VALUE ADDED TAX EVASION, AUDITING AND TRANSACTIONS MATCHING*

ARINDAM DAS-GUPTA
NANYANG TECHNOLOGICAL UNIVERSITY

IRA N. GANG
RUTGERS UNIVERSITY

JEL Codes: H26, H25
Keywords: value added tax, tax evasion

(Revised) JUNE, 1996

: The first draft of this paper was completed while Das-Gupta was a Visiting Fulbright Scholar at Rutgers University. He gratefully acknowledges support from the Fulbright Scholarship and the hospitality of Rutgers University. We also wish to thank Chung Gool Moon and Chuck Romeo for helpful discussion.

Address for correspondence:

Ira N. Gang
Department of Economics, Rutgers University
New Brunswick, NJ 08903-5055, USA
e-mail: gang@rci.rutgers.edu
phone: 908-932-7405
fax : 908-932-7416

AUDITING, TRANSACTIONS MATCHING AND VALUE ADDED TAX EVASION

ABSTRACT

This paper extends the standard theoretical model of tax enforcement by allowing for the cross-matching of transactions in addition to the auditing of taxpayers. For the Value Added Tax (VAT) the matching of purchase and sales invoices is an important enforcement technique. The paper examines the impact such enforcement on the revenue effectiveness and efficiency consequences of the VAT. Transactions matching is shown to have very different effects from auditing: Even when auditing alone is unable to induce non-zero taxpayer reports, and regardless of the expected success rate in auditing of the tax administration, sufficiently intensive cross-matching can induce truthful reporting. On the other hand, matching leads to distorted purchase and sales transactions. It can also distort input use and output decisions even if auditing alone has no adverse effects. In the model, conditions under which the VAT leaves input prices undistorted are found and the content of the often made claim, that a VAT is self-enforcing, is explored. The ability of the tax administration to enforce compliance with the VAT is shown to be sensitive to the knowledge that the tax administration has about the production technology.

AUDITING, TRANSACTIONS MATCHING AND VALUE ADDED TAX EVASION

I. Introduction

Theoretical papers on tax evasion, beginning with Allingham and Sandmo (1972), have focused on only one aspect of the technology of tax enforcement, namely auditing.[1] Auditing, in these papers, is an activity which leads to the discovery of (all or part of) under-reporting of the tax base of the audited taxpayer by the tax authority. An important implicit assumption, which may be termed *the independent audit assumption*, is that auditing of one taxpayer does not systematically throw up information which can be used to detect evasion by other taxpayers. In practice, this assumption may be justified in many situations.[2] However, in important modern contexts, such as in the enforcement of the income-tax and the value added tax (VAT), the assumption is unrealistic. The underlying transactions structure of income creation (or value addition) implies that auditing of one taxpayer throws up important information on other taxpayers. This information gain arises essentially from matching the receipts (for example from sales) with the expenditures (for example on purchases) of different taxpayers.

Conclusions drawn from models which neglect this interdependence of information run the risk of drawing misleading conclusions with regard to important aspects of enforcement such as its effectiveness and the efficiency or equity implications of different enforcement strategies. *The major objective of this paper is to extend the standard treatment of tax enforcement through independent audits by including transactions matching, an enforcement activity which systematically throws up information of use in examining other taxpayers*. The extended model is then used to examine tax enforcement effectiveness and its efficiency implications.

Incorporating this extension into an abstract model will lead to less insight than a relatively

---

[1] For reviews see Cowell (1990) and the special supplement to *Public Finance/Finances Publiques*, 1994.

[2] For example, enforcement of property taxes, the retail sales tax, land taxes and import duties.

concrete situation in which the information structure can be more carefully specified than is usually done. We therefore focus on the VAT, since the opportunity for tax evasion under a VAT is thought to be crucially affected by transactions matching.[3]

VAT evasion is also a topic of independent interest since in recent decades it has replaced other forms of sales taxes in many countries of the world and continues to attract new converts. This popularity is partly because, in comparison with other sales taxes, a VAT is thought to have two important advantages.[4] First, intermediate goods are supposed to bear no net tax. For example, under the widely used invoice method of administering the VAT, tax paid by intermediate goods producers is rebated to final goods producers against purchase invoices leading to zero net taxation of intermediate goods. As a result, marginal conditions for production efficiency are undistorted by the VAT. As is well known since the work of Diamond and Mirlees (1971), zero taxation of intermediate inputs forms part of an optimal commodity tax mix under very general conditions in a second-best world. That a consumption type VAT, along with a retail sales tax, leads to zero taxation of intermediate goods in a second best world is an important advantage claimed for it over other forms of the sales tax. Does this advantage survive in a third best world where tax evasion is possible?[5]

The existence of both purchase and sales invoices for the same transaction leads to the second important advantage claimed for a VAT commonly termed "self-enforcement".[6] This arises from the possibility of matching sales invoices against purchase invoices, making it difficult for intermediate goods sellers to understate sales, especially since purchasing firms have an incentive, other things equal, to declare purchase invoices to the sales tax administration (STA) and receive rebates.

---

[3] See, for example, Sandford and Godwin (1990).

[4] See, for example, Tait (1988) for a comprehensive discussion of the VAT.

[5] For recent work on optimal taxation under conditions of tax evasion see, for example, Cremer and Gavhari (1994). So far as we are aware, no paper has yet examined the optimal treatment of intermediate goods in the presence of tax evasion.

[6] See Sandford and Godwin, *op. cit.*

Thus, a second objective of this paper is to examine the effectiveness and efficiency implications of a VAT, implemented by the widely employed invoice method, in the presence of evasion and to study the extent to which "self-enforcement" takes place. Since, of course, no tax can magically enforce itself, we must first define self-enforcement. Essentially, we identify potential self-enforcement with a situation in which the optimal voluntary declaration by taxpayers is increasing not only in the level of enforcement by the STA but also with respect to voluntary reports by other taxpayers. For self-enforcement, furthermore, we require that taxpayers make positive reports in equilibrium so that taxpayers do have an impact on the voluntary disclosures of other taxpayers rather than only a potential impact.

We construct a two industry partial equilibrium model of VAT evading firms to examine these questions. STA enforcement is both through *auditing* and *cross-matching of purchase and sales invoices*, the latter activity capturing the interdependence of enforcement information on different firms. Our main findings are as follows:

i. Auditing and cross-matching have very different effects on the compliance behaviour of firms: Even in situations where auditing alone would lead to *zero reports* by firms, cross-matching can lead to *zero tax evasion* essentially due to the self-policing property of the VAT.

ii. Unlike auditing, cross-matching alters book-keeping and purchasing behaviour of firms. This may have adverse implications for allocative efficiency in the presence of scale economies in making purchases or sales through, for example, bulk orders.

iii. Auditing and cross-matching have independent effects on production efficiency via input prices: Distorted input prices may obtain due to cross-matching even when auditing alone leaves input prices undistorted and vice versa.

iv. In situations where auditing alone has no effect on the *output* decision of firms, a problem studied by Marelli (1984), cross-matching can still affect output decisions thus leading to a violation of the product-mix efficiency conditions (i.e., Marginal Rate of Substitution = Marginal Rate of Transformation) of the economy.

The very different implications we find in comparison to a model which only allows for auditing show that extension of the standard model to incorporate cross-matching is of importance. For the VAT, a comparison of (i) with (ii)-(iv) above suggests a possible trade-off between effective enforcement, and therefore the revenue generating ability of the VAT, and its impact on production efficiency and product-mix efficiency. Our findings also raise the possibility that the VAT is not part of an optimal commodity tax mix in a third-best world with tax evasion and costly enforcement.[7] This question, however, is beyond the scope of this paper.

In Section II we specify our two-industry framework in the absence of evasion. Evasion and enforcement are introduced in Section III. The specification of the information structure in Section III is intentionally detailed and based on a parable, so that the key differences between auditing and cross-matching in our model are readily apparent to the reader.

Our analysis begins in Section IV with an examination of the implications of cross-matching for the account-keeping, purchase and sales behavior of firms. Expected utility (of profits) maximizing firms are shown to prefer splitting input purchases between all available intermediate-input sellers, implying a possible loss of scale economies when such economies are present. This, however, leads endogenously to a simplification of the model, useful for further analysis.

The model is then used to examine the main VAT questions, the impact of evasion and enforcement on input prices and self-enforcement. In general, the VAT is neither self-enforcing nor leaves input prices undistorted (Section V). Somewhat surprisingly however, *it is possible for the VAT to be both self-enforcing and to leave input-prices undistorted with sufficient enforcement effort, regardless of effectiveness of the STA in carrying out audits* as measured by the (ex ante) expected success rate. Such high levels of enforcement effort may, of course, be too costly to implement especially if the STA is relatively inept at auditing. In other situations expected revenue maximizing audit and cross-matching may

[7] Alternatively, if a VAT continues to be optimal, taxation of inputs must form part of a third best optimal commodity tax.

lead to input price distortion even if there is no difference in per unit enforcement costs across industries (Section VI). Having dealt with input price distortion, in Section VII we describe necessary and sufficient conditions for the VAT to be self-enforcing. The conditions amount to merely requiring non-zero reports by all firms, so that our paper provides theoretical support for the claim that the VAT is largely self-enforcing in the presence of cross-matching.

