated with zero recovery (or $\rho = 0$), partial recovery is 50% (or $\rho = 0.5$) and full taxation is implied from 100% recovery (or $\rho = 1$). The VAT rate is assumed to be low, medium and high as 5%, 10% and 15% under exempt treatment (for revenue neutral analysis). Payroll tax, t_w, capital income tax, t_k and the investment allowance, κ, are set at reasonably moderate rates as 10%, 15% and 75% in the main estimates. For METR computations, some of these values are varied.

There is no direct estimates available for the share of bank's external inputs z and s, the price index for bank inputs μ, the amount of demand for administrating and monitoring as proportion of loans, b and fixed cost of entry, f_e. As such reasonable values are chosen for these parameters, which also allows simulations to converge to an

[25]The figure also shows that if the magnitude of a shift in ZCP is lower than the shift in FE, the productivity threshold would be higher than the initial case. Such a case does not happen at the equilibrium of this model. This is true by the setup given the technical formulation of the equations (ZCP and FE) ensures that this never happens.

equilibrium. These are 0.30 for s and z, 0.1 for μ, 50 for b[26] and finally, 10 for f_e.

Results are robust to a relatively wide range of these parameters and initial guesses. $fsolve$ procedure is used in MATLAB for the computation of the equilibrium and the convergence rule (for the convergence procedure) is further narrowed as a robustness check of these results.

7 Results

Results are summarized in two parts. The first part discusses the simulation results from the policy exercise: moving from exempt treatment ($\rho = 0\%$) to full taxation ($\rho = 100\%$), holding tax revenue constant. The exercise is repeated with partial recovery ($\rho = 50\%$) to illustrate intermediary cases between the two extremes, exempt treatment and full taxation. In both exercises, the reference case is exempt treatment for a given tax revenue and payment service intensity parameter. Through out the analysis, t_w and t_k are set to some reasonable rates, as 10% and 15%, and κ is assumed to be 0.75. As a simple sensitivity check, each scenario is simulated under low and high payment service intensities (e.g. $a_{low} = 25\%$ and $a_{high} = 75\%$), as well as for low, medium and high tax revenues. In a separate consideration, the model is simulated for different VAT rates under the each tax regime, holding everything else constant, in an attempt to investigate the relation between tax revenues and level of VAT rates (i.e. laffer curve). All of these results are shown in tables 1 to 4 and in Figure 1.

The second part focuses on how policy changes affect the cost of borrowing captured by the marginal effective tax rate (METR). METRs are computed under different policies and the % change is obtained for three policy dimensions: (1) shifting the regime from exempt treatment to full taxation ($\%\Delta\ METR_1$) holding the VAT rate and t_k constant, (2) changing the VAT rate under the same regime and t_k ($\%\Delta\ METR_2$) and (3) changing t_k under the same regime and VAT rate ($\%\Delta\ METR_3$), holding everything else constant. All other exogenous parameters are kept as before, while conducting the exercise. Results are summarized in Table 5.

In order to provide some policy context, all the results are summarized in the context of Ontario's recent tax harmonization policy, which was effectively increasing the VAT rate under exempt treatment. All the results should be interpreted as a qualitative analysis of the equilibrium effects of the tax policies, rather than quantitative (measuring the magnitude of the effect) since a formal calibration is not feasible due to data limitations at the moment. Nevertheless, one can still use these the numerical results to rank different tax policies (in an ordinal sense) in terms of the size of the GE effects.

[26]This number implies that the demand for administrating and monitoring a $100 loan is $2.

7.1 Main Aggregates

Overall results show that a policy shift from exempt treatment to full taxation, holding tax revenue constant, leads to a decrease in the loan interest rate (Table 1). While the decrease in the loan interest is considerably small, the number of firms at the equilibrium rises quite a lot given that the productivity threshold is now lower. This expands total output and consumption. All of these changes are found to be greater in magnitude when the exercise is repeated at higher levels of tax revenue. For instance, the decrease in the loan interest rate is as big as 0.1% point, when moving from exempt treatment to full taxation under the highest tax revenue. Results obtained from the industry with low payment service intensity ($a_{low} = 25\%$) are qualitatively very similar to the ones computed under a high payment service intensity ($a_{high} = 75\%$). It is important to note that the magnitude of the change in loan interest does not vary depending on payment service intensity, as it does not directly interact with the loan market. However, the impact of a change in loan interest rate on main real aggregates due a shift in the tax policy is greater in the case with higher payment service intensity. Repeating the exercise for intermediary cases produces qualitatively similar results. Yet, the magnitude of the changes is relatively smaller.

