UNITED NATIONS
UNIVERSITY
UNU-WIDER

WIDER Development Conference

Public economics for development

5-6 July 2017 | Maputo, Mozambique

Value-added Tax Remittance Responsibility: Impact on Firm Compliance and Production

Preliminary - Please do not cite

Bassirou Sarr, PSE[1]

June 19, 2017

In this paper, I study the effects of ubiquitous withholding VAT schemes on the tax compliance and productivity of firms. With insufficient administrative resources, third-party reports and the existence of paper trails are not sufficient conditions for compliance. Therefore, tax authorities in developing countries use withholding VAT as an advance payment to prevent fraud, evasion, and avoidance. This paper contributes to a crucial area of tax systems research which investigates the relevance of remittance responsibility, as well as the relative importance of withholding and information reporting on compliance. However, withholding also makes firms claimants of excess VAT credits. Credit reimbursement delays create distortions on resource allocation and adversely impact the productivity of businesses. I present stylized models which address these issues and use administrative data from Senegal, a country which terminated withholding, to test some of the hypotheses.

Keywords: Tax Remittance Responsibility, Withholding, Tax Evasion, Production, Value Added Tax

JEL Codes: H260, H250, H21

[1] Paris School of Economics, E-mail: bassirou.sarr@psemail.eu, sarrbassirou@gmail.com Address: Ecole Normale Supérieure, 48 Bvd Jourdan, 75014, Paris, France. I would like to thank my advisor Denis Cogneau and Francois Gerard for their support, advice and encouragement. I also received valuable comments from Joel Slemrod and Alexander Persaud. I thank the Office of Tax Policy Research at the University of Michigan for its hospitality and financial support. Oumar Diop Diagne at DGID has been over the years a great collaborator and advisor on topics pertaining to the Senegalese tax system, I am grateful to him for his support and advice. I am also grateful to DGID for the collaboration which made the data available. All errors and omissions are mine.

1 Introduction

Despite the widespread adoption of the value-added (VAT) tax in developing countries over the last fifty years, its revenue performance is often stalled by design and implementation issues. Chief among these implementation issues is the administration of withholding at source schemes, which are prevalent on business-to-business transactions in many African and Latin American countries. Withholding VAT is a system through which suppliers only receive a fraction – or sometimes none – of the VAT on sales to designated withholding agents[2]. Instead, they receive withholding certificates, which become credits in their VAT accounting. In some countries, banks accept these vouchers as collateral for short-term loans, albeit with a haircut. Without timely refunds, the system also contributes to the formation of excess-credits. The tax authority owes money to firms for long periods (See Ebrill et al. 2001, chapter 15). Needless to say, withholdings on firms' sales may create large opportunity costs. Businesses denounce these excess-credits as interest-free loans to the state. Credit formation due to withholding also hinders firm production via constraints on their operating capital for input expenditures. In Senegal, for instance, Figure 5 shows that credits rolled for more than three months are, on average, 12% of monthly turnover. Hence, withholding policies may introduce distortions which affect the allocation of production factors and the productivity of firms. In particular, withholding VAT may create substantial inefficiencies for input-intensive sectors. But, under optimal commodity taxation models, such production inefficiencies may be tolerated if administrative costs are taken into account (Dharmapala et al. 2011, Yitzhaki 1979, Diamand and Mirrlees 1971).

From the state's point of view, the goals of a withholding VAT system are threefold. First, as an advance payment, it maximizes collection and ensures that tax revenues reach the Treasury promptly. Second, as large business based remittance, withholding creates economies of scale and reduces administrative costs (Dharmapala et al. 2011, Kopczuk and Slemrod 2006). The tax authority can, therefore, conduct cost-efficient audits by processing information from large firms. In general, large corporations are more likely to possess superior bookkeeping methods compared to smaller and less sophisticated ones. Kleven et al. (2016) argue that, as fiscal intermediaries, large firms are unable to misreport true liabilities because verifiable book evidence and headcount act as a joint deterrence to evasion. With a large number of employees, any of them can denounce existing collusions about the information reported to the tax authority. Bachas and Jensen (2017) make a similar argument by documenting the existence of firm size effects in tax enforcement and compliance. They show that the size gradient in enforcement is highest for less developped countries. Yet, enforcement tools must be credible. And the tax authority should have enough

[2] Stated differently, designated withholding agents (buyers) are responsible for the remittance of VAT.

technical capacity to exploit third-party reports or a mechanism to raise the probability of detecting evasion (Naritomi 2016). This is not often the case in developing countries. Kuchumova (2017) presents a theoretical framework in which there is a trade-off between information reports and audits, arguing that an optimal allocation of resources between the two can improve tax compliance. However, this result is unlikely to be useful in developing countries where both tools tend to be weak. With sub-optimal enforcement, even the proper processing of third-party information and the existence of paper trails are not sufficient conditions in curbing tax evasion. When audits uncover discrepancies through cross-checks, firms can offset the increase in reported turnover by a similar increase in other margins (expenses), which are not subject to tax filling (Carrillo et al. 2016, Slemrod et al. 2015). Hence, the third goal of a withholding system. It creates a compliance default on the extensive and intensive margins. As the value-added tax is retained and remitted by the supplier, withholding reduces evasion possibilities for businesses outside the VAT system or those which are identified in the registry but would still fail to remit VAT [3]. On the intensive margin, withholding reduces evasion with the *post collection* audit of requests for credit reimbursements [4]. Despite its prevalence in developing countries, withholdings on inter-firm transactions have, so far, been subject to little analytical consideration. In the recent public finance literature, only Brockmeyer and Hernandez (2017) have studied the introduction of a withholding scheme in Costa Rica's sales tax. In many other countries, though, policy advice seems to go in the opposite direction, arguing for the elimination of withholding VAT.

In this paper, I use a reform removing withholding from Senegal's VAT design to address three related issues on both theory and empirics. The first problem is about firm compliance when we eliminate withholding from a tax system. Under such a scenario, the tax remittance responsibility shifts from the buyer to the seller. Even though optimal tax theory models assume the irrelevance of remittance responsibility when the tax authority observes liability at little cost, this assumption breaks down with evasion possibilities (Slemrod 2008). There are two hypotheses on the effects of withholding VAT on compliance. The first one is that VAT generates information trails which are sufficient in reducing evasion, regardless of withholding. Thus, if the tax authority has the technology to cross-check information reports, withholding VAT does not change compliance decisions. However, under weak technological capacity, it could be more relevant than information reports. Thus, the second hypothesis contends that withholding VAT creates a compliance default and its termination leads to greater evasion. When it is in place, the tax authority locks in tax payments before an audit. So, if removed, heterogeneous responses could be observed based on audit probability, which is often increasing in turnover. As a result, large taxpayers may not respond at all to the termination of withholding VAT. But small and medium-size firms with lower audit

[3] This is often the case for withholdings on employee earnings.
[4] Fraudulent claims are important sources of revenue losses.

probabilities will evade more. Moreover, because of limited resources devoted to tax audits, the evasion of smaller companies will be on the extensive margin. This second hypothesis justifies decisions to keep withholding as a feature of VAT design. To my knowledge, despite the widespread use such designs, no study has explored the compliance effects of withholding VAT. African countries such as Mali, Togo, and Senegal have removed it from their tax codes. But these policy changes are yet to be evaluated. With the administrative data from Senegal, I investigate whether the termination of withholding leads to greater non-compliance on the extensive and intensive margins. *[Insert findings on issue 1 here and discuss intensive and extensive margin mechanisms].*

The second issue relates to the behavior of the withholding agent. With the management and issuance of certificates as well the remittance responsibility, they bear part of the administrative costs. I propose a theory on the circumstances under which withholding agents fail to fully remit withholdings. I hypothesize that under weak audit probabilities or low penalties, withholding agents might remit less than they are supposed to. This simple result speaks to many anecdotes of withholding agents who are also part of the state apparatus. They are subject to weak enforcement tend to keep the withheld sums.

The third issue is that we expect withholding to introduce distortions in the production decisions and productivity of firms. There is a tension between the revenue maximization objective of the state and the necessity to run a VAT system which minimizes distortions. Indeed, withholding regulations and costly refund claims are major impediments to the neutrality of the value-added tax in many developing countries (See Chambas 2014 for a review on the subject). In Senegal, for instance, withholding affects a relatively high share of firms. Before the reform, on average, about 20% of Large Taxpayer Unit(LTU) or Medium Taxpayer Unit (MTU) firms were subject to withholdings every year. When withholding generates excess-credits which are reimbursed with lengthy delays, VAT can, indeed, become a tax on production instead of a tax on final consumption [5]. Firms often consider this as an impediment to efficiency because they become claimant of credits[6] ,which reduce their operating capital. Several questions need related to this issue merit attention. For instance, how does withholding affect the use of different inputs in production? In the case of production for domestic consumption, sectors that use intermediate inputs more intensively could grow less as result of the policy. Therefore, firms in these industries could benefit from the reduction of credit formation through the cancellation of withholding on VAT. *[Insert findings on issue 2 here].*

This paper is organized as follows. In section 2, I review the literature and point to the semantic distinction between withholding and third-party information reports and more generally

[5] Because of this, Ebrill et al. 2001 went so far as describing the dysfunctions in the refund process as the "Achilles heel" of the VAT (See Chapter 15 in [?]).

[6] Interest-free loans to the state

to address the contributions of this paper on the role of withholding in tax systems. Section 3 presents stylized facts about withholding VAT. In section 4, I present theoretical models on the compliance and production decisions of firms. Section 5 presents the institutional background in Senegal while Section 6 describes the data. In Section 7, I present preliminary empirical results.

2 Related Literature

In this study, I contribute to the new wave of research on optimal tax systems (Slemrod and Gilitzer 2013). This approach considers issues in three research areas. The first one is on the design of optimal tax rates and bases. To address issues related to this strand of the literature, many contributions on taxation in developing countries have used administrative data to cover the design of optimal rates and bases in the context of weak enforcement capacity (Best et al. 2015, Waseem 2015, Best 2014). The second area covers enforcement rules in taxation such as the use of non-rate instruments, including third-party information reports, audits and enforcement actions (Pomeranz 2015, Brockmeyer et al. 2017). And the third one is about remittance rules, the designation of taxpayers responsible for collecting and remitting taxes to the tax authority, the frequency of the transfers as well as the related reporting requirements. Here I address this third branch of the literature on tax systems. Optimal tax models are relatively silent about the role of remittance responsibility in revenue collection. However, with evasion possibilities and limited enforcement resources of revenue authorities, remittance responsibility could play a role in increasing compliance and reducing administrative costs. In other words, it matters whether it's a supplier or a buyer which remits VAT in a firm-to-firm transaction. When third-party reports have a limited impact on compliance or when certain taxpayers (small firms) internalize low probabilities of audits, withholding could play a role in reducing evasion. Few studies have addressed the role of remittance responsibility in VAT design. And this is perhaps because it tends to be bundled with the effect of third-party information reports on evasion. I discuss this issue in 2.1.