In Section VIII, we turn to output distortion and demonstrate that output decisions of risk-averse firms may be affected by enforcement through cross-matching even under conditions, as in Marelli (1982), where auditing alone has no impact on output choice.

VAT evasion is examined in the paper under the strong assumption that the input-output structure of production is common knowledge. In a brief extension (Section IX), the importance of the STA's information about this structure for its ability to collect revenue from the VAT is demonstrated. Whether or not enforcement effectiveness is undermined, it is argued, depends crucially on the legal structure, specifically, whether the *onus of proof that an invoice is not a fake is on the taxpayer or the government.*

The model in the paper is highly simplified and leaves out many features of the real world VAT. To close our discussion, limitations and extensions of our analysis are discussed (Section X). In particular, we argue that our simplifying assumptions as to market structure and also the particular procedure for cross-matching that we assume, are unlikely to qualitatively affect our conclusions about split transactions, input distortion and self-enforcement.

II. Basic Industry Structure

The basic model we specify is very simple as the extension to cross-matching will lead to added complexity. Nevertheless, we believe that our model leaves out nothing crucial. Furthermore, even in our simple model important differences arise between the effects of different enforcement activities.

There is a final goods industry producing a homogenous good **F** and an intermediate goods industry producing a homogenous input, **I**. Final and intermediate goods producing firms are referred to as f-firms and i-firms respectively. The intermediate goods industry is assumed to be perfectly competitive with many identical i-firms. The number of i-firms is determined endogenously. The final goods industry is assumed to consist of **n** identical firms, each of whom is a monopolist in its own region, where regions are assumed to be non-overlapping and to possess identical downward sloping demand curves for the final good. The revenue of a representative f-firm is denoted **R(F)**. F-firms may buy inputs from more than one i-firm, the number of i-firms they purchase from being determined endogenously. Likewise, i-firms may sell inputs to more than one f-firm.

The cost of production for a representative i-firm is **W(I)**. **W(I)** has a U-shaped average cost curve reaching a minimum at $\mathbf{I}^*$. The cost of production for the representative f-firm consists of two parts, the cost of primary inputs, **C(F)**, and the cost of intermediate inputs, α**wF**, where α, the input-output coefficient, is assumed to be constant and **w** is the price per unit of input purchased (and α**F** units of input are purchased).[8] The marginal primary cost function is assumed to be positive and increasing. Revenue and cost functions are assumed to be at least twice differentiable. It is assumed that no inventories are held by f-firms or i-firms: all intermediate purchases are used up in production in the same period and all final and intermediate goods produced are sold.[9] Consequently, in the absence of taxation, profits are π_I**(I) = wI - W(I)** for i-firms and π**(F) = R(F) - C(F) -** α**wF** for f-firms. Given competition, π_I **= 0** and **I =** $\mathbf{I}^*$ in long-run equilibrium.[10] Without loss of generality, units of input are chosen so that α**=1**. This is a partial

---

[8] The extension of the analysis to include capital purchases requires an explicitly dynamic model as treatment of capital purchases may differ under different variants of a VAT. Either primary or intermediate inputs may, however, be interpreted as including the cost of capital services, depending on which variant of the VAT is being examined, without affecting the analysis. For a discussion of VAT variants see, for example, Due and Friedlaender (1973) and Due (1988, Chapter 16).

[9] This assumption closes off one possible channel of VAT evasion (see, for example, Tait (1988)).

[10] We assume the existence of stable and unique equilibria for all sets of tax and enforcement parameters that arise in the course of this analysis without further comment.

equilibrium model as factor supplies underlying cost curves and the demand for final goods are taken as given.

We assume throughout the paper that all firms in a taxed industry are liable to pay sales taxes - there are no exempt firms.[11] The ad valorem rates of tax across industries are taken to be identical and denoted by **t**. Clearly, the average cost curves of i-firms are shifted up by a constant amount in the presence of the VAT. This implies that zero profit equilibrium will continue to occur at a per i-firm production of $\mathbf{I}^*$. The equilibrium price of the intermediate good under the VAT is determined by the relation $\mathbf{w_v(1\text{-}t) = w}$. The number of i-firms in equilibrium is denoted **m**, where **m** is determined by equating demand for intermediate goods, **nF** with supply $\mathbf{mI}^*$.

Profits of a representative f-firm under the VAT are

$$\pi(F) = R(F)\ (1-t)\ -\ C(F)\ -\ w_v F(1-t). \qquad \textbf{(1)}$$

In (1) the price of intermediate goods is reduced by the tax $\mathbf{w_v t}$ rebated to producers per unit of intermediate goods purchased (against purchase invoices), to offset the tax paid by i-firms on their sales. Consequently, the net of rebate price of inputs continues to be **w**.

III. Incorporating VAT Evasion and Enforcement Activity

Penalties for evasion and the information structure are now specified in some detail. Even so, the main innovation is in the specification of the structure of cross-matching with other assumptions being similar to those in earlier work. In modeling evasion behavior, we first ensure that penalties on over-reported purchases and under-reported sales do not, by themselves, lead to differing incentives to engage in these two activities.

---

[11] In practice, various types of firms are exempt from tax leading to possible distortion in output and further enforcement problems.

A1. There is no penalty for not reporting purchases (firms may report purchases if they wish).

F-firms will, of course, need to report purchases to the extent that they wish to claim VAT rebates.

A2. Penalty for tax evasion is levied on net underpaid taxes detected at a constant rate $f>0$.

This ensures that independently varying reported sales or purchases has no impact on penalties provided total tax evaded is unaffected.[12]

Now turn to the structure of accounts.

A3. Firms are required by law to issue and keep copies of sales invoices which bear the names of both the buyer and the seller. Tax returns need not, however, be supported by copies of invoices. These must only be produced if the firm is audited.[13]

Thus, realistically enough, the STA will not be able to infer the identity of sellers of intermediate inputs to an f-firm or the identities of purchasers of inputs from a given i-firm in the absence of an audit.

A4. Invoices in books shown to the STA by an audited firm are (at least) equal in value to rebates claimed by the f-firm, or sales voluntarily declared by the i-firm, so that accounts and voluntary reports are consistent.

This assumption, which simplifies the analysis, may not be entirely innocuous.[14] Finally, we make an assumption which permits us to focus on the case of i-firms and f-firms which deal with each other at arms length. In comparison with the other assumptions made here, the study of situations where this assumption does not hold may be an important task for the future.

A5. Firms deal at arms length. In particular, there is no collusion between i-firms and f-firms to conceal transactions from tax authorities.

---

[12] To exclude gratuitous reports, we also assume that no firm reports purchases or sales that it is not required to report. This could be ensured without affecting our results by explicitly introducing transactions costs incurred by firms if sales or purchases are reported. Such costs are associated, for example, with additional book-keeping requirements.

[13] See, for example Tait (1988), Chapters 13 and 14. South Korea appears to be the only country that required invoices to be sent to the tax office. In such a case invoices pertaining to voluntarily reported sales can be matched.

[14] See the discussion in the concluding section.

We must now describe the information structure of the sales tax administration. The formal statement of our assumptions is followed by a description of the organizational parable we have in mind with this specification.

A6. Prices of the i-good and f-good as well as the input-output coefficient are common knowledge.

In a brief extension, we show that the STA's knowledge about the input-output coefficient is crucial to its ability to enforce VAT compliance.[15] Next, consider tax audits by the STA.

A7. Firms to be audited are all selected prior to the cross-matching process: no resources are available for a second round of audits after cross-matching.[16]

A8. The probability of auditing i-firms and f-firms are **q** and **p**, respectively, **$0 \leq p,q \leq 1$**.

Audit probabilities are taken to be equal to the fraction of firms audited. The number of firms audited in an industry is a policy variable.

A9. With probability **e, $0 \leq e \leq 1$**, all sales come to light in an audit, while with probability **1-e**, no unreported sales are revealed by the audit. **e** is technologically given.

This all or nothing assumption is standard in the literature as there is little to be gained from allowing for partial discovery of evasion in an audit. It will be obvious later that STA technological ability differing across industries, or differing per unit enforcement costs, would have made our task easier by enhancing the possibility of input price distortion with VAT evasion. To ensure that some evasion takes place we place an upper bound on **e**, similar to the standard condition imposed on the audit probability in the literature:

---

[15] Knowledge of the input-output coefficient implies that the STA can establish in the appropriate court of law that no more than α**F** units of the intermediate input are required to manufacture **F** units of output. An example is the case of a grocery store which merely acts as a regional outlet for various consumer goods. An example of a case where the STA's knowledge will be limited is a tailoring establishment: different tailors will use differing amounts of material to make similar suits for identical customers. A second example is where two different processes are in parallel use to make a product - the inefficient process not having been completely phased out.

[16] The possibility of additional audits may be important in practice. The assumption is innocuous here, given identical firms.