We now turn our attention to changes in the main prices of the model, Table 2. As mentioned before, following a downward shift in the productivity threshold, new firms enter the market. Entering firms increase the demand for labor, which drives the wage rate up although the increase is not high enough to stop aggregate price from decreasing. In other words, aggregate price decreases due to increase in competition. Finally, the equilibrium levels of labour and capital are both higher under full taxation. As before the magnitude of changes are greater under a policy shift from exempt treatment to full taxation compared to the same shift from exempt treatment to partial recovery. And that magnitude is also larger with greater payment service intensity.

Results from revenue neutral welfare comparisons are presented in Table 3. They show that full taxation generates the highest level of welfare for any level of tax revenue and payment services intensity parameter. For instance, moving from exempt treatment to full taxation (in Table 3), holding revenue constant, welfare increases by .07% and this number is as high as 2% at the highest tax revenue under low payment intensity. Repeating the same exercise for high payment intensity at high tax revenue renders about a 7.4% increase in welfare. This number is 4.9% for a similar analysis of a policy shift from exempt treatment to partial recovery.

Tax revenue deserves a separate discussion as it has been at the centre of current policy discussions with regard to taxation of financial services under VAT. Firstly, I obtained simulation results from welfare natural comparisons of tax revenue.[27] According to these

[27]These results are not reported here as they are qualitatively similar to the ones reported but they

results, full taxation is revenue enhancing compared to exempt treatment for a given level of welfare. Secondly, an analysis of the source of government revenue (using the revenue neutral simulation results) is also performed in Table 4. Start with the low payment intensity case and low tax revenue (i.e. 5% under exempt treatment). Payroll tax generates the biggest portion of the total tax revenues in the model. VAT revenues (from both payment services and consumption) picks up as the VAT rate increases and eventually becomes the main source of revenue. As B2C transactions (i.e. payment services) are under-taxed under exempt treatment and partial recovery cases, revenue collection from payment services produced by the bank is quite low compared to full taxation where B2C transactions are fully taxed. Similarly, under exempt treatment and partial recovery policies, B2B transactions (i.e. loan services) are over-taxed. This generates a positive VAT revenue from loan services despite the fact that B2B transactions should not be taxed according to the main principles of VAT. These revenues are zero under full taxation.

VAT rate leverage across different policies is also investigated. It is evident from the results that total tax revenue does not always increase as the VAT rate increases (e.g. referred as VAT rate leverage), holding everything else constant. In fact, the revenue curve is concave in VAT rate, Figure 1; this is the well known as Laffer Curve in the literature. The curve displays revenue generation by different VAT rates (holding everything else constant) under full taxation with a low payment intensity industry. It shows that tax revenues from payroll and capital income taxes decrease as VAT rate increases. Tax revenue from payment services and purchases of final goods (for consumption) increases at a decreasing rate as VAT rate increases and eventually, starts decreasing. Total tax revenue increases up to 17% VAT rate and reflects to a decreasing trend after this point. It is important to note that for a low payment service intensity case this results does not change under exempt treatment. However, for high payment intensity case (e.g. $a_{high} = 75\%$), exempt treatment records a higher VAT rate for the reflection point than full taxation. In other words, VAT rate leverage under exempt treatment is higher than it is under full taxation. This is mainly due to the fact, B2C transactions are taxed effectively at a higher rate under full taxation and the tax base is larger for high payment intensity case.

7.2 Marginal Effective Tax Rates

As pointed earlier, the loan interest rate is higher under exempt treatment due to embedded VAT and the magnitude of this difference in the rate between exempt treatment and full taxation is greater for high VAT rates. This is well outlined in Table 5, which presents loan interest rates and METRs computed under 5% and 25% VAT rates for

are available upon request from the author.

different values of t_k. Overall, METRs are greater for high VAT rates under exempt treatment, which is not the case under full taxation. This is again due to the fact that VAT is not embedded in the loan interest rate under full taxation, as B2B transaction are taxed at zero rate. Finally, METRs increase in t_k for a given VAT rate under the same tax regime.

More specifically, $\%\Delta\ METR_1$ presents percentage change in METR due to a shift in tax regime from exempt treatment to full taxation, holding everything else constant. METRs are lower under full taxation. More importantly, the magnitude of the percentage change in METR is relatively greater under high VAT rates and it reduces as t_k increases. $\%\Delta\ METR_2$ displays percentage change in METR in response to a change in VAT rate under the same tax regime and t_k. Results show that METRs are greater for high VAT rates under exempt treatment that is not the case under full taxation. Finally, $\%\Delta\ METR_3$ shows that METRs increases in t_k for a given VAT rate under the same tax regime.