The second area where I seek to make a contribution relates to the evasion of indirect taxes such as VAT or goods and services taxes (GST). Recent studies have notably focused on the evasion of income taxes (Kleven et al. 2011, Kleven et al. 2015 and Kleven and Waseem 2013) and corporate income taxes (Best et al. 2015). But very few contributions have considered the issue of indirect tax evasion. Except for Keen (2008) and Emran and Stiglitz (2005), issues about the design and implementation of VAT in developing countries have received less theoretical and empirical consideration. In particular, despites the widespread presence of withholding in the VAT design of developing countries, its effects on compliance and production decisions of firms subject to withholding (withholdees), as well as withholding agents are not well understood. Finally, this study also relates to the literature on the effects of policy distortions intra-firm input allocation and the distribution of resources across firms.

2.1 Withholding vs. Third-party information reports

The 2012 tax reform in Senegal is unique insofar as it provides an opportunity to shed more light on the relative merits of withholding and third-party information reports in explaining observed levels of compliance. Before the change, they could have acted jointly to reduce evasion. But with the reform, only the information trail channel operates. Thus, the setting is suitable for finding evidence on the importance of each instrument. This is a novel contribution because the actual effect of information reports is not always clear when operates alongisde withholding.

Therefore, before I address withholding in the public finance literature, I find it necessary to stress the semantic distinction between withholding and third-party reported information. This distinction is important because the performance of modern tax systems is often attributed to the latter when in practice both policies, though different, are often implemented together. Therefore, it's worth pointing that a particular tax can include elements of third-party reported information but no requirement on withholding. Take the basic design of VAT in a country with high tax enforcement, say a developed country. The invoice-credit nature of the mechanism requires firms in intermediate stages of production to report verifiable information on their transactions. VAT can, therefore, have self-enforcement properties without any form of withholding.

Now, consider personal income taxes. Not only do they include third-party reported information insofar as the firm declares wages and salaries but, on top of that, they are also withheld at source. Thus, even with the ability to verify employer reports on income taxes, it's unclear whether evasion rates are low because of information reporting or withholding, two different aspects of tax design. For instance, Kleven et al. (2011) conduct an experiment in Denmark and claim that while there is substantial evasion for self-reported income, there is almost no evasion for third-party reported income. They extend the Allingham and Sandmo (1972) model to account for the heterogeneity in sources of income and the sources of information reports and audit probabilities associated with each type of revenue. In particular, they claim that third-party reporting resolves the puzzle in the Allingham-Sandmo model, which predicts high evasion rates when audit probabilities are small. However, this argument is flawed as it attributes high compliance rates to information reports when in fact personal income tax withholding at source could preempt employee evasion, even with no information reports.

Also, to make a point a point on the distinction between the two instruments, it's also important to recall the history of the introduction of the former in modern tax systems. For example, Slemrod 2008 shares few historical episodes in which withholding are introduced to increase tax compliance. When Catholic and other Irish farmers refused to remit tithes to the Protestant church, a law required landlords to withhold and remit the payments. As a result, compliance increased. And perhaps, the most successful example of tax withholding's effect on compliance is the withheld at source personal income taxes. The revenue-efficient nature of withholding at source is seen as early as 1803 when Henry Addington, with the breakdown of the Treaty of Amiens, used personal income taxes withheld at source to finance war efforts. About the more recent history of withheld taxes, Dusek and Bagchi 2016 use variations in the adoption of income tax withholding at source by U.S. states from the 1940s to the 1970s to show that holding tax rates constant, and there is 24% increase in income tax revenue after the implementation of withholding policies.

Specifically, for VAT, the literature is yet to show substantial evidence on the relative importance of third-party reporting and information trails when withholding is an instrument in its design. The only existing studies have claimed positive effects of information trails on compliance but in contexts where withholding is also present. For instance, considering the role of third-party information trails on the compliance decision of Chilean firms subject to VAT, Pomeranz 2015 has studied the response to an increase in perceived audit probability in the presence of paper trails. The paper presents two conflicting hypotheses on the role of paper trails. On the one hand, given equal evasion opportunities in all transactions, firms would exhibit a greater response to audit probability in the presence of paper trails. On the contrary, in the presence of paper trails, lower responses to increased audit probability can also indicate that,ex ante, the underlying cross-bookkeeping of the VAT mechanism was itself a deterrent to evasion. In particular, because of the paper trail generated through the invoice-credit system, in theory, VAT has self-reinforcing properties due to opposing fraud incentives in intermediate states of the production chain. The incentive structure of VAT only breaks down in the final sale to the consumer,who unlike firms, has no incentive to ask for receipts. Pomeranz (2015) differentiates between inter-firm transactions and sales to final consumers in evaluating the impact of audit signals on evasion. With this distinction, the study finds that announcing additional audits has a limited impact on the reporting of VAT transactions with paper trails. But the role of withholding (administrative substitution of the entity responsible for VAT remittance) on compliance is not mentioned at all in this argument proposing paper trails as a sufficient deterrent to evasion. In fact, lower responses to greater audit probabilities could be observed not only because of the paper trails but also because there were not evasion opportunities, to begin with. Indeed, the Chilean tax code allows the designation of some buyers as responsible for retaining VAT on their purchases.

2.2 Withholding, optimal taxation, and tax evasion

Standard optimal tax models are silent on the effects of remittance responsibility. However, when these models consider evasion possibilities, the irrelevance assumption about the entity (buyer or seller) responsible for remittance breaks down (Slemrod 2008 and Kopczuk et al. 2016). For instance, in the implementation of VAT, if changes in the withholding system switch the remittance responsibility from all firms to a smaller set (or vice versa as is the case in Senegal), the irrelevance assumption no longer holds. In the more recent literature using administrative data, Carillo et al. 2012 and Brockmeyer and Hernandez (2017) have explored the effect of withholding on compliance. In particular, Brockmeyer and Hernandez (2017) study the importance of information and withholding on the compliance to a business sales tax in Costa Rica. The authors document compliance gaps on the extensive and intensive margins and to a lesser extent on the payment margin. These differences provide arguments for the use of withholding as an enforcement tool. They find that coverage by business sales tax withholding is associated with higher reported taxable income. Doubling the withholding rate leads to an increase of 40% in the tax payments of firms covered by the withholding scheme. The mechanisms for the higher payments are an incomplete reclaim of withheld amounts and lower misreporting. However, the contribution of the present paper is different from Brockmeyer and Hernandez (2017) because I focus on VAT which has been the most popular tax instrument in developing countries over the last fifty years albeit with significant impediments such as excess-credits, often due to withholding schemes. Second, I study the removal of a withholding scheme as a natural experiment rather than its reinforcement, as is the case in Brockmeyer and Hernandez (2017).

With the Senegalese reform, I will present a theory which elicits the effect of withholding on the evasion and production of firms and test empirical predictions with administrative data. I provide and extension of Yaniv (1999) which shows that while withholding is irrelevant under the expected utility theory models of evasion (Allingham and Sandmo 1972), the size of advance payment affects evasion decisions under a prospect theory model. In particular, prepaid taxes can substitute for costly enforcement. The model I present explains enforcement conditions (audit probability and penalties) under which withholding in a VAT system can deter evasion. For a theoretical model of the withholding agent's behavior, I adopt a similar framework to Yaniv (1998) which considers conditions under which an employer fails to remit personal income tax withholdings.

2.3 VAT Withholding, resource misallocation, and productivity

Withholding VAT also fits into the literature dealing with the effects of government policies on resource allocation and firm productivity. Beyond differences in the speed of technology adoption, aggregate differences in total factor productivity (TFP) depend on the distribution of production factors within production units but also across heterogeneous ones. Recent contributions to this literature have sought to explicitly establish a relationship between distortions introduced by particular institutions or policies to resource misallocations and lower aggregate total factor productivity (TFP). The diversion of resources from high productivity firms to unproductive ones could explain large cross-country differences in growth outcomes. Following the seminal contribution of Restuccia and Rogerson (2008), Hsieh and Klenow (2009) provide the first quantitative evaluation of the extent of resource misallocation across manufacturing establishments in China and India compared to the United States. They show that substantial differences in the marginal products of labor and capital exist within industries. In a counterfactual measurement, reallocating capital and labor to equalize their marginal products to the extent observed in the U.S. would result in $30-40\%$ TFP gains in China and $40-60\%$ TFP gains in India. In the same vein, other contributions by Cirera et al. (2017) have replicated the measurement of resource misallocation in Africa. They find that a reallocation equalizing marginal returns across different firms within the manufacturing sector would increase manufacturing productivity by 163% in Kenya and 31% in Côte d'Ivoire.

Based on a survey spanning many literatures, Restuccia and Rogerson (2017) describe three categories of factors resulting in resource misallocation. First, they list statutory provisions, such as features of the tax system which vary with firm characteristics (audit probability based on size, minimum turnover taxes), regulations on business size or market access. Second, there are distortionary provisions introduced by the state or banks to penalize or favor particular firms. Examples include tax incentives, procurement rules or selective tax enforcement. And third, we have market imperfections and specifically market structures (monopoly power), market frictions and the enforcement of property rights which create disincentives for individual firms. These idiosyncratic policies and market distortions can build wedges in the prices of each production unit, but the aggregate capital accumulation and aggregate relative prices are constant. Other frictions that have been subject to theoretical and empirical work are related to trade policies (Melitz 2003), imperfect credit markets (Midrigan and Xu 2014, Kalemli-Ozcan and Sorensen 2012) and capital adjustment costs (Asker et al. 2014). However, so far, the effects of specific aspects of tax systems on resource allocation and productivity are yet to be explored using the analytical tools of this literature (IMF 2017 has provided insights on productivity and tax-related distortions). Withholding on VAT is one such policy, which many anecdotes describe as having distortionary effects on the production decisions of firms. It is thought to create disincentives to input intensive firms which provide goods and services to designated withholding agents. For those businesses, withholding contributes to the formation of excess-credits which, in turn, affect resource allocation and productivity.

When withholding has excess-credits as its byproduct, it can introduce differences in the marginal products of capital. To illustrate this point, consider a hypothetical example like the one presented in Hsieh and Klenow (2009). Suppose we have two firms with identical technologies. But one was not subject to withholding (say a state-owned bank) and the other one which is subject to withholding. When the latter faces credit formation with reimbursement delays, its implicit cost of capital can be higher compared to the firm without excess-credits (imperfect credit markets and withholding certificates accepted with a haircut). If both businesses equate their marginal product of capital (MPK) with their cost of capital, the MPK of the firm without excess-credits is lower. This situation produces resource misallocation since total output per worker and TFP would be greater if capital is transferred from the production unit with low MPK to the one with high MPK.