A10. **e(1+f) < 1**.

Next, turn to cross-matching of transactions. Invoices brought to light in audits form the information base for cross-matching.[17] South Korea is, to date, the only country which has attempted to implement a mechanism to match *all* purchase and sales invoices that its administration was aware of (Tait, 1988). Consequently, we allow for partial cross-matching of a subset of transactions on which the STA has information.

A11. Assume that a fraction **s (S)** of i-firm invoices known to the STA are matched with f-firm (i-firm) reports.

The information potentially available to the STA for cross-matching under our assumptions is summarized in Table 1.

To be concrete, imagine the following procedure for matching. After auditing, invoices are sent to the STA's matching division. The matching division must sort invoices received from each i-firm according to the f-firm named in the invoice. Before these invoices are received, the f-firm's tax file contains (a) the tax return for unaudited f-firms, (b) for audited f-firms, the return; a record of additional sales detected on audit; and invoices of purchases from different i-firms that are revealed voluntarily or discovered by the STA. The information for cross-matching of i-firm returns is similar.

[17] Over-reporting of purchases under a VAT will never be optimal for f-firms if the input-output coefficient is common knowledge. Fake invoices are considered in the extension in Section VII.

<table>
<tr><th colspan="4">Table 1: Information Potentially Available to the STA for Cross-Matching</th></tr>
<tr><th colspan="2">INFORMATION ON F-FIRMS</th><th colspan="2">INFORMATION ON I-FIRMS</th></tr>
<tr><th>F-firm audited unsuccessfully</th><th>F-firm not audited</th><th>I-firm audited unsuccessfully</th><th>I-firm not audited</th></tr>
<tr><td>I-firm audited successfully:
Purchases from i-firm reported by f-firm against sales to f-firm by i-firm under the VAT.</td><td rowspan="2">Total purchases imputed from sales reported by f-firm against total reported or discovered sales by audited i-firms to f-firm.</td><td>F-firm audited successfully:
Sales to f-firm reported by i-firm against purchases from i-firm by f-firm.</td><td rowspan="2">Total reported sales by i-firm against total reported or discovered purchases by audited f-firms from i-firm.</td></tr>
<tr><td>I-firm audited unsuccessfully:
Purchases from i-firm reported by f-firm against sales to f-firm reported by i-firm under the VAT.</td><td>F-firm audited unsuccessfully:
Sales to f-firm reported by i-firm against purchases from i-firm reported by f-firm.</td></tr>
</table>

This structure renders it impossible for the STA to limit in advance the sorting of invoices according to firms named in invoices: i-firm (f-firm) invoice records must be sorted even for f-firms (i-firms) that have been successfully audited. However, the STA can choose to sort records of only a subset of audited firms.[18] Furthermore, after sorting is complete, the matching wing can select a subset of assessment files to actually carry out matching or tallying of sales and purchase totals. So assume that:

A12. Invoices of a randomly chosen subset of audited firms are sorted and all sorted records are matched.

There are thus four enforcement activities: matching of invoices for firms in either industry and auditing of firms in either industry. This structure essentially captures the informational advantage of a VAT administration over other sales tax administrations in detecting sales by intermediate goods industries. Information on tax evasion by final goods firms may, however, be common to a larger class of taxes (such as the multi-point cascade sales tax).

[18] It should also be possible to first select invoices of firms that have been successfully audited. This confers no particular advantage.

IV. Consequences of cross-matching for account-keeping, purchase and sales transactions

Here three questions are answered. What fraction of purchases from an i-firm will the f-firm record in the accounts it shows to the STA, given the fraction of total sales it voluntarily declares? How much will an f-firm purchase from each i-firm given the number of i-firms it deals with and its total planned sales? How many i-firms will the f-firm make purchases from? For each of the three decisions, risk-neutral firms will act so as to minimize the expected detection of unreported sales.

Consider a representative f-firm. Since the input-output coefficient is normalized to unity, the quantity of input purchased is equal to the quantity of output sold. For this section total purchases are also normalized to one unit. The fraction of actual sales reported to the STA by an f-firm is denoted by Θ. Since the input-output coefficient is known, the STA can infer that at least Θ units have been purchased by the f-firm.

First examine the purchase by the f-firm from each i-firm and the proportion it records in its account books, taking as given the total number of i-firms, **z**, from which the f-firm makes purchases. The intuition here is strong. Given the fraction of sales reported by the f-firm, suppose that **k** of the i-firms from whom purchases are made are audited. Since any combination of **k** out of the **z** i-firms can be picked with equal probability under random auditing, roughly equal quantities should be purchased from each firm and a constant fraction, Θ, of these purchases should be recorded in the books shown to the STA if the f-firm is audited. With purchases of unequal size, the number of i-firms that have to be audited and matched to detect under-reporting could fall below that with equal sized purchases making the detection of under-reporting more likely. Of course, the number of i-firms that have to be audited and matched to detect evasion by the f-firm will depend on the fraction of sales reported by i-firms (denoted Φ). We state the result as a Lemma and relegate the proof to the Appendix.

Lemma: *If a risk-neutral f-firm under-reports its sales to the STA, then for any set of values of (a) the total sales of an f-firm, (b) the proportion of sales it reports to the STA and (c) the number of i-firms it makes purchases from, no other pattern of purchases and bookkeeping rules leads to lower expected detection of under-reporting than equal purchases from all i-firms and equal amounts of each purchase recorded in the f-firm's books of account.*

Other patterns of purchases and accounts may do just as well as equal purchases, provided amounts are not too different, but cannot do better. So an alternate way to state the lemma for risk neutral firms would be in terms of uniform convergence to equal purchases in the limit as the number of i-firms grows large. For risk averse f-firms, unequal purchases lead to increased risk[19] without lowering expected detection of under-reporting. *Consequently, risk averse firms will strictly prefer equal purchases and recorded amounts.* With regard to sales by i-firms, an identical argument *(mutatis mutandis)* shows that sales to f-firms of equal size and a constant fraction of each sale recorded in the i-firm's books dominates other patterns of sales with cross-matching. Consequently, without any real loss, we assume from here on that all purchases by f-firms are of equal size and that equal amounts of each purchase (sale) are recorded in books of account by f-firms (i-firms).

Now turn to the third question, the number of i-firms from which an f-firm makes purchases, given equal sized purchases. The answer is not obvious *a priori*, since there is a trade-off between a lower probability of discovery and a higher fraction discovered from each i-firm if purchases are made from fewer i-firms. Let the number of i-firms audited and then randomly chosen for cross-matching be $\mathbf{a} = \mathbf{As} < \mathbf{m}$, where **m** is now the number of i-firms in the long run equilibrium with evasion, **A** is the number of firms audited and **s** is the fraction selected for (sorting and) cross-matching. Since all i-firms have an equal chance of being audited or selected for matching, the probability that **k** of the **z** i-firms from which

[19] Increasing risk in the sense of a mean-preserving spread. See Rothschild and Stiglitz (1970).

an f-firm makes purchases are audited and selected for matching is given by:

$$h(m,a,z,k)=\frac{\Gamma(z,k)\ \Gamma(m-z,a-k)}{\Gamma(m,a)}, \qquad \textbf{(2)}$$

where, for example, $\Gamma(z,k) \equiv z!/k!(z-k)!$, is the number of combinations of **k** objects out of **z** objects. (2) is a hypergeometric distribution with parameters **m, a** and **z**.[20] The upper and lower limits of this distribution are respectively **Y=min(a,z)** and **y=max(0, a-(m-z))**. The second term in **y** merely recognizes that, if **z** is large, then some of the **z** i-firms from whom purchases are made will always be audited.

Given that **k** i-firms are audited, what is the distribution of the number of i-firms whose sales are detected? For such an i-firm the fraction of sales to any f-firm that is revealed is **Φ/z**. Additionally, since the probability of detection on audit of any i-firm is **e**, the probability of detection of the sales of **j** i-firms has a binomial distribution $b(e,k,j) = \Gamma(k,j)e^{j}(1-e)^{k-j}$, for **j = 0,1,2,...k.** Without confusion, this probability is denoted **b(k,j)**. With a successful audit, the remaining **(1-Φ)/z** of sales by the i-firm to the f-firm come to light.