These results are qualitatively robust to considering an intermediary case (e.g, partial recovery with $\rho = 50\%$), however, the magnitudes of the changes are slightly smaller. Secondly, different values of the investment allowance, κ, (i.e. 25% and 50%) are considered. Results with respect to $\%\Delta\ METR_1$ and $\%\Delta\ METR_2$ are qualitatively the same except that the magnitude of the change is slightly smaller under relatively lower values of κ. However, the results with respect to $\%\Delta\ METR_3$ are different. That results indicate that % increase in METR as t_k raises is greater under relatively lower values of κ for a given level of VAT under the same tax regime. Lower κ implies a deduction scheme over a long period that reduces the present value of a dollar tax deduction allowance for capital investment, which acts as an indirect tax on capital. This simply increases the spread between the hurdle rate of interest and the loan interest rate that implies a greater METR.

7.3 Case Study: PST versus HST in Ontario

The federal Goods and Services Tax (GST), 5%, and the Provincial Sales Tax (PST), 8%, were combined into one consumption tax that is Harmonized Sales Tax (HST), 13% effective by July 1, 2010 in the province of Ontario, Canada. Prior to harmonization most financial transactions were subject to GST but not PST. Under the legislation most financial services are exempt from HST, where financial institutions and insurance companies cannot charge HST on most of their outputs, nor do they receive input credits for tax paid on their inputs. The policy effectively has increased the VAT rate on financial transactions in Ontario from 5% to 13% under exempt treatment. As discussed above, this has several direct and indirect implication for the real economy.

First, it would be expected to increase the loan interest rate by about 0.01 percentage

points according to the simulations presented above. That translates to an increase in the effective cost of capital (METR) by 0.23%.[28] This increase in the cost of capital would not occur under full taxation. Although it is hard to make claims about the magnitude of the impact without proper calibration, based on the work presented here, the effect is expected to be strong enough to cause a decrease in overall welfare, investment and tax revenue.

Overall, based on the previous findings one can claim that HST reform in Ontario leads to a rise in the effective cost of producing financial services that are exempt from VAT. And this increase is reflected in the entire economy through financial services (e.g. loan intermediation and payment services). This shifts the productivity threshold up and forces small firms to exit (and many others not to even enter) the market. This effect is expected to be more severe in capital intensive industries and also in industries that intensively rely on payment services for sales of their products.

One may argue that most of these effects may be mitigated in real life by opening the economy to international borrowing. However, while this might be true for loan services it is not entirely true for the payment services. Moreover, international borrowing would still not help all the financial institutions, such as credit unions. Credit unions which mostly serve and only accept deposits from local consumers and businesses (mostly micro and small businesses), may still suffer from the negative consequences of HST policy in Ontario.

Lastly, the simulation results also suggest that the magnitude of the negative consequences of the HST policy in Ontario can be mitigated by increasing the recovery rate. It is therefore important to repeat here that the higher the recovery rate, the smaller the distortion due to exempt treatment will be.

8 Conclusions

The treatment of financial institutions under the VAT is a very complex issue in public finance. For various technical and practical reasons, financial services are generally exempt from VAT in most countries. This creates distortions in the real economy. While some related literature has discussed the importance of these distortions, the precise transmission mechanisms of these distortions and their net impact on the real economy and welfare have not been explored in a general equilibrium setting.

This paper presents a dynamic general equilibrium framework with financial institutions and government that can be used to identify these mechanisms and evaluate many of the aforementioned distortions. The model accommodates both extensive and intensive margin movements. The model is simulated at the steady state equilibrium under

[28] In the METR computation, t_k is taken 25% and κ is set equal to 50% and the rest is same as before. VAT rate is shifted from 5% to 13% and the percentage change in METR is computed.

reasonable parameter assumptions and a numerical comparative steady state analysis is undertaken. The results of the numerical analysis qualitatively illustrate the impact of different VAT policies on the real economy, and can also ranks those policies in terms of their impact on welfare.