3 Withholding VAT: A brief introduction and stylized facts

In most countries, withholding VAT is a credit system through which government institutions, major corporations; parastatal organizations act as VAT collection agents on behalf of the tax authority. When invoiced by suppliers – regardless of the vendor's tax registration status – withholding agents only pay a price net of VAT. They keep the VAT component of the gross price and remit it to the tax authority. Suppliers receive credit certificates. They can use the vouchers as credits in their VAT accounting and later as a supporting document in the refund process.

To present the inner workings of a typical withholding VAT system, let's consider a transaction between two economic agents, A and B. Let us assume a state institution, say the Ministry of Agriculture, is designated as a withholding agent [7], call it A, purchases a 100 million worth of goods and services comprising a VAT exclusive price of 90 million and 10 million of VAT from its supplier, call it firm B. Under a 50% withholding rate, A only pays firm B 95 million. As a withholding agent, A withholds and remits 5 million to the tax authority. Firm B is entitled to a withholding certificate worth 10 million.

Now, also assume that at the end of the month, firm B determines that it has a net VAT due of 1 million. Then, it has VAT credit claim of 4 million from the tax authority, resulting from the difference between the withheld amount of 5 million and its VAT liability of 1 million. In this scenario, Firm B would submit a refund request of 4 million. Figure 1 illustrates the case of withholding VAT with a credit claim.

But in most cases, it's worth noting that firm B faces delays and uncertainty before repayment. Audits and checks on submitted documents – including certificates received from

[7] The Ministry of Agriculture is a suitable example, as it does not sell goods and services. It would only be responsible for remitting VAT withholdings to the tax authority. This simplifies the examples, which follow.

withholding agents – or just administrative inefficiency often extends statutory repayment periods. It is also important to note the cash flow implication for firm B when there is no withholding policy. As Figure 2 illustrates, firm B would have received a 100 million VAT inclusive price from A. B would only have to transfer directly the net VAT due of 1 million to the state.

The cases described herein are from the simplest design of a withholding VAT scheme. Tax codes in Latin America and Africa include more complex designs with parameters on the computation of the withheld amount, the goods and services to which the policy applies and the ability to roll forward excess-credits and obtain a refund. The different schemes for selected countries in Africa and Latin America are summarized in Table 1.

4 Analytical Framework

In this section, I present an analytical framework for the evasion and production decisions of firms. Both the withholding agent and its supplier can seek to evade taxes. The former commits fraud to compensate for administrative costs it bears on behalf of the tax authority. The latter could be non-compliant because the withholding system generates excess-credits which limit its operating capital. First, I present stylized models on the evasion decisions of the withholdee and the withholding agent. Then I analyze production decisions and productivity under a withholding system.

4.1 The compliance of the withholdee

Consider a firm subject to withholding (seller, B in Figure 1) operates in perfect competition with free entry and exit, has a production function H and has expenditures E. Its value-added is, therefore, $VA_B = H - E$. Assume that the state sets three policy parameters, namely the VAT rate τ, a withholding rate δ on the VAT amount to be remitted by the buyer (withholding agent) and $\lambda_B > 1$, the penalty parameter if evasion is detected. The withheld amount on sales is, therefore, $W = \delta \cdot \tau \cdot H$. The firm can choose to report a value-added of $X_B \leq VA_B$. Firm B's net income in the non-detected and detected states of evasion write as follows

$$Y_B^{nd} = (VA_B - W) + (W - \tau \cdot X_B)$$
$$Y_B^{d} = (VA_B - W) + (W - \tau \cdot X_B - \lambda_B \cdot \tau(VA_B - X_B))$$

Note that in both cases, without withholding or when the tax authority reimburses credit right away, W drops out of the model. We fall back to the classic tax evasion model. However, assume that the state does not reimburse credits in a timely fashion. Then, the firm's fraud decision depends on whether it has a net credit $(W > \tau \cdot VA_B)$ or net debit position $(W < \tau \cdot VA_B)$ vis-à-vis the tax authority[8]. In particular, the company cares about its net cash position before filling its monthly VAT return. In other words, its evasion decision is influenced by how much cash it will have for its operations as a result of the advance payment. The net position in the non-detected and detected states of the world write

$$\Delta Y_B^{nd} = W - \tau \cdot X_B$$
$$\Delta Y_B^{d} = W - \tau \cdot X_B - \lambda_B \cdot \tau(VA_B - X_B)$$

In particular, similar to Yaniv (1999), under prospect theory, the value function of the firm $v(\cdot)$ is concave for net debit positions and convex for net credit positions. The firm's reference point is $VA - W$. It solves

$$\underset{X_B}{\mathrm{M}}\mathrm{ax} V = v(\Delta Y_B^{nd}) + p_B v(\Delta Y_B^{d}) \qquad (1)$$

Where, according to common practice in the decision to audit VAT returns, firms with consistently net credit positions are likelier to have their VAT returns audited. Hence, I assume $p_B(W - \tau \cdot VA_B)$ and $p_{B'}(\cdot) > 0$. In 1, when $W > \tau X_B$, the firm has a certain net credit position. So consistent with prospect theory, this certain refund is associated with a probability of unity.

Proposition 1. When withholding VAT is above the actual liability (net credit position, $W > \tau \cdot VA_B$), the entry condition into evasion is $\frac{dV}{dX_B} < 0$ at $X_B = W$. This condition is summarized by

$$p_B \lambda_B < \frac{v'(W - \tau VA_B)}{v'(0)^-} < 1$$

The entry condition in **Proposition 1** depends on the size of the withheld amount W. When W are set sufficiently high, the tax authority can reduce evasion on both the intensive and extensive margins, even when it has limited enforcement capacity. So even as p is close to null, evasion can be prevented by setting withholding sufficiently high. This is illustrated in Figure 3. Given that the penalty parameter λ_B is set by statute, when limited administrative resources constraint audit probabilities to small values, only withholdings above the true liability can act as an enforcement tool. In the case of Senegal, the high withholding rates seemed to reflect this theoretical result. Except for LTU firms, withholding agents paid VAT exclusive prices and remitted the full tax to the tax authority.

Proposition 2. When withholding VAT is below the actual VAT liability ($W < \tau VA$), with low penalties or audit probabilities, the firm faces a lower marginal cost of evasion and will evade as

[8] Net credit means an overpayment while in net debit, the firm must make an additional payment.

long as $p_B \lambda_B \leq \frac{v'(0)^+}{v'(0)^-}$.

According to **Proposition 2**, under low audit probabilities or small penalties, firms no longer subject to withholding VAT will remit less tax. However, in the administrative organization of revenue authorities in many countries where withholding VAT is a feature of the tax system, taxpayer segmentation by turnover also creates a notched audit probability. For instance, Large Taxpayer Unit firms tend to have higher audit probabilities. Thus, the removal of withholding VAT may have no effect on their reporting and payment behavior. However, when withholdings are below the actual VAT liability (including the case of withholding policy removal), the risk of evasion will be higher for medium and small businesses. Senegal eliminated withholding VAT in two successive steps, first for large taxpayers and then for medium-size taxpayers. This is a useful context for a careful evaluation of the result presented in **Proposition 2**.

4.2 The remittance decision of the withholding agent

Assume that the withholding agent (the buyer, firm A in Figure 1) operates in a competitive market with a single good. Let F be its output and H its costs which are also what it purchased from the withholdee (sales of supplier and firm B in Subsection 4.1). The withholding agent's value-added is $VA_A = F - H$. It withholds an amount $W = \delta \cdot \tau \cdot H$ from the purchase transaction with the withholdee. However, the firm can choose to remit less than the full withheld amount. It can do so through an under-reporting of the costs of inputs it paid to its supplier, which allows it to capture part of the VAT withheld at source. The withholding agent has a decision set consisting of a menu of two cost functions $\{H, \underline{H}\}$, respectively represent an accurate reporting and under-reporting. Such a decision can, for instance, be justified by a desire to compensate for the costs it incurs to administer the collection and reporting on behalf of the tax authority. Hence it declares the withheld amount $W_A = \delta \cdot \tau \cdot \underline{H} \leq W = \delta \cdot \tau \cdot H$. Correspondingly, the declared value-added amount $X_A = F - \underline{H} \leq VA_A = F - H$.

It's also worth mentioning that the firm could, in principle, use other channels to evade VAT. It can underreport or overestimate its costs. Assuming the withholding agent is a large firm whose operations are visible, it might be strategic on its part to fraud on withholdings, anticipating that with limited resources, the tax authority is unlikely to audit its actual input costs via the books of its smaller suppliers. If caught after an audit, the withholding agent faces a penalty $\lambda_A > 1$. When the withholding fraud goes unnoticed, the firm maximizes the following objective function

$$Y_A^{nd} = VA_A + (W - W_A) - \tau X_A$$

$$Y_A^{d} = VA_A + (W - W_A) - \tau X_A - \lambda_A (W - W_A)$$

Firm A's net payment position in the non-detected and detected cases write

$$\Delta Y_A^{nd} = (W - W_A) - \tau X_A = \delta\tau(H - \underline{H}) - \tau(F - \underline{H})$$

$$\Delta Y_A^{d} = (W - W_A) - \tau X_A - \lambda_A(W - W_A) = \Delta Y_A^{d} = \delta\tau(H - \underline{H}) - \tau(F - \underline{H}) - \lambda_A(\delta\tau(H - \underline{H}))$$

Next, consistent with the practice of taxpayer segmentation in many developing countries, I assume that to reduce its administrative costs and to maximize its coverage of smaller suppliers, the audit probability of withholding agents is contingent on size. More formally, $\frac{\partial p_A}{\partial F} > 0, p_A(0) = 0$ and $\lim_{F \to \infty} p_A = 1$. Furthermore, I assume the manager is risk-averse with utility function $v(\cdot)$. The firm's manager solves

$$\underset{\underline{H}}{\mathrm{Max}} V = (1 - p_A) v(\Delta Y_A^{nd}) + p_A v(\Delta Y_A^{d})$$

Considering an interior solution, setting first-order condition equal to zero yields

$$p_A \lambda_A < \frac{1-\delta}{\delta}$$

Hence, the evasion decision depends on the withholding rate δ as well as enforcement parameters, namely the audit probability p_A and the penalty parameter λ_A. Below are two propositions which follow from this result.

Proposition 3 An increase in the withholding rate δ raises the relative price of the marginal benefit of a truthfully reported unit cost of inputs (declare and remit less than what was withheld forom the supplier).

Proposition 4. Everything else the same, a decrease in the penalty parameter λ_A or a decrease in the audit probability p_A will result in fewer remittances by the withholding agent.

Proposition 4 is an important result since it emphasizes relevant parameters in the compliance decision of withholding agents. Though the risk of non-remittance by withholding agents is overlooked, it's an important element to consider in the implementation of a withholding VAT system. In particular, when withholdings agents are other state institutions with significant economic or political clout, the penalty parameter is no longer relevant. Non-remittance could be rampant.