From Table 1, for unaudited f-firms, the STA can only compare the total value of sales invoices to an f-firm from audited i-firms with the total purchases reported by the f-firm. Denote the minimum number of i-firms that need to be detected to establish under-reporting by the f-firm, given that **k** relevant i-firms are audited by **J(z,k,Φ)**. For a given value of **Θ**, **J(z,k,Φ)** is clearly weakly increasing in **z** and weakly decreasing in **k** and **Φ**. So the expected amount of under-reporting that will be detected, when **k** relevant i-firms are audited is

$$\sum_{j=J(z,k,\Phi)}^{k} b(k,j)\left[\frac{k\Phi}{z}+\frac{(1-\Phi)j}{z}-\Theta\right]. \qquad \textbf{(3)}$$

Consequently, if $J(z) = \min_k J(z,k,\Phi)$ is the minimum number of i-firms that need to be audited to detect

[20] In the usual urn analogy, the urn contains **z** black balls out of a total of **m** balls. The number of balls drawn without replacement from the urn is **a.** The hypergeometric distribution describes the number of black balls out of the **a** balls drawn from the urn.

under-reporting (**J(z)** is the smallest integer that is not less than **zΘ**), the expected amount of under-reporting that will be detected is given by

$$\sum_{K=J(z)}^{Y} h(z,k) \sum_{j=J(z,k,\Phi)}^{k} b(k,j)[\frac{k\Phi}{z} + \frac{(1-\Phi)j}{z} - \Theta]. \qquad \textbf{(4)}$$

Consider next the case of an audited f-firm. From Table 1, expected under-reporting that will be detected when **k** relevant i-firms are audited and selected for matching is given by **k[Φ + e(1-Φ) - Θ]/z** (noting that the mean of the relevant binomial distribution is **ke**).

We can now put the pieces together. The probability of not being audit is **(1-p),** the probability of an unsuccessful audit is **p(1-e)** and the probability of a successful audit is **pe**. If **Θ < min[Y/z, Φ + e(1-Φ)]**, expected under-reporting detected is, therefore, given by:

$$pe(1-\Theta) + p(1-e)\frac{a[\Phi + (1-\Phi)e - \Theta]}{m} + (1-pe)\sum_{k=J(z)}^{Y} h(z,k) \sum_{j=J(z,k,\Phi)}^{k} b(k,j)[\frac{k\Phi}{z} + \frac{(1-\Phi)j}{z} - \Theta]. \qquad \textbf{(5)}$$

In (5) the fact that the mean of the hypergeometric distribution is **az/m** has been used in the second term. If **Θ ≥ min[Y/z, Φ + e(1-Φ)]**, then cross-matching is irrelevant for either audited or unaudited firms and either the third term or the second term of (5) (or both) drops out. Given its report, **Θ**, a risk neutral f-firm will seek to choose **z** to minimize the quantity in (5). The following proposition can now be stated.

Proposition 1: *If the expected amount of under-reporting by an f-firm that will be discovered by the STA is given by (5) then the purchase of intermediate inputs by an f-firm from all* ***m*** *i-firms weakly (strictly) dominates purchasing from fewer i-firms for risk neutral (averse) f-firms.*

The proof is in the Appendix. The weak dominance result in the case of risk-neutral firms is, in fact, somewhat stronger than is apparent from the proposition. When **m** is not too small, there will exist a number **M≤m** such that purchases from at least **M** firms strictly dominates purchases from fewer firms: Weak dominance is "almost strict". In view of this we assume, henceforth, that f-firms purchase

intermediate goods from all i-firms.

With this assumption, the distribution in the third term of (5) collapses to a binomial distribution **b(a,e,j)**. The distribution of i-firm sales detected through cross-matching the VAT is similarly a binomial distribution where the number of independent trials corresponds to the number of audited f-firms selected for cross-matching.

There are two important implications of the analysis in this section. Firstly, the dominance of purchases from all i-firms may no longer be true if there are economies of scale to be reaped from bulk purchases. Even so, the results of this section establish that cross-matching may cause purchases to be smaller than optimal given evasion gains thus leading to lost scale economies. Second, since multiple purchases will only be feasible if the number of i-firms is large, the loss in surplus due to the sacrifice in purchase economies will be less severe in industries with few i-firms and absent in a monopoly, an effect running opposite to the usual pattern of deadweight losses due to suboptimal output decisions.

V. Evasion decisions by firms and input price distortion

Our model is related to existing models of sales tax evasion by monopoly firms (Marelli, 1982) and competitive firms (Virmani, 1989). Two new factors come into play in comparison with earlier work. First, it is possible to use cross-matching of invoices to detect evasion by firms who go through the audit round unscathed. Second, f-firms have an incentive to over-report purchases (that is, report fake purchases) in addition to under-reporting sales. Over-reporting is ruled out by the assumption that the input-output coefficient is common knowledge. The consequence of relaxing this assumption will be explored later. To simplify the analysis of evasion decisions, we follow Virmani (1989) and assume that firms are risk neutral.

Given the structure of expected detection in (5) and the implications of Proposition 1, the expected

profit of a representative i-firm is given by

$$\pi_I(I,\Phi)=w_v I-W(I)-tw_v I[\Phi+\lambda(1+f)], \tag{6}$$

where,

$$\lambda \equiv qe(1-\Phi)+q(1-e)pS[\Theta+e(1-\Theta)-\Phi] +(1-q)\sum_{J_I(n,\Theta,\Phi)}^{Snp} b(Snp,e,j)[\Theta Sp+\frac{(1-\Theta)j}{n}-\Phi], \quad for \quad \Phi<\min[\Theta+e(1-\Theta),Sp] \tag{7}$$

is the expected detection by the STA of under-reported sales. In (7), $\mathbf{J_I(.)}$ is, as before, the minimum number of f-firms that need to be audited to establish evasion by unaudited i-firms through cross-matching and **S** is the fraction of audited f-firms selected for cross-matching. The first term in (7) is the expected detection of under-reporting from auditing. The second term gives the expected detection of under-reporting through cross-matching if the i-firm is unsuccessfully audited given that f-firms report a fraction Θ of their sales. The term is zero if $\mathbf{\Phi \geq \Theta+e(1-\Theta)}$. The last term is the expected detection of under-reporting through cross-matching for unaudited i-firms, given that $\mathbf{\Phi < Sp}$.

Similarly, for f-firms, expected profits are given by

$$\pi(F,\Theta)=R(F)-C(F)-w_v F-[\Theta+\Omega(1+f)][R(F)-w_v F]t, \tag{8}$$

where

$$\Omega \equiv pe(1-\Theta)+p(1-e)qs[\Phi+(1-\Phi)e-\Theta] +(1-p)\sum_{J(m,\Phi,\Theta)}^{sA} b(sA,e,j)[\Phi sq+\frac{(1-\Phi)j}{m}-\Theta] \quad for \quad \Theta<\min[\Phi+e(1-\Phi),sq]. \tag{9}$$

The interpretation of the three terms in (9) is similar to the case of i-firms.

From (6) and (7) the fraction of sales that will be reported by an i-firm is independent of the level of output. Consequently, given any report, the average cost curve will be shifted up vertically by the amount of the effective tax (i.e. the expected tax cum penalty) in comparison with the no tax cost curve.

Thus, the long run output of an i-firm will still be **I***.[21] The number of i-firms will continue to be determined by the market equilibrium condition

$$nF = mI^* \qquad \textbf{(10)}$$

Equations (6) to (10) bring together the equations of the model being analysed in this section. From (7) and (9) it can be seen that the fraction of sales reported by i-firms and f-firms are mutually interdependent. Consequently, equilibrium requires, in addition to mutually consistent profit maximizing output decisions, equilibrium expectations concerning the reported sales fractions, Φ and Θ. This can be found by solving "reaction functions" of representative i- and f-firms for Φ and Θ. The examination of interior equilibria is taken up in Section VII below.

First, we define self-enforcement. The VAT will be said be *potentially self-enforcing* if the optimal report of f-firms is increasing in the optimal report of i-firms and vice-versa for a given set of enforcement parameters. If, in addition, firms reports positive sales in equilibrium, the VAT will be said to be *self-enforcing*. The latter requirement ensures that reports by firms do, in fact, influence reports by other firms rather than merely having the potential to do so. We now turn to an examination of conditions under which input prices remain undistorted. It is clear that for input prices not to be distorted by the VAT, i-goods must bear no net tax and receive no net subsidy. Our concern here is, therefore, in what enforcement regimes this situation obtains. We argue, first, that the model has the following, rather surprising, property.

Proposition 2*: For any value of the efficiency parameter* ***e****, 0<e<1, the STA can ensure that firms report their sales truthfully with sufficiently intensive cross-matching and auditing, due to the self-enforcement*

---

[21] This differs from the result of Virmani (1989) who concludes that tax evading competitive firms will produce below the minimum efficient scale. This is because he assumes that concealment of sales is costly with costs increasing with the *proportion* of sales the firm attempts to conceal from tax auditors.

property of the VAT.

Note that, given the assumption **e(1+f)<1**, both i-firms and f-firms will make zero reports if there is no cross-matching since, in this case, expected profits for both types of firms are decreasing in the fraction of sales reported regardless of audit rates. Thus positive voluntary reports by firms, if they obtain, must be due entirely to the additional effect of cross-matching.