The results show that the price of financial intermediation is higher under exempt treatment relative to both alternative cases (e.g. full taxation or partial recovery). This forces some firms out of the market and shrinks the aggregate demand for the products of the remaining firms, which lowers overall welfare and tax revenues. In contrast, moving from exempt treatment to full taxation generates a welfare improvement and/or is revenue enhancing. In this regard, if full taxation is not feasible for technical reasons, then any movement away from exempt treatment and closer to full taxation (with higher recovery rates), will potentially attain a higher welfare and revenue. Examples in this regard include the granting of partial input credits for VAT paid on financial inputs, and the zero-rating of B2B transactions.

The negative consequences of exempt treatment is expected to be more severe for small and micro enterprises, as they rely more on loan financing with limited access to stock and money markets, as well as industries with a high reliance on financial payment services (e.g. online payments, credit cards, etc.) and capital intensive sectors.

There are several potential extensions of the paper. Incorporating a heterogeneous payment service intensity parameter would introduce new sources of distortions into the model. Finally, the current model considers a closed economy and opening the model to international markets may prove informative.

References

Auerbach, A. J. and R. H. Gordon (2002, May). Taxation of financial services under a vat. *American Economic Review 92*(2), 411–416.

Bird, R. M., P.-P. Gendron, and J. L. Rotman (2005). Vat revisited a new look at the value added tax in developing and transitional countries. *USAID Report*.

Boadway, R. and M. Keen (2003). *Theoretical Perspectives on the Taxation of Capital Income and Financial Services*, Chapter 2, pp. 31–80. World Bank and Oxford University Press.

Büttner, T. and K. Erbe (2012). Revenue and welfare effects of financial sector vat exemption. *TaxFACTs*.

Caminal, R. (2002). Taxation of banks: A theoretical framework. UFAE and IAE Working Papers 525.02, Unitat de Fonaments de l'Anàlisi Econòmica (UAB) and Institut d'Anàlisi Econòmica (CSIC).

Chia, N.-C. and J. Whalley (1999, November). The tax treatment of financial intermediation. *Journal of Money, Credit and Banking 31*(4), 704–19.

Chisari, O., A. Estache, and G. Nicodème (2013, January). Efficiency and equity effects of taxing the financial sector: Lessons from a cge model for belgium. Working Papers ECARES ecares 2013-01, ULB – Universite Libre de Bruxelles.

Commision, E. (2010). Taxation papers: Financial sector taxation. Technical report, European Commission.

Denis, D. J. and V. T. Mihov (2003). The choice among bank debt, non-bank private debt, and public debt: evidence from new corporate borrowings. *Journal of Financial Economics 70*(1), 3–28.

Feenstra, R. C. (1994, March). New product varieties and the measurement of international prices. *American Economic Review 84*(1), 157–77.

Gendron, P.-P. (2007). Value added tax treatment of financial services: An assessment and policy proposal for developing countries. International Tax Program Papers 0701, International Tax Program, Institute for International Business, Joseph L. Rotman School of Management, University of Toronto, http://ideas.repec.org/p/ttp/itpwps/0701.html.

Greenwood, J., Z. Hercowitz, and G. W. Huffman (1988). Investment, capacity utilization, and the real business cycle. *The American Economic Review 78*(3), pp. 402–417.

Grubert, H. and J. B. Mackie III (2000). Must financial services be taxed under a consumption tax? *National Tax Journal 53*(1), 23–40.

Helpman, E., M. J. Melitz, and S. R. Yeaple (2004, March). Export versus fdi with heterogeneous firms. *American Economic Review 94*(1), 300–316.

Houston, J. and C. James (1996). Bank information monopolies and the mix of private and public debt claims. *Journal of Finance 51*(5), 1863–89.

Huizinga, H. (2002). A european vat on financial services.

IMF (2010). A fair and substantial contribution by the financial sector. Report, IMF.

Jack, W. (2000). The treatment of financial services under a broad-based consumption tax. *National Tax Journal 53(4, Part 1)*, 841–851.

Johnson, S. A. (1997). An empirical analysis of the determinants of corporate debt ownership structure. *Journal of Financial and Quantitative Analysis 32*(1), 47–69.

Kerrigan, A. (2010, March). The elusiveness of neutrality – why is it so difficult to apply vat to financial services? *International VAT Monitor 21*(2), 103–112.

Kleven, H. J., W. F. Richter, and P. B. Sørensen (2000). Optimal taxation with household production. *Oxford Economic Papers 52*, 584–594.

Lockwood, B. (2010). How should financial intermediation services be taxed? Working Paper 1014, Oxford University Centre for Business Taxation, http://ideas.repec.org/p/btx/wpaper/1014.html.