4.3 Withholding VAT, excess-credits and intra-firm input allocation

Now, I turn to the effect of withholding VAT on the production decisions of firms which receive prices net of VAT, the withholdees.

Consider a withholdee firm which accumulates credits after delays in the reimbursement of excess-withholdings. The firm's budget is,therefore, expenditure constrained. If credit markets were perfect with the acceptance of withholding certificates as collateral for bank loans, the effect of excess-withholdings and delays in reimbursement would matter less. But with imperfect credit markets [9] and a financial sector unwilling to accept withholding certificates as collateral for credit, then the build-up of excess-withholdings can lead to adjustments in input use and output production [10]. I use the approach from models in the agricultural economics literature to present expenditure-constrained profit functions. Short-run and long-run credit constraints are known to affect farm production(Ciaian et al. 2011).

Assume that the withholdee's technology is characterized by a nonempty, compact and convex set with free disposal of inputs and outputs. Further assume that it operates with two inputs, labor L and capital K. Let w and R be their corresponding input prices, which are strictly positive.

Assuming $P > 0$ as the vector output prices, the withholdee's profit function then writes

$$\pi = PH - w \cdot L - R \cdot K$$

The constraint on expenditures on inputs is modeled through the inclusion of parameters κ_L and $\kappa_K \in \{0,1\}$. When there is no withholding, $\delta = 0$ and $\kappa_L = \kappa_K = 0$. The firm's expenditure function writes

$$E(\delta) = \kappa_L \cdot w \cdot L + \kappa_K \cdot R \cdot K$$

Where the upper bound on expenditures, E, is a function of the withholding rate δ. A reduction in the withholding rate reduces excess-credits and therefore relaxes the expenditure constraints. In formal terms, $E(\delta)$ and $\frac{\partial E}{\partial \delta} < 0$ [11].There are several cases to consider in the analysis of the effect of expenditure constraints on input use. Depending on the level input intensity and the subsequent build-up of excess-credits or the operational cash at its disposal to pay for non-creditable inputs such as labor, the firm may face different scenarios. I first analyze the case where no such constraints exist, either because there are no excess-credits or if the state reimburses them in due time. In general, The firm solves a constrained optimization problem as below

$$\begin{array}{ll} \text{Max} & \pi = PH - w \cdot L - R \cdot K \\ \text{subject to} & \kappa_L \cdot w \cdot L + \kappa_K \cdot R \cdot K \leq E(\delta) \quad (\mu > 0) \end{array}$$

Assuming only interior solutions, the optimum is characterized by

[9] Strong information asymmetries and credit rationing

[10] To isolate the hypotheses on the effect of expenditure constraints on input use and production decisions, we must assume that the level of access to external financing does not change during the period under consideration. This is a strong assumption, which must be considered when interpreting model results.

[11] Of course, we abstract away from other factors, which can also affect the expenditure constraints. Factors specifically related to VAT design include (I) the existence of multiple rates, especially when those applicable to a firm's purchases are higher than those on its sales, (ii) credit formation during an investment phase and (iii) credit formation to a focus on export-oriented operations or (iv) the existence of a substantial share of zero-rated items on a sale.

$$PH_L = (1+\mu\kappa_L)w$$
$$PH_K = (1+\mu\kappa_K)R$$

Now, I present present several cases under different propositions.

Proposition 5. Without a withholding induced expenditure constraint (i.e. $\delta = 0, \kappa_L = \kappa_K = 0$), production decisions only depend on the relative market prices and the relative market value products of inputs. In other words, the equality $\frac{PH_L}{PH_K} = \frac{w}{R}$ characterizes the equilibrium. This result is the same as in standard unconstrained profit maximization problem.

Proof. Without an expenditure constraint, the problem reverts to the standard unconstrained optimization problem for the firm. The interior optimum is characterized by $PH_L = w$ and $PH_K = R$, from which the result follows.

Proposition 6. When withholding leads to the build-up of excess-credits to the point of symmetric expenditure constraints (i.e. $\kappa_L = \kappa_K = 1$), the marginal value products of each input is greater than its per unit input cost ($PH_L > w$ and $PH_K > R$). Also, input use and output decisions are affected by the upper bound on expenditures as well as the relative market prices of labor and capital. Relaxing the upper limit on costs increases both the use of both inputs, as well as the production level. The extent to which either input increases depends on the relative market prices of labor and capital.

Proof. See Appendix 8.

Proposition 7. Under a symmetric expenditure constraint, any increase in the withholding rate leads to a decrease in output. Inversely, a reduction in the withholding rate leads to an increase in production.

Proof. From the proof of Proposition 6, we know that $\frac{\partial H}{\partial E} > 0$. Since the level of expenditure is decreasing in the withholding rate, we also have $\frac{\partial E}{\partial \delta} < 0$. It follows that $\frac{\partial H}{\partial \delta} = \frac{\partial H}{\partial E}\frac{\partial E}{\partial \delta} < 0$.

The case described in Proposition 7 may arise for a capital intensive firm. A transaction with a withholding agent, combined with a delay in the reimbursement of excess-credits, can reduce the company's ability to both acquire new inputs and or increase its labor force.

4.4 Withholding VAT, resource allocation across firms and productivity

Now, I turn to the effect of withholding on the resource allocation across firms. First, assume that the economy has a set of heterogeneous firms in a monopolistic competition market structure.

These production units have different efficiencies and face different distortions on output or input use based on the excess credits induced by the withholding policy. In contrast to the case analysis in the previous subsection, for simplicity, I assume that the Revenue Authority reimburses all excess credits with delays. This extension on the effects of withholding on resource allocation follows the presentation in Hsieh and Klenow 2009. A representative firm produces a single final good H using a Cobb-Douglas technology while using the outputs of I other sectors as its inputs. Hence, we have

$$H = \prod_{i=1}^{I} H_i^{\theta_i}$$

With θ_i the contribution of sector i in the value-added and $\sum_{i=1}^{I}\theta_i = 1$. The sector specific output H_i is itself characterized by CES technology, such that:

$$H_i = [\sum_{j=1}^{J} H_{ij}^{\frac{\sigma-1}{\sigma}}]^{\frac{\sigma}{\sigma-1}}$$

Where σ is the intra-industry elasticity of substitution. The production function of the differentiated product H_{ij} writes

$$H_{ij} = A_{ij} K_{ij}^{1-\alpha_i} L_{ij}^{\alpha_i}$$

Now, I assume that withholding can introduce two distortions on capital and labor, either jointly reflected in the value of output or through their relative marginal products. An output distortion ψ_H changes the marginal products of the two inputs by the same proportion. In the context of withholding VAT, firms with delays in the reimbursement of excess credits face a constraint which hinders their turnover growth, for instance with limitations on the purchase of new inputs or hiring decisions. Thus, ψ_H would be high for them and low for firms which are unaffected. The second distortion, ψ_K, increases the marginal product of capital relative to the marginal product of labor. Again, firms affected by excess withholding have a higher cost of capital. Bank accept their withholding certificates as collateral for credits but with a haircut. Their implied cost of capital and, therefore, ψ_K is higher compared to their peers. With these distortions and if the cost of capital R and the wage w are constant, the objective function of the representative firm in sector i and producing good j with price P_{ij} can be formalized as

$$\pi_{ij} = (1-\psi_{H_{ij}})P_{ij}H_{ij} - wL_{ij} - (1+\psi_{K_{ij}})RK_{ij}$$

Profit maximization yields the following marginal products in 2 and 3. The marginal revenue product of capital $MRPK$ is proportional to the turnover to capital ratio. The marginal revenue product of labor $MRPL$ is proportional to the average turnover per employee. For each production unit, idiosyncratic distortions on capital and labor drive resource allocation decisions which, in turn, lead to differences in marginal revenue products between firms. Firms affected by withholding policy, either through higher output distortions or cost of capital, have higher marginal revenue products.

$$MPRK_{ij} = \frac{R(1+\psi_{K_{ij}})}{(1-\psi_{H_{ij}})} \tag{2}$$

$$MRPL_{ij} = \frac{w}{1-\psi_{H_{ij}}} \tag{3}$$

Again, following Hsieh and Klenow 20009, a firm's revenue productivity is the geometric mean of the factor marginal revenue products and writes:

$$TFPR_{ij} = \frac{P_{ij}H_{ij}}{K_{ij}^{\alpha_i}(wL_{ij})^{1-\alpha_i}} \propto (MPRK_{ij})^{\alpha_s}(MRPL_{ij})^{1-\alpha_i} \propto \frac{(1+\psi_{K_{ij}})^{\alpha_i}}{1-\psi_{H_{ij}}} \tag{4}$$

From 4, it's clear that high marginal revenue products of labor and capital are associated with high $TFPR$. When a firm's excess withholding raises its marginal revenue product of capital, it produces at a sub-optimal level. So, when the removal of the withholding policy reduces excess credits, we would expect the marginal revenue products of firms affected by the reform to decline in the post-reform period. They would, therefore, experience revenue, capital and employment growth. Finally, with assumptions on input shares and on the elasticity of substitution, the distortions and productivity parameters can be estimated as follows.

$$1+\psi_{K_{ij}} = \frac{\alpha_i}{1-\alpha_i}\frac{wL_{ij}}{RK_{ij}}$$

$$1-\psi_{H_{ij}} = \frac{\sigma}{1-\sigma}\frac{wL_{ij}}{(1-\alpha_i)P_{ij}H_{ij}}$$

$$A_{ij} = \frac{(P_{ij}H_{ij})^{\frac{\sigma}{\sigma-1}}}{K_{ij}^{\alpha_i}L_{ij}^{1-\alpha_i}}$$

With these results and assumptions on the elasticity of substitution as well as factor shares, I can measure capital and output distortion before and after the withholding VAT policy reform, as well as infer about its impact on the productivity of firms.

5 Institutional background in Senegal

5.1 Administrative organization

Senegal's domestic tax system is administered by the Directorate General of Taxes and Domains (Direction générale des impôts et des domaines - DGID). DGID collects revenue on all direct and indirect taxes as well as registration rights. DGID has central services and operational services. The latter is divided into 20 tax centers, three of which are specialized tax units. A Large Taxpayer Unit (LTU) is in charge of all taxpayers whose turnover is greater than 1 billion CFA Francs, firms operating in sectors of strategic importance to Senegal's economy, public institutions and businesses in cross-ownership arrangements. A Medium Taxpayer Unit (MTU) covers all businesses whose turnover is between 200 million CFA Francs and 1 billion CFA Francs. A third specialized tax unit administers taxpayers in regulated professions and which require their members to identify themselves through a registration process (lawyers, notaries, pharmacists, etc.). In general, specialized tax units have jurisdiction in Dakar, the administrative and economic capital of Senegal. All other tax centers cover different geographic blocks across the country.