To prove the proposition, suppose, initially, that cross-matching of unaudited firms is absent, so that the third terms in (7) and (9) drop out. Differentiate (6) with respect to **Φ** using (7). The derivative is **(1+f)[qe +q(1-e)pS]-1**. Clearly, for sufficiently large **q,p** and **S** this will be positive. In such a case expected profit maximization will require that i-firms make the report **Φ=Θ+e(1-Θ)** so that no evasion can be detected through matching. They will not report any higher since auditing alone is unable to deter under-reporting. Similarly, with sufficiently intensive auditing and cross-matching f-firms will report **Θ=Φ+e(1-Φ)**, from (8) and (9). But, these reaction functions intersect only at **Θ=Φ=1** which, therefore, must hold in equilibrium! Clearly, the argument will go through even if we allow, in addition, cross-matching of unaudited firms. This completes the argument. The precise condition for this full compliance equilibrium is, of course,

$$\min[qe+q(1-e)pS\ ,\ pe+p(1-e)qs](1+f)\ \geq\ 1. \qquad \textbf{(11)}$$

The interpretation of this rather unexpected finding is that large-scale cross-matching, possibly with the aid of high speed computers, can compensate to a large extent for lack of ability to detect evasion in traditional audits. Of course, large scale auditing be too costly to implement, though this is less likely with high speed computers.

Now consider the opposite case where auditing and matching rates are low enough that both types of firms make zero reports. From (6)-(9) a sufficient condition for zero reports to be optimal for both types of firms is

$$\max[qe+q(1-e)Sp+(1-q)\ ,\ pe+p(1-e)sq+(1-p)](1+f)\ <\ 1. \qquad \textbf{(12)}$$

This follows since zero is the optimal response of each type of firm, *given* a zero report by the other type if the condition holds. With zero reports by both i-firms and f-firms, (6) and (8) reduce to

$$\pi_I(I,0)=w_v I-W(I)-tw_v Ie[q+(1-qe)Sp](1+f), \qquad \textbf{(13)}$$

and

$$\pi(F,\ 0)=R(F)-C(F)-w_v F-te[p+(1-ep)sq](1+f)[R(F)-w_v F]. \qquad \textbf{(14)}$$

In long run equilibrium, the price if i-goods can be determined from (13) to be

$$w_v\ =\ \frac{w}{1\ -\ te(1+f)[q\ +\ (1-eq)Sp]}. \qquad \textbf{(15)}$$

If **S=s=1** or **q=p** and **S=s**, the denominator of (15) is identical to the net of rebate cost of inputs in (14) implying that input costs continue to be undistorted in the presence of evasion when enforcement effort is sufficiently weak (that is, (12) holds) and there is complete cross-matching.

A third case of undistorted input prices results from equal reported sales fractions for i- and f-firms and can be found by inspection of (6)-(7) and (8)-(9) to be where **S=s, p=q and n=m** in equilibrium[22]. We have thus shown that

Proposition 3*: Sufficient conditions for a VAT to leave input prices undistorted are that*

[i] *(11) holds; or*

[ii] *either* ***S=s=1*** *or* ***q=p***, ***S=s*** *and, furthermore, (12) holds; or*

[iii] ***S=s, p=q*** *and* ***n=m***.

Self-enforcement also occurs in case [i].

[22] The claim can also be seen from the reaction functions derived in section VII.

Stronger sufficient conditions or even necessary and sufficient conditions can be stated, but we have been unable to find any with intuitively appealing interpretations. One obvious and fairly general way to examine empirically if input prices are distorted on account of the VAT is to examine if any net revenue is raised from intermediate goods or if a net subsidy accrues to intermediate goods under the VAT. In practice, however, problems may arise in appropriately accounting for depreciation of capital goods.

VI. Revenue maximizing allocation of audit resources and input price distortion

Our concern in this section is to demonstrate that STA revenue maximization can require that audit probabilities and matching rates to differ across industries even if there is no difference in the cost of auditing or cross-matching across industries. As discussed above, this will strongly imply a possible conflict between the revenue goal and input-price neutrality. Furthermore, the analysis enables us to shed some light on factors that should influence the design of audit and matching strategy. To carry out the analysis, we assume that the STA is able to commit to an audit-cum-matching policy. A principal-agent analysis is then appropriate. We restrict attention to the case where audit costs are high enough so that (12) holds and firms find it optimal to make zero reports. Since a non-zero level of enforcement requires some auditing our examination can be restricted to checking if, at interior optima, **S=s** and **p=q** hold or if, at the optimum, **S=s=1**.

The equation for government revenue, **G**, is given by

$$G = n[R(F) - w_v F]t_v + nw_v F t_i - (np+mq)g_a - (Sp+sq)nmg_M \qquad \textbf{(16)}$$

where $\mathbf{t_i}$ is the effective tax rate on i-firms equal to **te(1+f)[q+Sp(1-eq)]** (so that $\mathbf{w_v = w/(1-t_i)}$); $\mathbf{t_v}$ is the effective tax rate on f-firms equal to **te(1+f)[p+sq(1-ep)]**; **q=A/m** and $\mathbf{g_a}$ and $\mathbf{g_M}$ are respectively the cost per audit and per matched transaction assumed to be identical across industries and constant per audit. The

equations determining **m** and **F** are (10) and the first order condition derived from (14) respectively.

The first order necessary conditions for revenue maximization can be found to be

$$\begin{aligned} (1-qe)\xi_1 - nmg_M &= 0 \\ (1-pe)\xi_2 - nmg_M &= 0 \\ [(\frac{s}{q})\frac{\partial G}{\partial s}] + (1-Spe)\xi_1 - mg_a &= 0 \\ [(\frac{S}{p})\frac{\partial G}{\partial S}] + (1-sqe)\xi_2 - ng_a &= 0 \end{aligned} \qquad \textbf{(17)}$$

where

$$\begin{aligned} \xi_1 &\equiv te(1+f)w_v\frac{(1-t_v)}{(1-t_i)}[nF+\frac{G_F}{D}], \\ \xi_2 &\equiv te(1+f)[n(R(F)+w_vF) + \frac{(R'(F)-w_v)G_F}{D}], \qquad \textbf{(18)} \\ D &\equiv (1-t_v)R''-C''+te(1+f)\frac{q}{F}[w_v(1-Spe)\frac{(1-t_v)}{1-t_i}+(R'(F)-w_v)s(1-ep)] < 0. \end{aligned}$$

In (18), $\mathbf{G_F}$ is the partial derivative of **G** with respect to **F**. The negative sign of **D** follows if the equilibrium in the 2 market model is stable. From (17), **p=q** and **S=s** will be a solution if and only if **n=m** and $\xi_1=\xi_2$. A necessary condition for the solution **S=s=1** can be found to be $\mathbf{g_A/g_m > (nm)^{0.5}}$. Thus audit and matching rates depend on there being an equal number of firms in the two industries besides appropriate cost and demand conditions. They will not, in general, leave input prices undistorted.

VII. Self-Enforcement

In Section V, we defined a VAT to be self-enforcing, naturally enough, as a situation in which optimal reports by each type of firm are increasing functions by reports made by the other kind of firm given that each makes a non-zero report in equilibrium. Clearly, there is no further scope for enforcement of any kind if firms report truthfully even in the absence of matching. Thus we are left with situations in which firms not only make non-zero reports but in which they would not report truthfully in the absence

of matching as candidates for equilibria in which self-enforcement obtains. One case in which the VAT is self-enforcing, where auditing is intensive enough to cause (11) to hold, has been examined in section V. Are more general results available? In fact a condition related to (12) characterizes cases of equilibrium self-enforcement under the VAT. The condition merely ensures that all-firms in the economy make non-zero reports. The analysis is, however, complicated by the fact that optimal reported output fractions will take on only a discrete set of values corresponding to the "kink points" of (7) and (9). To simplify the analysis assume that the number of i-firms in equilibrium and the number of f-firms are large enough so that a differentiable approximation to the binomial distribution **b(.)**, to be denoted **β(.)**, can be employed. In this case, (7) and (9) are replaced by:

$$\lambda \equiv qe(1-\Phi)+q(1-e)pS[\Theta+e(1-\Theta)-\Phi] + (1-q)\int_{J_I(n,\Theta,\Phi)}^{Snp} b(Snp,e,j)[\Theta Sp+\frac{(1-\Theta)j}{n}-\Phi], \quad for \quad \Phi<\min[\Theta+e(1-\Theta),Sp] \tag{19}$$

and

$$\Omega \equiv pe(1-\Theta)+p(1-e)qs[\Phi+(1-\Phi)e-\Theta] + (1-p)\int_{J(m,\Phi,\Theta)}^{sA} b(sA,e,j)[\Phi sq+\frac{(1-\Phi)j}{m}-\Theta] \quad for \quad \Theta<\min[\Phi+e(1-\Phi),sq]. \tag{20}$$

In (19) and (20), J_I and **J**, are no longer integers but are given by $J_I=n(\Phi-\Theta Sp)/(1-\Theta)$ and $J=m(\Theta-\Phi sq)/(1-\Phi)$ (recall that, for example, J_I is the minimum number of f-firms that need to be audited and matched to detect evasion by an unaudited i-firm).