McKenzie, K. J. (2000). Taxing banks. Taxation Discussion Paper 1, ATAX, University of New South Wales.

McKenzie, K. J. and M. Firth (November 3, 2011). The gst and financial services: Pausing for perspective. Technical report, The GST at 20: The Future of Consumption Taxes in Canada, School of Public Policy, University of Calgary.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica 71*(6), 1695–1725.

Piggott, J. and J. Whalley (2001). Vat base broadening, self supply, and the informal sector. *The American Economic Review 91*(4), 1084–1094.

Poddar, S. (2003). *Chapter 12 Consumption Taxes: The Role of the Value-Added-Taxes*, Chapter 12, pp. 345–380. World Bank and Oxford University Press.

PWC (2011). How the eu vat exemptions impact the banking sector. *PWC Report*.

Roussiang, D. J. (2002). Should financial services be taxed under a consumption tax? probably. *National Tax Journal 55*(5), 281–291.

Russ, K. N. and D. Valderrama (2012). A theory of bank versus bond finance and intra-industry reallocation. *Journal of Macroeconomics 34*(3), 652 – 673.

Schenk, A. S. (2009). Taxation of financial services (including insurance) under a united states value added tax. Technical Report 1520704, http://ssrn.com/paper=1520704.

Table 1: Revenue Neutral Comparision of Main Aggregates, %Δ

Recovery	Total Rev.	Loan Int. Rate	Productivity Threshold	Num. of Firms	Output	Consumption
Payment Service Intensity (a) =25% - low case						
0%						
50%	0.219	-0.02%	-0.00033%	0.05%	0.05%	0.04%
100%		-0.04%	-0.00065%	0.10%	0.09%	0.08%
0%						
50%	0.240	-0.04%	-0.00065%	0.22%	0.18%	0.17%
100%		-0.07%	-0.00127%	0.43%	0.36%	0.33%
0%						
50%	0.247	-0.05%	-0.00097%	1.67%	1.34%	1.27%
100%		-0.10%	-0.00187%	2.89%	2.31%	2.19%
Payment Service Intensity (a) =75% - high case						
0%						
50%	0.150	-0.02%	-0.00034%	0.36%	0.29%	0.27%
100%		-0.04%	-0.00065%	0.68%	0.55%	0.52%
0%						
50%	0.165	-0.04%	-0.00068%	1.40%	1.12%	1.06%
100%		-0.07%	-0.00127%	2.58%	2.05%	1.95%
0%						
50%	0.169	-0.06%	-0.00107%	7.17%	5.66%	5.40%
100%		-0.10%	-0.00187%	10.86%	8.55%	8.16%

0 % recovery is "exempt treatmen"t, 50 % recovery is "partial recovery" and 100 % recovery is "full taxation". Payrol tax and capital (interest) income tax are assumed to be 10 % and 15 %. Depretiation allowance is 75 %. Base case in all computations is exempt treatment.

Table 2: Revenue Neutral Comparision of Prices and Inputs, %Δ

Recovery	Total Rev.	Wage Rate	Agg. Price	Labour	Capital
Payment Service Intensity (a) =25% - low case					
0%					
50%	0.219	0.019%	-0.007%	0.025%	0.062%
100%		0.035%	-0.013%	0.048%	0.120%
0%					
50%	0.240	0.088%	-0.021%	0.109%	0.232%
100%		0.171%	-0.041%	0.212%	0.454%
0%					
50%	0.247	0.712%	-0.117%	0.829%	1.602%
100%		1.222%	-0.202%	1.427%	2.774%
Payment Service Intensity (a) =75% - high case					
0%					
50%	0.150	0.151%	-0.027%	0.178%	0.348%
100%		0.287%	-0.050%	0.337%	0.661%
0%					
50%	0.165	0.595%	-0.099%	0.695%	1.333%
100%		1.091%	-0.182%	1.275%	2.452%
0%					
50%	0.169	3.019%	-0.480%	3.517%	6.707%
100%		4.530%	-0.716%	5.284%	10.172%

0 % recovery is "exempt treatmen"t, 50 % recovery is "partial recovery" and 100 % recovery is "full taxation". Payrol tax and capital (interest) income tax are assumed to be 10 % and 15 %. Depretiation allowance is 75 %. Base case in all computations is exempt treatment.