5.2 Value-added Tax (VAT) and withholding VAT

Senegal adopted its first value-added tax in the fiscal year 1980. Since then, the legislation has gone through multiple changes aiming to increase VAT performance with a simplification of the base, reductions of applicable rates and a greater inclusion of wholesalers [12]. Since August 1, 1994, the West African Economic and Monetary Union (WAEMU) treaty set harmonizes indirect tax policy with statutory tax rates and tax bases. Later, the 1998 WAEMU directive set guidelines for VAT policy.Since then, the maximum VAT rate applicable in any member country has to be between 15 and 20%. Senegal chose a VAT rate of 18%, closer to the upper bound and reflecting in part the narrowness of its VAT. A reduced rate of 10% applies to accommodations and food sales in the tourism industry. There is no official threshold for VAT registration. But most of the VAT registered firms are medium-size and large businesses, with the bulk of revenue collected from the latter group.

By the 15^{th} of every month, firms are required to declare all transactions in the preceding month subject to VAT. Firms fill a form which includes information on turnover a given month, sales not subject to VAT, applicable deductions and the amounts withheld at source by withholding agents, excess credits rolled over for the next period and finally the value of refund claims in process. Besides, withholding agents also submit a monthly list of suppliers for which they applied

[12] Sometimes they are loosely defined as the informal sector.

withholding VAT, as well as the amounts withheld for each firm. If VAT operations result in a net credit, the amount can be rolled over to the following month and subtracted like other tax-deductible business expenses. If the firm does not exhaust its excess credits, it may request a quarterly refund and at the latest within two years of the initial transaction which created the credit. Credits which are not reclaimed within two years are lost. Exporting companies, firms executing government contracts, public entities, and enterprises financed with aid or external debt can immediately ask for refunds in the month following the transaction which resulted in an excess credit.

So, it is in this context that a reform came into effect in 2013 to eliminate VAT management complexities . Senegal reformed its VAT system with the gradual removal of withholding . From 2004 to 2012, a withholding rate of 50% of the VAT amount was applied to firms registered at Senegal's LTU while 100% of the VAT component of eligible transactions was withheld at source for all others. With the reform, first, for the LTU firms, the withholding policy was removed on January 1st, 2013. Then, on January 1st, 2015 the policy was canceled for MTU businesses.

6 Data

The data comes from the database tables of DGID, generated with a software called SIGTAS (Standard Integrated Government Tax Administration System).The database provides information on firm characteristics such as economic sector, business activity, location, incorporation type, the relevant tax center (LTU, MTU or other tax centers) and payments made for each tax and in every period. The data covers monthly and annual returns between 2010 and 2015, including 2 years before the reform, took effect for LTU firms and three years for MTU firms. In this paper, I use the following tables:

VAT Returns I use repeated cross-sections (monthly) of VAT returns before and after the policy change. This dataset also includes withholding VAT for every reporting period. The main outcomes from this table are the payment and declaration (filling of returns) of VAT. This data table also includes standard information such as monthly turnover, exempt-sales, excess-credits and input deductions.

Income Tax. Firms also file annual income tax returns, which they submit along with their financial statements. The annual income tax returns provide information about on annual turnover, the accounting profit of the firm, its taxable profit, physical capital formation, fixed and variables costs.

PAYE. Every month, taxpayers also submit a form with the personal income tax withheld at source for each employee (Pay-As-You-Earn - PAYE data). These forms provide information on the wage bill and labor force of the firm. I use the PAYE data for information about the firm's wage bill, which is used in estimating the effect of withholding on productivity.

I combine all these sources of information into a single dataset used for the analysis. Table 2 displays summary statistics on the characteristics of firms registered at the LTU or the MTU. The majority of LTU firms are limited-liability corporations or partnerships with 95.69% of them categorized as legal persons. They are also more likely to be in manufacturing and have higher gross profit margins. On the other hand, MTU firms are more likely to be sole proprietorships with 12.15% of businesses in this category, compared to 3.05% for LTU firms.Tables 4 and 5 present the descriptive statistics on the VAT returns of LTU and MTU firms for the $2010-2015$ period.

7 Preliminary results on the effect of withholding reform on compliance and productivity

To assess firms' compliance behavior after the removal of withholding VAT, I exploit the fact that the reform creates a quasi-experiment, setting itself for a difference-in-differences estimation. Table 2 displays the evolution of the withholding VAT policy parameters over time. The first wave of the reform occurs on Jan 01, 2013 for LTU firms, followed by a second change for MTU firms on Jan 01, 2015. The policy remains in place for all others. Using the staggered roll-out of the reform, I can define treatment and control groups in both periods. At each point, all the firms which are affected by the change are defined as treatment groups while the all other unaffected firms constitute a control group. The identification relies on a common trend assumption between firms in treatment and control groups. To test this hypothesis, I check whether the trends in the primary outcome variables are similar across the two groups during the years preceding the reform. For instance,Figure 7 plots the average share of firms which made positive VAT payments in the treatment and control groups for the LTU wave of the policy change. The two lines move in tandem, lending strong support to the validity of the parallel trends assumption. To estimate the effect of the reform on compliance, I run the following difference-in-differences specifications.

$$y_{it} = \alpha_i + \gamma_t + \mu_i \cdot t + \beta_1 Treat_i \cdot Post_t + \beta_2 Treat_i \cdot W_i + \beta_3 W_i \cdot Post_t + \delta Treat_i \cdot Post_t \cdot W_i + \varepsilon_{it} \qquad (5)$$

Where y_{it} is firm i's outcome in period i, $Treat_i$ is its treatment status, $Post_t$ denotes the post-reform period. The specification also includes an interaction term between the treatment status (=1 if affected by the reform) and an indicator variable for firms with withholding-related transactions in the two years preceding the reform. That is $W_i = 1\left(\sum W_{it\,t=-24}^{\;t=0} > 0\right)$. Finally, α_i and γ_t are firm and month fixed effects, while μ_i is a linear time-trend and ε_{it} the error term. The treatment effect for firms affected which were affected by withholding VAT before the reform is is $\beta_1 + \delta$. The treatment effect fo firms which had no positive withholding VAT is β_1.

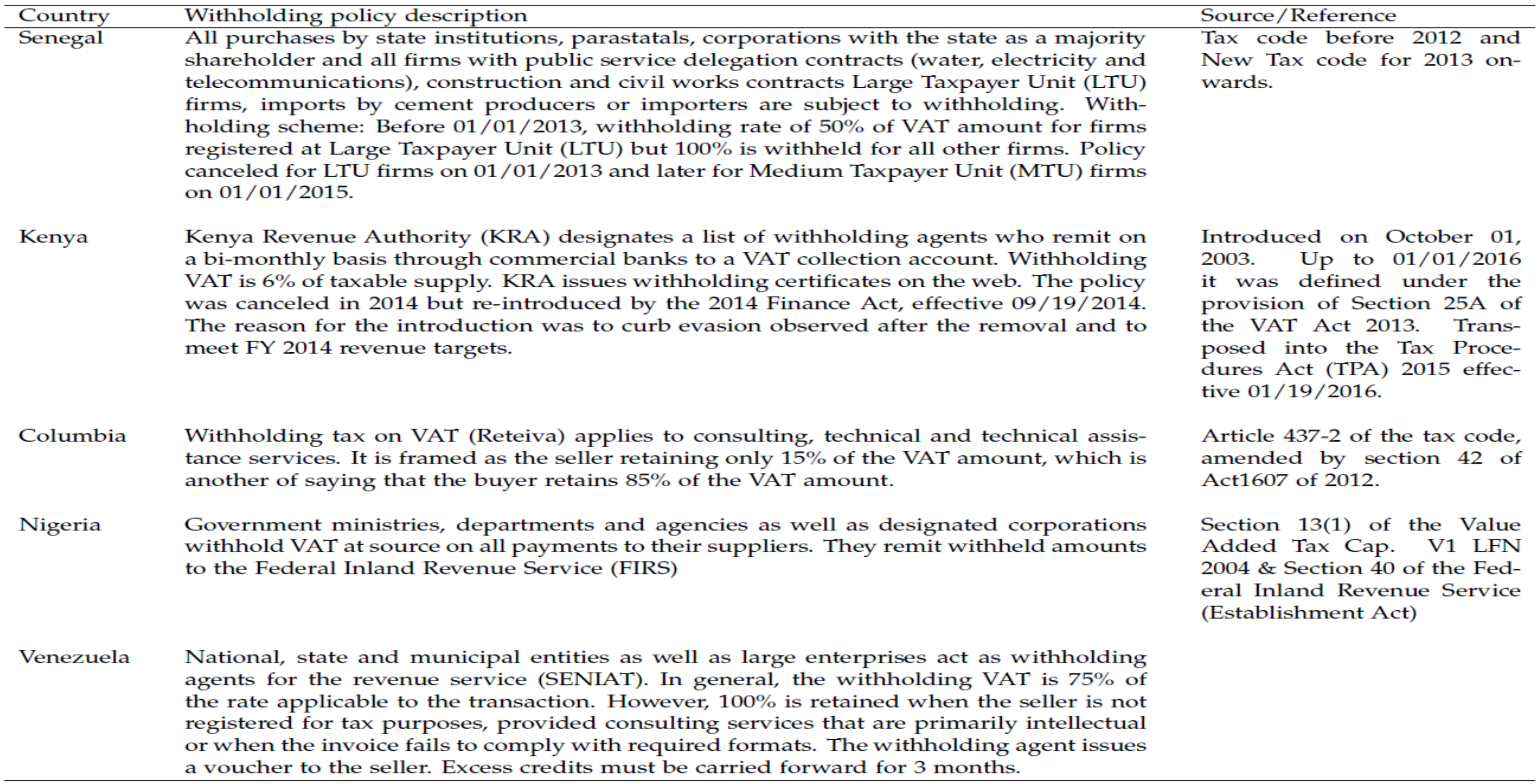
Tables 4-6 report very preliminary results of these two specifications in linear probability models for the payment and reporting margins of LTU firms. In model (1) of Tables 4 and 5, I find that the reform has a negative and significant effect $(p < 0.05)$ on both the extensive and intensieve margin of VAT payment for LTU firms which experienced withholding VAT before, lending support to the argument that the termination of withholding leads to greater evasion. This result runs counter to our theoretical model which posited that the removal of withholding would have no effect on LTU firms because of higher audit probability. However, to make a definite conclusion on this, we would need to confirm those audit probabilities for LTU firms are indeed high. In case they are not, the result would not be surprising.

Models (1) and (2) of Tables 4 and 5 shows that after the reform has no adverse effect on the behavior of MTU firms which previously had positive withholding VAT. The result holds for both the intensive and extensive margins of VAT payment. Consistent with this result, I also find no effect on the filling behavior of these firms.

Table 6 provides preliminary results on the impact of withholding policy termination on a very basic measure firm productivity. The outcome variable is the log of output per employee. I find that, for firms which previously experienced withholding, the reform had no effect on this basic measure of productivity. This result holds for medium-size businesses as well. In future versions of this paper, I will focus on more elaborate measures of productivity and their relationship with withholding VAT as presented in subsection 4.4. Updates will also include an empirical section on the behavior of the withholding agents, as well as robustness checks on regression results.