With this simplification we can state:

Proposition 4: *Suppose that the expected profits of i-firms are given by (6) along with (19) while the expected profits of f-firms are given by (8) and (20). Then the VAT is self-enforcing if and only if*

$$\min[qe+q(1-e)Sp+(1-q)\ ,\ pe+p(1-e)sq+(1-p)](1+f)\ >\ 1. \qquad \textbf{(21)}$$

To prove the proposition we first show that (21) is necessary and sufficient for non-zero reports by both kinds of firms. We next argue that the report of each type of firm are increasing in the reports of the other, provided the report is not already maximal. We present the argument for f-firms only, since an analoguous argument holds for i-firms. Differentiating (8) with respect to **Φ** using (20) and the expression for **J(.)** yields the expression in (22):

$$-1+(1+f)\ [pe+p(1-e)qs+(1-q)\int_{J(m,\Phi,\Theta)}^{sA}\beta(sA,e,j)dj\ ]. \qquad \textbf{(22)}$$

It is easily seen from (22) that a necessary and sufficient condition for a zero report to be optimal for f-firms is that **(1+f)[pe+p(1-e)qs+(1-p)]≤1**. An analogous condition holds for i-firms. Consequently, (21) is necessary and sufficient for i-firms and also f-firms to make positive reports. Furthermore, setting the first-order condition in (22) to zero (the second order condition is easily verified), we see that this implies **J(.)** being constant regardless of **Φ**. From the expression for **J(.)** above, this implies that **Θ** is increasing with **Φ**. Thus, for interior solutions to (22) the proof is complete. On the other hand, if (22) has no interior solution, then we will have, from (22), **Θ≥sq** at which values cross-matching of unaudited f-firms cannot detect additional evasion. Second, also from (22), it is the case that **(1+f)[pe+p(1-e)qs]≥1**. In this case, as seen in Section V above, it is optimal for the f-firm to set **Θ=Φ+e(1-Φ)** so that, once more, **Θ** is increasing in **Φ** and **Φ** is positive. This completes the argument.

The nature of self-enforcement equilibria, at the intersection of reported sales reaction functions of i- and f-firms is shown for the case where **(1+f)max[pe+p(1-e)qs,qe+q(1-e)pS]<1** in Figure 1. Two possible cases arise for each type of firm, corresponding to **sq≤e** and **sq>e** for f-firms and **Sp≤e** and **Sp>e** for i-firms. The two cases correspond to the reaction functions that start from the **Θ*** and **Θ*'**or **Φ* and Φ*'**. From (22), furthermore, reaction functions are given by **J=constant** for f-firms and, analogously, $\mathbf{J_I}$**=constant** for i-firms. Consequently, from the expressions for **J** and $\mathbf{J_I}$ above, the reaction functions are

straight lines. The discontinuity in the reaction functions occurs at the point where cross-matching of audited firms starts to have an effect or when **sq=Φ+e(1-Φ)**. Self-enforcement is illustrated by the fact that the reaction functions are positively sloped so that higher reports by one kind of firm induce higher reports by the other type of firm. The four possible equilibria are labelled a, b, c and d in the diagram. A second type of equilibrium with self-enforcement, analysed in Section V, corresponds to the case where (11) holds so that **Φ=Θ+e(1-Θ), Θ=Φ+e(1-Φ)**. This equilibrium, will occur where both types of firms report truthfully which corresponds to the north-west corner of the box in Figure 1.

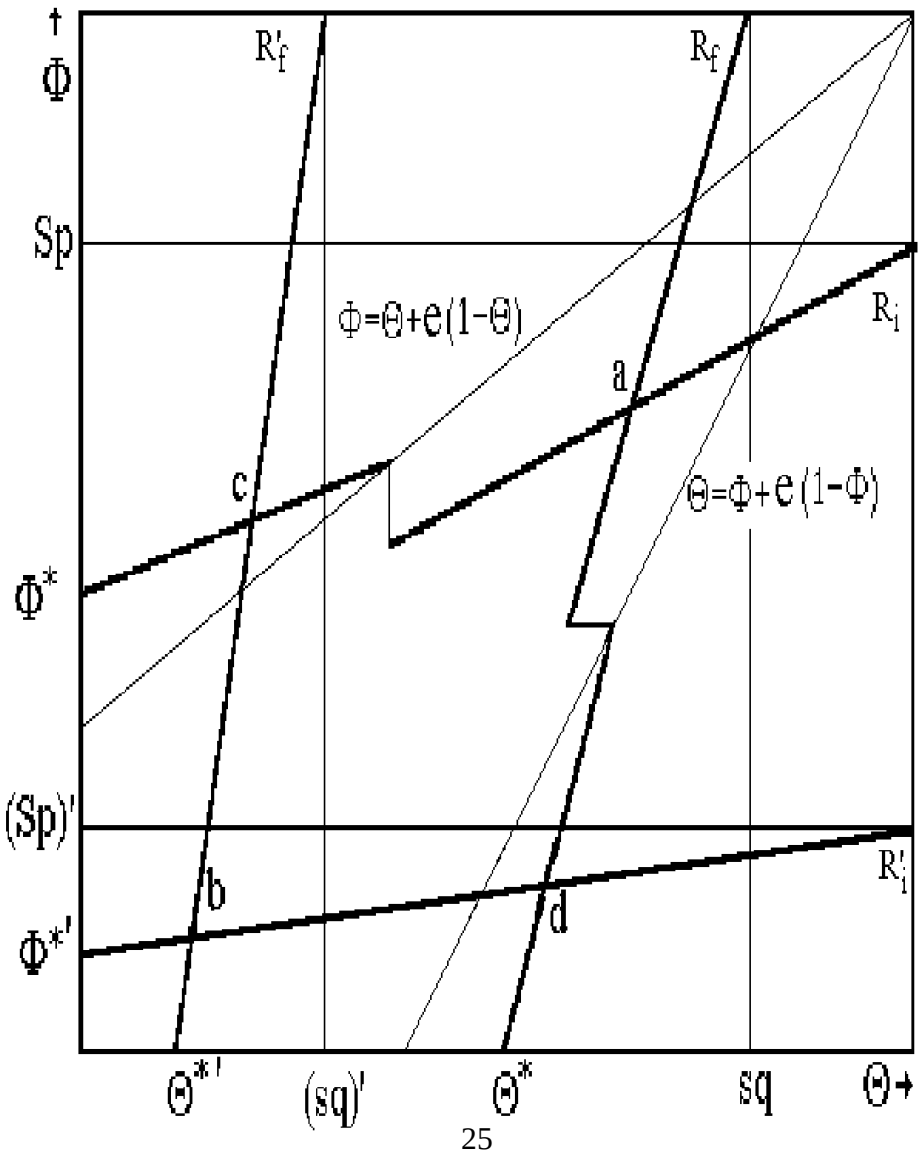

Figure 1: Self-Enforcement Equilibria

VIII. Cross-Matching and Output Distortion

We now turn to an additional effect of cross-matching: for risk averse f-firms, ***output*** decisions can be distorted by enforcement parameters even when auditing alone has no impact on output choice, a case examined by Marelli (1982). Consider first, the reporting and output decisions of f-firms when cross-matching is absent. Assume that firms' preferences are described by twice differentiable von Neuman - Morgenstern utility functions $\mathbf{U(\pi)}$. The expected utility of profits of a representative f-firm is given by

$$EU[\pi(F)] = peU[(R(F) - w_v F)\{1 - t - tf(1-\Theta)\} - C(F)] + (1-pe)U[(R(F) - w_v F)\ \{1 - t\Theta\} - C(F)]. \qquad \mathbf{(23)}$$

This is essentially the model of Marelli (1982) except that $\mathbf{R(F)}$ in his model is replaced by $\mathbf{R(F)-w_vF}$ here. Assuming an interior report, the first order condition for a maximum with respect to Θ is given (after simplification) by

$$peU'[(R(F) - w_v F)\ \{1 - t - tf(1-\Theta)\} - C(F)]tf + (1-pe)U'[(R(F) - w_v F)\ \{1 - t\Theta\} - C(F)]t\ =\ 0. \qquad \mathbf{(24)}$$

If (24) is substituted into the first order condition with respect to output derived from (23), the latter reduces to $\mathbf{[R'(F) - w_v](1-t) - C'(F) = 0}$ which is exactly the condition - aside from any input price distortion - that holds in the absence of evasion: As in Marelli (1982), tax evasion and enforcement through auditing alone has no effect on the output decision of risk averse firms making an interior report.

To simplify the extension assume, as in Section VII, that the number of i-firms is large enough so that $\mathbf{b(e,a,j)}$, can be approximated by a continuous and differentiable density function. Use the notation $\mathbf{x}$ to denote the fraction of output that the STA discovers (through auditing and cross-matching) in state of the world $\mathbf{x}$. Let the associated probability density be $\mathbf{\sigma(x)}$. The exact expressions for $\mathbf{x}$ and $\mathbf{\sigma(x)}$ will not be needed. Then, the expected utility of a risk averse firm can be written as

$$EU(F,\Theta) = \int_{\Theta}^{1} U[\pi^* - \Theta t\{R(F) - w_v F\} - (x-\Theta)t\{R(F) - w_v F\}(1+f)]\sigma(x)dx + U[\pi^* - \Theta t\{R(F) - w_v F\}]\left[\int_{0}^{\Theta}\sigma(x)dx\right], \tag{25}$$

where $\pi^* \equiv R(F) - C(F) - w_v F$. Substituting the first order condition for interior Θ derived from (25) into the first order condition for **F** and simplifying gives

$$R'(F) - C'(F) - w_v - \frac{t(1+f)\{R'(F) - w_v\}\int_{\Theta}^{1} U_x x\sigma(x)dx}{\int_{0}^{1} U_x\sigma(x)dx} = 0. \tag{26}$$

In (26) the notation U_x denotes marginal utility in state **x**. (26) shows that allowing for states of the world with partial discovery of under-reporting, which is here the outcome of cross-matching, leads to enforcement having an effect on the output decision of f-firms. Intuitively, with only auditing, there are two possible states of the world. Firms can use the two instruments available to them, Θ and **F**, to separately address risk and returns. With additional states of the world, such separation may no longer be possible and output decisions can be affected.