Table 3: Revenue Neutral Welfare Comparison

	a=25% (low case)			a=75% (high case)		
Recovery	TR	Welfare	%Δ	TR	Welfare	%Δ
0%	0.219	0.70757		0.150	0.53333	
50%		0.70783	0.037%		0.53463	0.245%
100%		0.70807	0.071%		0.53581	0.465%
0%	0.240	0.60554		0.165	0.45379	
50%		0.60647	0.153%		0.45817	0.964%
100%		0.60735	0.299%		0.46182	1.768%
0%	0.247	0.52120		0.169	0.38855	
50%		0.52714	1.141%		0.40764	4.913%
100%		0.53145	1.968%		0.41732	7.405%

0 % recovery is "exempt treatmen"t, 50 % recovery is "partial recovery" and 100 % recovery is "full taxation". Payrol tax and capital (interest) income tax are assumed to be 10 % and 15 %. Depretiation allowance is 75 %. Base case in all computations is exempt treatment.

Table 4: Revenue Neutral Comparison of Source of Tax Revenue , %Δ

Recovery	TR	Loan S. VAT Rev.	Payment S. VAT Rev	Consumption VAT Rev.	Payroll Tax Rev.	Capital Income Rev.	Total Tax Revenue
	Payment Service Intensity (a) =25% - low case						
0%		0.08%	0.50%	35.37%	40.23%	23.82%	100%
50%	0.219	0.04%	1.08%	34.79%	40.25%	23.84%	100%
100%		0.00%	1.65%	34.23%	40.27%	23.85%	100%
0%		0.11%	0.74%	51.91%	29.52%	17.72%	100%
50%	0.240	0.05%	1.65%	50.96%	29.58%	17.76%	100%
100%		0.00%	2.53%	50.04%	29.64%	17.79%	100%
0%		0.12%	0.89%	61.49%	23.31%	14.18%	100%
50%	0.247	0.06%	2.01%	59.86%	23.68%	14.40%	100%
100%		0.00%	3.07%	58.45%	23.94%	14.55%	100%
	Payment Service Intensity (a) =75% - high case						
0%		0.08%	1.46%	34.81%	39.60%	24.05%	100%
50%	0.150	0.04%	3.08%	33.02%	39.73%	24.13%	100%
100%		0.00%	4.54%	31.42%	39.84%	24.20%	100%
0%		0.11%	2.17%	50.91%	28.96%	17.85%	100%
50%	0.165	0.05%	4.64%	47.91%	29.33%	18.07%	100%
100%		0.00%	6.80%	45.30%	29.65%	18.26%	100%
0%		0.12%	2.60%	60.18%	22.82%	14.28%	100%
50%	0.169	0.06%	5.47%	54.98%	24.33%	15.16%	100%
100%		0.00%	7.92%	51.35%	25.11%	15.62%	100%

0 % recovery is "exempt treatmen"t, 50 % recovery is "partial recovery" and 100 % recovery is "full taxation". Payrol tax and capital (interest) income tax are assumed to be 10 % and 15 %. Depretiation allowance is 75 %. Base case in all computations is exempt treatment.

Table 5: % Change in METR under Different Policies

t_k	ρ	VAT	Loan Int. Rate	METR	%Δ $METR_1$	%Δ $METR_2$	%Δ $METR_3$
25%	0%	5%	17.468%	18.90%			
30%			17.843%	20.60%			9.01%
35%			18.275%	22.48%			9.12%
40%			18.780%	24.56%			9.26%
25%		25%	17.490%	19.00%		0.55%	
30%			17.865%	20.70%		0.48%	8.94%
35%			18.297%	22.58%		0.42%	9.05%
40%			18.802%	24.65%		0.36%	9.20%
25%	100%	5%	17.462%	18.87%	-0.15%		
30%			17.837%	20.58%	-0.13%		9.03%
35%			18.269%	22.46%	-0.11%		9.14%
40%			18.774%	24.54%	-0.10%		9.28%
25%		25%	17.462%	18.87%	-0.69%	0.00%	
30%			17.837%	20.58%	-0.61%	0.00%	9.03%
35%			18.269%	22.46%	-0.53%	0.00%	9.14%
40%			18.774%	24.54%	-0.46%	0.00%	9.28%

%Δ $METR_1$ is the % change in METR due to a policy change from exempt treatment to full taxation; %Δ $METR_2$ is the % change in METR due to a change in the VAT rate; %Δ $METR_3$ is the % change in METR due to change in $t_{k,}$ holding everything else constant. Depretiation allowance is 75 %. 0 % recovery is "exempt treatmen"t and 100 % recovery is "full taxation".

Figure 1: Laffer Curve, Revenue Generation by VAT rate (Under Full Taxation)