Tables

Table 1: Stylized facts on withholding VAT in selected African and Latin American Countries



Country	Withholding policy description	Source/Reference
Senegal	All purchases by state institutions, parastatals, corporations with the state as a majority shareholder and all firms with public service delegation contracts (water, electricity and telecommunications), construction and civil works contracts Large Taxpayer Unit (LTU) firms, imports by cement producers or importers are subject to withholding. Withholding scheme: Before 01/01/2013, withholding rate of 50% of VAT amount for firms registered at Large Taxpayer Unit (LTU) but 100% is withheld for all other firms. Policy canceled for LTU firms on 01/01/2013 and later for Medium Taxpayer Unit (MTU) firms on 01/01/2015.	Tax code before 2012 and New Tax code for 2013 onwards.
Kenya	Kenya Revenue Authority (KRA) designates a list of withholding agents who remit on a bi-monthly basis through commercial banks to a VAT collection account. Withholding VAT is 6% of taxable supply. KRA issues withholding certificates on the web. The policy was canceled in 2014 but re-introduced by the 2014 Finance Act, effective 09/19/2014. The reason for the introduction was to curb evasion observed after the removal and to meet FY 2014 revenue targets.	Introduced on October 01, 2003. Up to 01/01/2016 it was defined under the provision of Section 25A of the VAT Act 2013. Transposed into the Tax Procedures Act (TPA) 2015 effective 01/19/2016.
Columbia	Withholding tax on VAT (Reteiva) applies to consulting, technical and technical assistance services. It is framed as the seller retaining only 15% of the VAT amount, which is another of saying that the buyer retains 85% of the VAT amount.	Article 437-2 of the tax code, amended by section 42 of Act1607 of 2012.
Nigeria	Government ministries, departments and agencies as well as designated corporations withhold VAT at source on all payments to their suppliers. They remit withheld amounts to the Federal Inland Revenue Service (FIRS)	Section 13(1) of the Value Added Tax Cap. V1 LFN 2004 & Section 40 of the Federal Inland Revenue Service (Establishment Act)
Venezuela	National, state and municipal entities as well as large enterprises act as withholding agents for the revenue service (SENIAT). In general, the withholding VAT is 75% of the rate applicable to the transaction. However, 100% is retained when the seller is not registered for tax purposes, provided consulting services that are primarily intellectual or when the invoice fails to comply with required formats. The withholding agent issues a voucher to the seller. Excess credits must be carried forward for 3 months.	

Sources: Tax code provisions and press coverage on withholdings in listed countries

Table 2: Withholding policy over time

	Before 01/01/2013	After 01/01/2013	After 01/01/2015
Large Taxpayers	50%	0%	0%
Medium Taxpayers	100%	100%	0%
Small Firms	100%	100%	100%

Note: This table summarizes the evolution of Senegal's withholding VAT policy over time, as well as its applicable parameters. Firms are segmented into three sub-populations. Large Taxpayers are firms with turnover greater than 1 billion CFA Francs, firms in sectors deemed strategic to the economy (telecom, mining), public institutions and firms in cross ownership arrangements. Before the reform, these firms were subject to a withholding rate of 50%, meaning that only half of the VAT was retained whenever they transacted with withholding agents. Medium Taxpayers are businesses with turnover in $[0.2, 1)$ billion CFA Francs. Medium Taxpayers and all other firms were subject to a withholding rate of 100% up to January 2015, when the medium taxpayers were no longer subject to withholding. A 100% withholding rate continues to apply on the sales of small firms to withholding agents.

Impact on the extensive Margin of VAT payment

	(1)	(2)
	LTU	MTU
	=1 if VAT Paid>0	=1 if VAT Paid>0
Treatment*Post	-0.0408***	0.0401***
	(0.0137)	(0.0130)
Treatment*Withholding VAT	-0.0162	0.0391*
	(0.111)	(0.0222)
Withholding VAT*Post	0.0283***	0.0113**
	(0.00725)	(0.00483)
Treatment*Withholding VAT *Post	-0.0391**	-0.0400**
	(0.0182)	(0.0160)
Constant	0.787***	0.666***
	(0.0231)	(0.00200)
Month FE	Yes	Yes
Firm FE	Yes	Yes
Observations	116,540	648,862
Number of firms	2,495	21,218
R-squared	0.007	0.001

Notes: This table displays the results on the extensive margin of VAT payment for LTU and MTU firms. Both (1) and (2) reflect the difference-in-difference specification in (5). For example, (1) shows the results for large taxpayer unit firms, which were the treatment group in the first wave of the reform. It Displays the treatment effect of the reform on the intensive margin of VAT payments. The treatment effect on firms which had positive withholding VAT before the reform is ($\beta_1 + \delta$) = (-0.450)+(-0.947)=-1.397. This treatment effect is negative and significant at the 5% level. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Impact on Intensive Margin of VAT payment

	(1)	(2)
	LTU	MTU
	Log (VAT Paid)	Log (VAT Paid)
Treatment*Post	-0.450**	0.778***
	(0.219)	(0.207)
Treatment*Withholding VAT	-0.647	1.036***
	(1.688)	(0.371)
Withholding VAT*Post	0.697***	0.329***
	(0.124)	(0.0648)
Treatment*Withholding VAT *Post	-0.947***	-0.772***
	(0.304)	(0.254)
Constant	12.01***	8.142***
	(0.353)	(0.0305)
Month FE	Yes	Yes
Firm FE	Yes	Yes
Observations	115,600	639,069
Number of firms	2,494	21,205
R-squared	0.008	0.002

Notes: This table displays the results on the intensive margin of VAT payment for LTU and MTU firms. Both (1) and (2) reflect the difference-in-difference specification in (5). Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Impact on the declaration margin

	(1)	(2)
	LTU	MTU
	=1 if VAT Declared	=1 if VAT Declared
Treatment*Post	0.0196*	-0.00142
	(0.0108)	(0.0116)
Treatment*Withholding VAT	0.0986**	0.0484***
	(0.0501)	(0.0162)
Withholding VAT*Post	-0.0125*	-0.0107***
	(0.00666)	(0.00352)
Treatment*Withholding VAT *Post	-0.0228	-0.00378
	(0.0144)	(0.0137)
Constant	0.807***	0.648***
	(0.0108)	(0.00172)
Month FE	Yes	Yes
Firm FE	Yes	Yes
Observations	116,540	648,862
Number of firms	2,495	21,218
R-squared	0.004	0.002

Notes: This table displays the results on the declaration margin of VAT for LTU and MTU firms. Both (1) and (2) reflect the difference-in-difference specification in (5). Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Impact on turnover per employee

	LTU	MTU
	Log (Turnover/Employee)	Log (Turnover/Employee)
Treatment*Post	0.417**	0.325*
	(0.180)	(0.189)
Treatment*Withholding VAT	1.708*	0.670**
	(1.027)	(0.286)
Withholding VAT*Post	-0.0910	0.0686
	(0.107)	(0.0510)
Treatment*Withholding VAT *Post	-0.235	0.0313
	(0.240)	(0.223)
Constant	12.24***	8.895***
	(0.218)	(0.0273)
Month FE	Yes	Yes
Firm FE	Yes	Yes
Observations	116,540	648,862
Number of firms	2,495	21,218
R-squared	0.011	0.003

Notes: This table displays the results on the treatment effects of the withholding reform for LTU and MTU firms. Both (1) and (2) reflect the difference-in-difference specification in (5). Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: VAT Return Summary Statistics - Large Taxpayer Unit, 2010-2015

	2010	2011	2012	2013	2014	2015	Average
Total Turnover	340.98	350.66	358.44	363.11	371.76	372.34	360.24
	(405.29)	(412.06)	(410.72)	(420.11)	(426.45)	(439.60)	(419.94)
	[169.75]	[177.02]	[182.96]	[182.19]	[188.08]	[177.81]	[179.67]
Exempt sales	109.91	121.13	124.59	123.11	126.64	112.24	119.78
	(175.68)	(183.66)	(185.62)	(188.70)	(192.09)	(187.31)	(185.98)
	[15.42]	[19.52]	[22.01]	[16.52]	[17.16]	[3.17]	[15.33]
Taxable Amount	175.99	183.13	171.57	175.96	180.14	175.14	176.95
	(235.76)	(240.20)	(234.69)	(240.60)	(245.18)	(245.86)	(240.63)
	[68.40]	[74.26]	[60.36]	[58.89]	[57.59]	[45.68]	[60.74]
Gross VAT	10.08	27.62	26.45	27.25	27.74	26.77	24.69
	(23.85)	(34.84)	(34.28)	(34.92)	(35.60)	(35.66)	(34.21)
	[0.00]	[11.31]	[9.86]	[10.05]	[9.44]	[7.49]	[6.24]
VAT Withheld	2.45	2.34	2.40	.	.	.	2.39
	(2.01)	(2.00)	(2.04)	(.)	(.)	(.)	(2.02)
	[2.05]	[1.68]	[1.85]	[.]	[.]	[.]	[1.88]
Previous month VAT credits	23.60	21.51	25.39	18.80	23.80	22.43	22.56
	(42.33)	(40.82)	(43.89)	(39.40)	(43.07)	(42.57)	(42.09)
	[0.18]	[0.00]	[0.22]	[0.00]	[0.00]	[0.00]	[0.00]
Total of deductions	75.19	73.58	76.33	76.28	74.03	66.85	73.61
	(103.92)	(103.41)	(104.26)	(107.57)	(107.88)	(103.46)	(105.19)
	[25.90]	[23.46]	[25.13]	[22.16]	[18.34]	[13.42]	[20.86]
Rolled-over credit	59.88	94.82	97.60	67.52	67.99	56.79	74.03
	(125.72)	(169.58)	(170.62)	(134.81)	(137.37)	(125.76)	(146.06)
	[2.07]	[2.34]	[3.18]	[2.38]	[2.22]	[0.51]	[1.97]
Number of firms	784	804	817	834	831	872	
Observations	7794	8650	9102	9324	9289	9689	

Notes: For each fiscal year, this table reports the means, standard deviation (in parentheses) and median (in brackets) of VAT return line items for LTU registered firms. All line items are winosrized at the top and bottom 1% All figures are expressed in millions of CFA Francs.