IX. Extension to Imperfectly Known Technology

Our final demonstration is to show that effective cross-matching is important for the VAT. Allowing for fake invoices, which vitiates the effectiveness of cross-matching, may seriously affect VAT revenue performance.[23] Specifically, we show that it is possible for f-firms to reduce their tax liability to

[23] Inflated claims of VAT refunds are a serious problem in practice. Tait (1988) points out that 44 percent of all VAT fraud in the Netherlands had to do with inflated refund claims. He points out that "businesses have been established solely to invent and print false invoices for sale to those wishing to defraud the revenue" (p 307). Furthermore, if capital purchases are allowed for and the VAT component of the cost capital goods qualifies for rebate, the problem can be much more serious.

zero (or less) by optimally over-reporting purchases if we relax the assumption that the STA knows the true input-output coefficient and, furthermore, if the *onus of proof that an invoice is fake is on the STA*. The question of onus of proof arises when an invoice which cannot be matched against any duplicate invoice from the other party named in the invoice. If the invoice itself is *prima facie* evidence of payment, then, in the absence of other evidence, the STA will have to bear the revenue loss. If, however, it is the taxpayer who must provide additional evidence of genuineness when a matching document cannot be found, then revenue loss from fake receipts will be curtailed by matching. Thus, under the assumptions above, when a fake invoice naming a particular i-firm is submitted by an f-firm, the STA may be unable to determine whether the f-firm is making a fake rebate claim or the i-firm is suppressing its sales unless, of course one of the firms has been successfully audited.[24]

Assume that the STA knows only that the true input-output coefficient is less than α^*, $\alpha^* > 1$. Furthermore, let the ratio of purchase invoices submitted to the STA to true invoices be $\mu > 1$. Since the total number of purchase invoices exceeds the actual number of invoices that can be verified through cross-matching, cross-matching will play no role in detecting over-invoicing by f-firms. Consequently, the expected payments to the STA by the f-firm will be given by

$$t\{\Theta R(F) - \mu w_v F + pe[R(F)(1-\Theta) - w_v F(1-\mu)](1+f)\}. \qquad (27)$$

Since expected taxes can be seen to be increasing in Θ (given $e(1+f)<1$), it will be optimal to set the reported sales fraction to μ/α^*. Making this substitution shows that expected taxes will be negative if $R(F)/F < \alpha^* w_v$. That is, the STA has to be unable to disprove that the cost per unit of intermediate inputs exceeds the price of the final goods. In this event, f-firms will over-report sales and pay less per unit of output than the rebate they receive. Clearly, while this may be possible for some firms in some periods, it cannot be considered an endemic shortcoming of the VAT. More generally, of course, greater evasion

[24] The device of naming a fictitious firm is precluded in most real world VATs as rebates are limited to invoices from intermediate goods suppliers registered with the STA. See Tait (1988) for a discussion.

is facilitated if firms can inflate refund claims since not only can sales be under-reported but net-of-rebate tax paid per unit of output will be less than in the case of known technology. Consequently, the case of known technology that we studied above, under which cross-matching works well, is the case in which VAT enforcement has its greatest relative effectiveness.

X. Conclusions and limitations

We have found that cross-matching can lead to distorted purchase decisions on the part of firms and distorted output decisions. Furthermore, lack of knowledge about production technology will limit the ability of the STA to raise revenue with a VAT since this reduces the effectiveness of matching and opens the door to excessive rebate claims. This depends on whether the onus of proof of the authenticity of purchase invoices is on the STA or the taxpayer. Input prices can be distorted under a VAT even with optimal (revenue-maximizing) enforcement policy. What is surprising, however, is that even if the STA is inefficient in carrying out audits, intensive cross-matching under the assumption of a known technology can lead to full-compliance with the VAT as a consequence of self-enforcement. This situation may become possible when the cost of cross-matching is low due to the deployment of high speed computers.

The industry structure of the model used to reach these conclusions is clearly rather special. So are some of the elements of the cross-matching procedure the STA is assumed to have. To what extent are our results likely to withstand generalization or modification? We believe generalizations in several directions can be accomodated.

Consider first, the industry structure. The result on purchase splitting (Proposition 1) is independent of industry structure and so will be unaffected if this is altered. The analysis of input price distortion depends essentially on there being a net tax or subsidy on intermediate industries. Once again, industry strcture has only a tangential role. The analysis could equally well have been for a competitive

or monopolistically competitive f-industry at the cost of having to take account of long-run zero profit conditions. What is affected, however, is the result pertaining to output distortion since Marelli's result cannot apply to a competitive or monopolistically competitive industry. Conversely, though there is no formal problem, our results on input price distortion may not be of great relevance for a single i-good seller or buyer given the improved matching opportunities.

Regarding the information structure for cross-matching, two assumptions need to be questioned. The first is assumption A4 that voluntary reports are reflected in the books maintained by firms. For i-firms, maintenance of accounts may not, in practice, be required in all countries. On the other hand, some countries impose a penalty on firms that do not maintain proper accounts. For f-firms, clearly, rebates claimed will be disallowed if not supported by invoices. The ability of the STA to use cross-matching to ensure compliance by i-firms is reduced if accounts do not support voluntary declarations. In such a situation self-enforcement will be adversely affected.

A second problem may be the sequence of actions by the STA in auditing and cross-matching that form part of our parable on cross-matching procedure. Other organizational procedures can lead to differences in the particular form of the equations of the model. Nevertheless, two insights from our analysis appear, intuitively, to have some claim to generality. The first is that more purchase or sales transactions makes matching difficult so that the incentive to split purchases that we found is not merely the result of our matching procedure. This is borne out, for example, by the analysis in Das-Gupta (1994) who shows that compliance is adversely affected in the presence of transactions splitting. The second insight is the interdependence of voluntary reports by firms which is likely to persist provided there is some cross-matching. Nevertheless, it may be worth examining alternative organizational arrangements explicitly in order to identify least cost organizational arrangements and distortionary consequences.

Despite the special structure of the model, we have been compelled to consider aspects of tax administration not usually dealt with in theoretical models of tax evasion. For example, the demonstration in Section VI involves studying the allocation of STA resources across enforcement activities, a dimension

of enforcement strategy that has not received attention in the literature. Other aspects of administration are still treated as "black boxes". These aspects relate, firstly, to the actual methods of detection of tax evasion on audit and secondly, to the STA's system for identifying taxpayers. Since the extensions we did incorporate have enabled a richer analysis of administrative activity, for example the problem of allocation of STA resources, further extensions may be worthwhile. Other aspects of tax administration include, for example, collection lags in an inflationary environment, policy for registration of firms and legal requirements or associated non-compliance penalties for maintenance of books of account.

The modelling of the VAT in this paper is also simplified and leaves out many features of a real world VAT. This is particularly true in the context of open economies and when considering investment decisions. Clearly, much additional work needs to be done in analyzing the VAT and sales taxes in general. A re-examination of the optimal commodity tax question, to be of some utility, must await a clarification of the major unresolved issues.

REFERENCES

Bagchi, Amaresh, Richard M. Bird and Arindam Das-Gupta (1995) "An Economic Approach to Tax Administration Reform" Working Paper, International Tax Program, University of Toronto.

Cowell, Frank A. (1990), *Cheating the Government*, Cambridge, Mass: MIT Press.

Cremer, Helmuth and Firouz Gahvari (1994) " Tax Evasion, Concealment and the Optimal Linear Income Tax" *Scandinavian Journal of Economics* 96, 216-239.

Das-Gupta, Arindam (1994) "A Theory of Hard to Tax Groups", *Public Finance*, 49(Supplement): 28-39.

Das-Gupta, Arindam, Dilip Mookherjee and D.P. Panta (1992) "Income Tax Enforcement in India: A Preliminary Analysis", mimeo, National Institute of Public Finance and Policy, New Delhi.

Diamond, Peter A. and James A. Mirlees (1971) "Optimal Taxation and Public Production I: Production Efficiency and II: Tax Rules", *American Economic Review*, 61: 8-27 and 261-278.

Due, John F. (1988) *Indirect Taxation in Developing Economies* (Revised Edition), (Baltimore: The Johns Hopkins Press).

Due, John F. and Ann F. Friedlaender (1973) *Government Finance: Economics of the Public Sector*, (Homewood, Ill: Richard D. Irwin, Inc).