Table 5: VAT Return Summary Statistics - Medium Taxpayer Unit, 2010-2015

	2010	2011	2012	2013	2014	2015	Average
Total Turnover	37.71	39.18	44.41	48.67	51.45	52.92	46.03
	(66.86)	(66.68)	(82.14)	(89.40)	(96.93)	(100.09)	(85.56)
	[22.09]	[24.51]	[25.38]	[26.19]	[26.01]	[25.51]	[24.92]
Exempt sales	11.81	13.40	15.19	15.62	16.50	15.68	14.79
	(39.39)	(40.59)	(46.86)	(50.55)	(53.83)	(54.55)	(48.34)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]
Taxable Amount	24.65	25.53	26.42	29.29	29.13	31.62	27.93
	(49.90)	(48.25)	(50.99)	(55.83)	(58.62)	(63.08)	(55.03)
	[11.49]	[12.89]	[12.35]	[12.85]	[11.15]	[12.27]	[12.19]
Gross VAT	2.59	4.20	4.47	5.07	5.08	5.53	4.56
	(9.08)	(7.51)	(8.56)	(9.39)	(9.91)	(10.73)	(9.33)
	[0.00]	[2.00]	[1.95]	[2.19]	[1.86]	[2.09]	[1.56]
VAT Withheld	1.62	1.73	1.77	1.75	1.85	.	1.74
	(1.69)	(1.77)	(1.80)	(1.78)	(1.80)	(.)	(1.77)
	[0.87]	[0.95]	[0.94]	[0.91]	[1.07]	[.]	[0.94]
Previous month VAT credits	8.32	7.45	8.48	7.91	10.37	9.51	8.69
	(20.23)	(18.80)	(20.90)	(21.61)	(24.61)	(23.83)	(21.84)
	[0.51]	[0.12]	[0.22]	[0.00]	[0.35]	[0.13]	[0.16]
Total of deductions	15.24	14.29	16.29	17.99	20.44	20.72	17.60
	(35.59)	(31.60)	(35.07)	(38.28)	(43.16)	(45.48)	(38.78)
	[4.71]	[4.62]	[4.99]	[4.73]	[4.81]	[4.74]	[4.76]
Rolled-over credit	18.77	48.92	37.01	15.88	18.95	18.98	26.54
	(63.18)	(124.12)	(104.86)	(41.07)	(49.77)	(51.36)	(79.79)
	[1.36]	[1.82]	[1.52]	[1.60]	[1.73]	[1.48]	[1.58]
Number of firms	1,101	1,139	1,127	1,102	1,062	1,219	
Observations	10,350	12,269	12,594	12,305	11,933	13,852	

Notes: For each fiscal year, this table reports the means, standard deviation (in parentheses) and median (in brackets) of VAT return line items for MTU registered firms. All line items are winosrized at the top and bottom 1%. All figures are expressed in millions of CFA Francs.

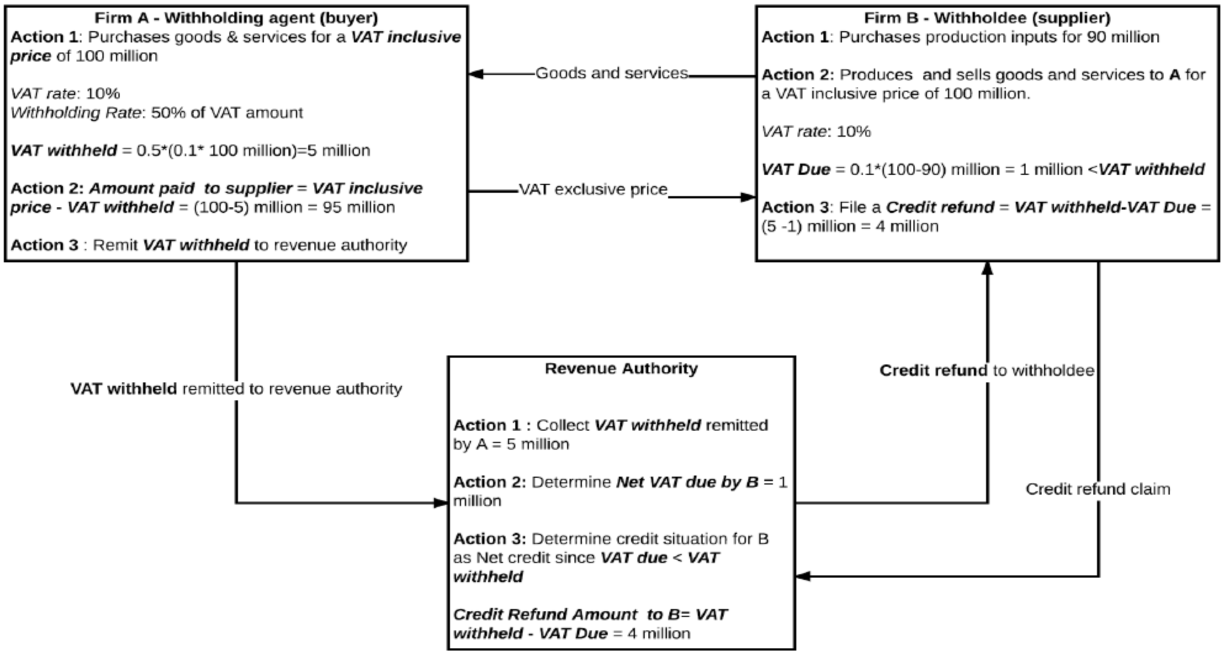

Firm A - Withholding agent (buyer)
Action 1: Purchases goods & services for a VAT inclusive price of 100 million
VAT rate: 10%
Withholding Rate: 50% of VAT amount
VAT withheld = 0.5*(0.1* 100 million)=5 million
Action 2: Amount paid to supplier = VAT inclusive price - VAT withheld = (100-5) million = 95 million
Action 3 : Remit VAT withheld to revenue authority
Goods and services
VAT exclusive price
Firm B - Withholdee (supplier)
Action 1: Purchases production inputs for 90 million
Action 2: Produces and sells goods and services to A for a VAT inclusive price of 100 million.
VAT rate: 10%
VAT Due = 0.1*(100-90) million = 1 million <VAT withheld
Action 3: File a Credit refund = VAT withheld-VAT Due = (5 -1) million = 4 million
VAT withheld remitted to revenue authority
Revenue Authority
Action 1 : Collect VAT withheld remitted by A = 5 million
Action 2: Determine Net VAT due by B = 1 million
Action 3: Determine credit situation for B as Net credit since VAT due < VAT withheld
Credit Refund Amount to B= VAT withheld - VAT Due = 4 million
Credit refund to withholdee
Credit refund claim

Figure 1: Illustration of VAT Withholding resulting in a Credit Refund

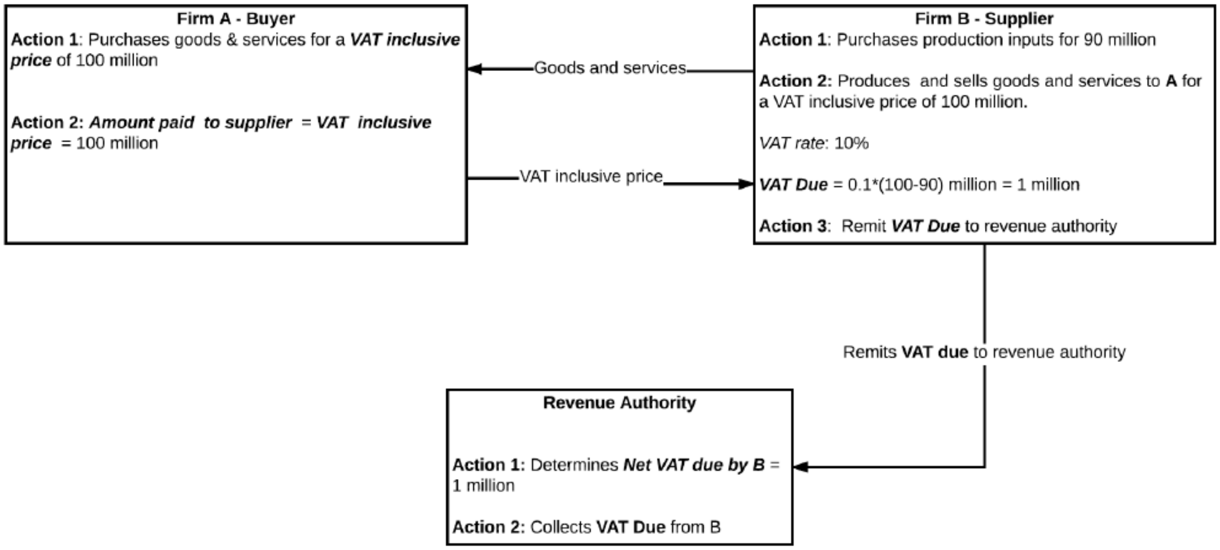

Firm A - Buyer
Action 1: Purchases goods & services for a VAT inclusive price of 100 million
Action 2: Amount paid to supplier = VAT inclusive price = 100 million
Goods and services
VAT inclusive price
Firm B - Supplier
Action 1: Purchases production inputs for 90 million
Action 2: Produces and sells goods and services to A for a VAT inclusive price of 100 million.
VAT rate: 10%
VAT Due = 0.1*(100-90) million = 1 million
Action 3: Remit VAT Due to revenue authority
Remits VAT due to revenue authority
Revenue Authority
Action 1: Determines Net VAT due by B = 1 million
Action 2: Collects VAT Due from B

Figure 2: Illustration of a transaction without a withholding policy

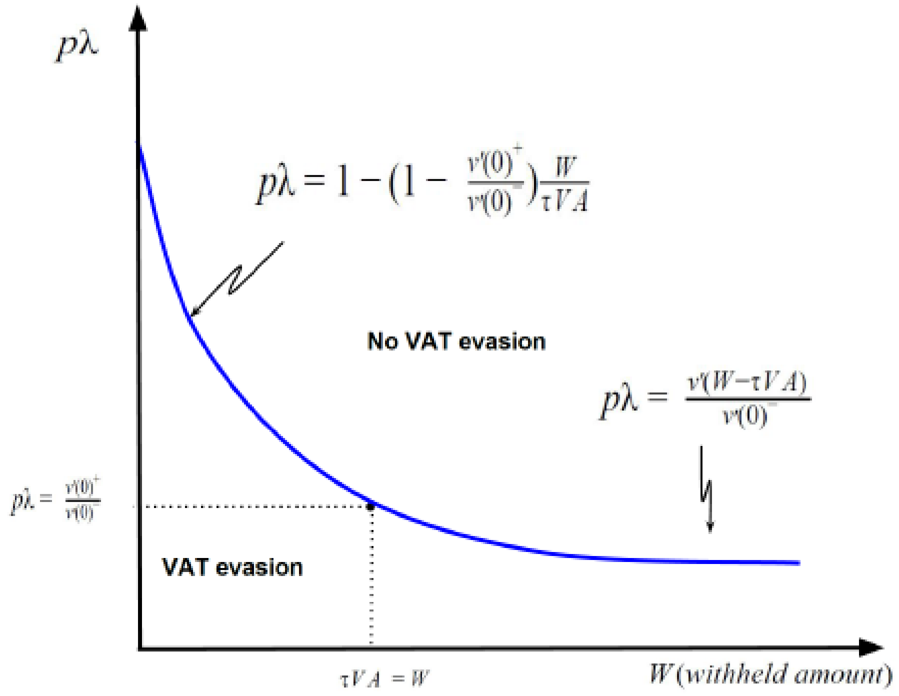

$p\lambda$
$p\lambda = 1-(1-\frac{v'(0)^{+}}{v'(0)^{-}})\frac{W}{\tau VA}$
No VAT evasion
$p\lambda = \frac{v'(W-\tau VA)}{v'(0)^{-}}$
$p\lambda = \frac{v'(0)^{+}}{v'(0)^{-}}$
VAT evasion
$\tau VA = W$
W (withheld amount)

Figure 3: Graphical illustration of evasion decisions

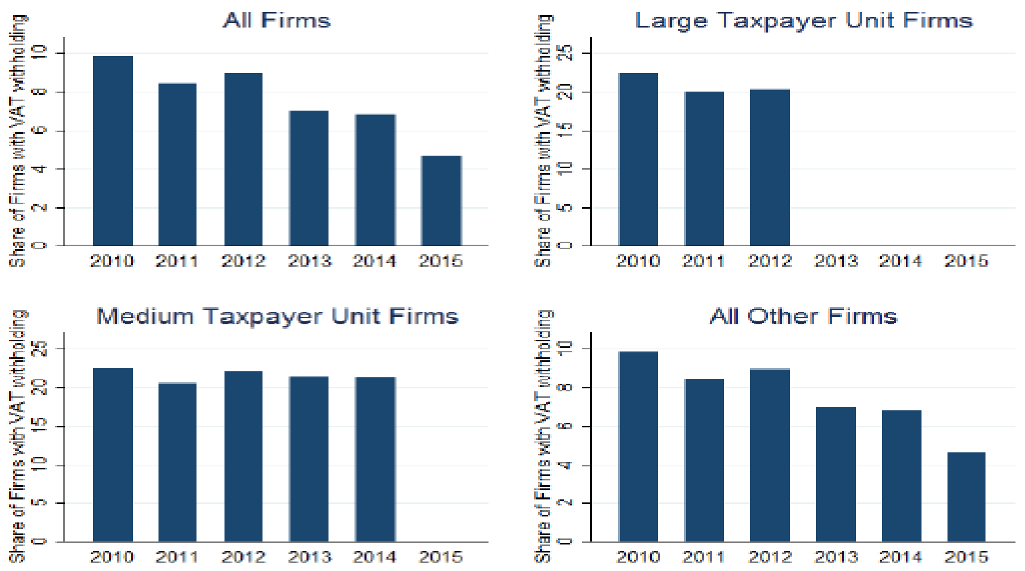


Figure 4: Share of Firms with Positive withholdings

[Graphs on parallel trends here]

References

Allingham, M. and Sandmo, A. (1972). Income tax evasion: A theoretical analysis. Journal of public economics, 1(3/4):323–338.

Asker, J., Collard-Wexler, A., and De Loecker, J. (2014). Dynamic inputs and resource (mis) allocation. Journal of Political Economy, 122(5):1013–1063.

Best, M. C. (2014). The role of firms in workers' earnings responses to taxes: Evidence from Pakistan. Unpub. paper, LSE.

Best, M. C., Brockmeyer, A., Kleven, H. J., Spinnewijn, J., and Waseem, M. (2015). Production versus revenue efficiency with limited tax capacity: Theory and evidence from Pakistan. Journal of Political Economy, 123(6).

Brockmeyer, A. and Hernandez, M. (2017). Taxation, information, and withholding: evidence from Costa Rica. World Bank Policy Research Working Paper, (7600).

Brockmeyer, A., Kettle, S., and Smith, S. D. (2017). Casting the tax net wider: experimental evidence from Costa Rica.

Carrillo, P., Pomeranz, D., and Singhal, M. (2016). Dodging the taxman: Firm misreporting and limits to tax enforcement. Technical report, National Bureau of Economic Research.

Carrillo, P. E., Shahe Emran, M., and Anita, R. (2012). Do cheaters bunch together? Profit taxes, withholding rates and tax evasion.

Chambas, G. (2014). Securiser les remboursements de credits de TVA dans les pays membres de l'uemoa. Document de Travail-CERDI.

Ciaian, P., Fałkowski, J., Kancs, d., and Pokrivcak, J. (2011). Productivity and Credit Constraints: Firm-Level Evidence from Propensity Score Matching. Factor Markets Working Paper No. 3, September 2011.

Dharmapala, D., Slemrod, J., and Wilson, J. D. (2011). Tax policy and the missing middle: Optimal tax remittance with firm-level administrative costs. Journal of Public Economics, 95(9):1036–1047.

Diamond, P. A. and Mirrlees, J. A. (1971). Optimal taxation and public production i: Production efficiency. The American Economic Review, pages 8–27.

Dusek, L. and Bagchi, S. (2016). Are efficient taxes responsible for big government? Evidence from tax withholding. Evidence from Tax Withholding (December 14, 2016).

Ebrill, L., Keen, M., Summers, V., and Bodin, G. (2001). The modern vat. The International Monetary Fund Publication.

Emran, M. S. and Stiglitz, J. E. (2005). On selective indirect tax reform in developing countries. Journal of Public Economics, 89(4):599–623.

International Monetary Fund (IMF). 2017. Fiscal Monitor: Achieving More with Less. Washington, April.

Keen, M. (2008). Vat, tariffs, and withholding: Border taxes and informality in developing countries. Journal of Public Economics, 92(10):1892–1906.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., and Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. Econometrica, 79(3):651–692.

Kleven, H. J., Kreiner, C. T., and Saez, E. (2015). Why can modern governments tax so much? an agency model of firms as fiscal intermediaries. Technical report, National Bureau of Economic Research.

Kleven, H. J., Kreiner, C. T., and Saez, E. (2016). Why can modern governments tax so much? An agency model of firms as fiscal intermediaries. Economica, 83(330):219–246.

Kleven, H. J. and Waseem, M. (2013). Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan. The Quarterly Journal of 45 Economics, 128(2):669–723.

Kopczuk, W., Marion, J., Muehlegger, E., and Slemrod, J. (2015). Does evasion break the law of tax incidence? Point of collection and the pass-through of state diesel taxes.

Kopczuk, W., Marion, J., Muehlegger, E., & Slemrod, J. (2016). Does tax-collection invariance hold? evasion and the pass-through of state diesel taxes. American Economic Journal: Economic Policy, 8(2), 251-86.

Kopczuk, W. and Slemrod, J. (2006). Putting firms into optimal tax theory. The American economic review, pages 130–134. Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica, 71(6):1695–1725.

Kuchumova, Y. P. (2017). The Optimal Deterrence of Tax Evasion: The Trade-off Between Information Reporting and Audits. Journal of Public Economics, 145, 162-180.

Midrigan, V. and Xu, D. Y. (2014). Finance and misallocation: Evidence from plant-level data. The American Economic Review, 104(2):422–458. Naritomi, J. (2016). Consumers as tax auditors. WP, LSE.

Pomeranz, D. (2015). No taxation without information: Deterrence and self-enforcement in the value-added tax. American Economic Review, 105(8).

Restuccia, D. and Rogerson, R. (2008). Policy distortions and aggregate productivity with heterogeneous establishments. Review of Economic dynamics, 11(4):707–720.

Restuccia, D. and Rogerson, R. (2017). The causes and costs of misallocation. Technical report, National Bureau of Economic Research.

Slemrod, J. (2008). Does it matter who writes the check to the government? The economics of tax remittance. National Tax Journal, 61:251–275.

Slemrod, J., Collins, B., Hoopes, J., Reck, D., and Sebastiani, M. (2015). Does credit card information reporting improve small-business tax compliance? Technical report, National Bureau of Economic Research.

Slemrod, J. and Gillitzer, C. (2013). Tax systems. MIT Press. Waseem, M. (2015). Taxes, informality and income shifting: Evidence from a recent Pakistani tax reform. WP.

Yaniv, G. (1988). Withholding and non-withheld tax evasion. Journal of Public Economics, 35(2):183–204.

Yaniv, G. (1999). Tax compliance and advance tax payments: A prospect theory analysis. National Tax Journal, pages 753–764.

Yitzhaki, S. (1979). A note on optimal taxation and administrative costs. The American Economic Review, 69(3):475–4

Appendix A: Proofs

Proof. The proof uses 3 elements, namely the two results from the FOCs and the expenditure constraint. First recall that with $\kappa_L = \kappa_K = 1$, we have

$$E = \kappa_L wL - \kappa_K RK \tag{1}$$

Differentiating the constraint gives:

$$\frac{dK}{dE} = \frac{1}{R} - \frac{w}{R}\frac{dL}{dE} \tag{2}$$

The FOCs are:

$$PH_L = (1+\mu)w \text{ (3)} \qquad\qquad PH_K = (1+\mu)R \text{ (4)}$$

From 3 and 4, we have the following result

$$\frac{PH_L}{w} = \frac{PH_K}{R}$$

Totally differentiating this result gives:

$$\frac{P}{w}[H_{LL}\frac{dL}{dE} + H_{LK}\frac{dL}{dE}] = \frac{P}{R}[H_{KL}\frac{dL}{dE} + H_{KK}\frac{dK}{dE}] \tag{5}$$

Replacing 2 into 5 yields:

$$\frac{dL}{dE}[H_{LL} - 2H_{LK}\frac{w}{R} + \frac{w}{R}\frac{w}{R}H_{KK}] = \frac{w}{R}\frac{H_{KK}}{R} - \frac{H_{LK}}{R}$$

Re-arranging this expression gives the first comparative static with respect to labor L

$$\frac{dL}{dE} = \frac{\frac{1}{w}(H_{KK} - H_{LK}\frac{R}{w})}{(H_L\frac{R^2}{w^2} - 2H_{LK}\frac{R}{w} + H_{KK})} > 0$$

Following the same steps, we arrive at the comparative static with respect to capital

$$\frac{dK}{dE} = \frac{\frac{1}{w}(H_{LL}\frac{R}{w} - H_{KL})}{(H_L\frac{R^2}{w^2} - 2H_{LK}\frac{R}{w} + H_{KK})} > 0$$

For the effect on output, from the chain rule, we have

$$\frac{dH}{dE} = \frac{\partial H}{\partial L}\frac{dL}{dE} + \frac{\partial H}{\partial K}\frac{dK}{dE} = H_L\frac{dL}{dE} + H_K\frac{dK}{dE}$$

Replacing the results for $\frac{dL}{dE}$ and $\frac{dK}{dE}$ into this expression yields

$$\frac{dH}{dE}=\frac{H_L(H_{KK}-H_{LK}\frac{\partial R}{\partial w})+H_K(H_{LL}\frac{\partial R}{\partial w}-H_{KL})}{w(H_L\frac{R^2}{w^2}-2H_{LK}\frac{R}{w}+H_{KK})}>0$$