Marelli, Massimo (1982) "On Indirect Tax Evasion" *Journal of Public Economics*, 25:181-196.

Rothschild, Michael and Joseph E. Stiglitz (1970) "Increasing Risk: I. A Definition" *Journal of Economic Theory*, 2: 225-43.

Sandford, Cedric and Michael Godwin (1990) " VAT Administration and Compliance in Britain" in *Value Added Taxation in Developing Countries*, Malcolm Gillis, et. al., Editors, The World Bank: Washington, D.C.

Slemrod, Joel and Shlomo Yitzhaki (1987) "On the Optimal Size of a Tax Collection Agency", *Scandinavian Journal of Economics*, 89: 183-92.

Tait, Alan A. (1988) *Value-Added Tax: International Practice and Problems*, (Washington D.C: International Monetary Fund).

Tait, Alan A. (ed.) (1991) *Value-Added Tax: Administrative and Policy Issues*, Occasional Paper No. 88, International Monetary Fund, Washington D.C., October.

Virmani, Arvind (1989) "Indirect Tax Evasion and Production Efficiency" *Journal of Public Economics*, 39:223-237.

Appendix: Proofs of the Lemma and Proposition 1

Proof of the Lemma

The result depends on showing that equal purchase and accounted fractions belongs to the set of undominated purchasing and accounting strategies. Note that the average purchase by an f-firm from an i-firm is **1/z**. Consider, first, the case where the f-firm has not been audited. For such firms, books of accounts remain unexamined and are irrelevant. Given random auditing, if **k** i-firms **($k > z\Theta$)** from which purchases have been made are audited and selected for matching, then the expected fraction of total purchases by the f-firm that will be identified by the STA is **$k\Phi/z$** regardless of the exact pattern of purchases from different i-firms. Again, since the probability of unreported sales by an i-firm being detected is a constant fraction **e**, (and thus independent of the probability that under-reported sales by any other i-firm are detected), the expected value of additional sales to the f-firm discovered from these **k** i-firms is **$ke(1-\Phi)/z$** regardless of its exact pattern of purchases. Adding the two parts gives the total expected discovery of unreported purchases when **k** i-firms are audited and matched as **$k[\Phi+e(1-\Phi)]/z$** - **Θ** independent of the pattern of purchases.

Now let **J** be the minimum number of i-firms that have to be audited and matched for under-reporting by the f-firm to be detected, given that equal amounts are purchased from all i-firms. Clearly, **J** is the smallest integer such that **$J \geq \Theta z$**. If **J-1** i-firms are audited and matched, then no under-reporting will be detected under equal purchases. However, with certain other purchase patterns, **J-1** transactions may be sufficient to detect under-reporting - such patterns will therefore be sub-optimal. Equally clearly, no unequal purchase pattern can give rise to a situation where more than **J** i-firms are required to detect evasion. Hence the Lemma is trivially true for risk neutral firms. Now examine the case of unsuccessfully audited f-firms. Note that exactly a fraction **Θ** of total purchases is recorded in the books of account of the f-firm (the firm will pay additional taxes if more is recorded and less is ruled out by assumption). The

expected discovery from each i-firm is now $[\Phi+e(1-\Phi) - \Theta_j]x_j/z$, where x_j/z is the actual amount purchased from the jth i-firm and Θ_j, $z+1\Theta_1x_1+..+\Theta_zx_z=\Theta$, is the amount of this purchase recorded in the f-firms books of account. Once again, the expected discovery if k i-firms are selected is independent of the exact pattern of purchases being equal to $k[\Phi+e(1-\Phi) - \Theta]/z$, so that equal purchases and accounts will do as well as any other pattern.

Remark: Bounds on optimal purchase patterns can be obtained as follows. Denote the maximum quantity purchased from any one i-firm by τ and the corresponding minimum purchase quantity by τ_m. Then it must be true that $\tau + [(J-2)(1-\tau)/(z-1)] \leq \Theta \leq J/z$. This follows from the restriction to J/z on the combined purchases from $J-1$ i-firms and the fact that τ will be at its maximum if all other transactions are as small as possible since any combination of them can be selected for audit. Since there are $z-1$ other i-firms that have made sales to the f-firm all other purchases must be equal to $(1-\tau)/(z-1)$. The left hand inequality states that the total detection given that $(J-1)$ i-firms are audited and matched must not exceed Θ. Rearrangement gives $\tau \leq [(2z-J)/z(z+1-J)]$ and $\tau_m \geq (z-J)/z(z+1-J)$. This upper bound on τ is increasing with J and does not impose any restriction on the size of τ if $J=z$ but yields an interior bound if $J<z$. That is if all z i-firms have to be audited to detect evasion, τ can be any number less than Θ. At the other extreme, if $J=2$, (z must be at least 2 for detection to be possible with fewer transactions) then $\tau = 2/z$.

Proof of Proposition 1

Since the first two term in (5) are independent of z, we need to show that $\delta(z,\Phi) \geq \delta(m,\Phi)$ for all $z<m$ where $\delta(z,\Phi)$ is defined as

$$\delta(z,\Phi) \equiv \sum_{k=J(z)}^{Y} h(z,k) \sum_{j=J(z,k,\Phi)}^{k} b(k,j)\left[\frac{k\Phi}{z} + \frac{(1-\Phi)j}{z} - \Theta\right] \qquad \textbf{(A1)}$$

and $\delta(m,\Phi)$ is as in (A1) except that, since purchases are made from all matched i-firms, the distribution of purchases discovered through cross-matching collapses to $b(a,e,j)$. Note first that, since $Y \leq a$, if $\Theta \geq a/m$, then by purchasing from m i-firms, the f-firm can ensure that no under-reporting is detected. Consequently, the proof is complete if the claim is shown to hold for $\Theta < a/m$.

We show, by induction on **z-J(z)**, that $\delta(z,\Phi) \geq \delta(z+1,\Phi)$ for $\Theta < a/m$. Now either **Y=z+1** with **z+1** transactions or **Y=a** with both **z** and **z+1** transactions. It suffices to consider the first case which is more stringent. Secondly, with **z+1** transactions, the lower limits of the binomial distributions in (A1) associated with each **k** are at least as great as with **z** transactions.

By direct computation, the inequality can be shown to hold for **J(z)=z**, noting that, $\Theta \geq z/(z+1)$ in this case.

Now assume the inequality is true for **J(z)=J'**.

The (cumulative) probability of at most **k** successes out of **z** possible successes with a hypergeometric distribution exceeds that of at most **k** successes out of **z+1** possible successes for **k≤z** since

$$h(z,k) - h(z+1,k) \;=\; h(z+1,k)\;\left[\frac{(z+1)a-(m+1)k}{(z+1)(m-z-a+k)}\right]. \qquad \textbf{(A2)}$$

That is, the term in square brackets is positive at **k=0** but decreasing in **k** and cumulative properties must sum to unity so that the graphs of the two cumulative distributions cannot intersect at **k≤z**. Consequently, the probability of at least **k** successes out of **z** possible successes is less than that out of **z+1** successes.

Expanding $\delta(.)$, we have, by assumption:

$$\sum_{k=J'}^{z} h(z,k) \sum_{j=J(z,k,\Phi)}^{k} b(k,j)\left[\frac{k\Phi}{z}+\frac{(1-\Phi)j}{z}-\Theta\right] \;\geq\; \sum_{k=J'}^{z+1} h(z+1,k) \sum_{j=J(z+1,k,\Phi)}^{k} b(k,j)\left[\frac{k\Phi}{z+1}+\frac{(1-\Phi)j}{z+1}-\Theta\right]. \qquad \textbf{(A3)}$$

Given (i) the property of the hypergeometric distribution discussed above and (ii) that the expected discovery given that **z+1** relevant firms are audited and matched exceeds that when fewer firms are audited and matched: the term involving **h(z+1,z+1)** can be apportioned among the remaining terms on the right hand side so that A3 can be rewritten as

$$\sum_{k=J'}^{z} h(z,k) \sum_{j=J(z,k,\Phi)}^{k} b(k,j)\left[\frac{k\Phi}{z}+\frac{(1-\Phi)j}{z}-\Theta\right] \;\geq\; \sum_{k=J'}^{z} h(z,k) \sum_{j=J(z+1,k,\Phi)}^{k} b(k,j)\left[\frac{k\Phi}{z+1}+\frac{(1-\Phi)j}{z+1}-\Theta\right] + \textit{Remainder}. \qquad \textbf{(A4)}$$

By the arguments above, the remainder is positive and decreasing in Θ. Clearly, therefore, A4 will

continue to hold if Θ is replaced by any smaller non-negative fraction, in particular values of Θ, say Θ^*, for which $\mathbf{J(z)=J'-1}$.

It only remains, therefore, to add terms for the case where **J'-1** firms are audited and matched to both sides of A3, noting that $\Theta=\Theta^*$. Since $\Theta<\mathbf{a/m}$, **J'** cannot exceed the smallest integer at least as great as **za/m** so that **J'-1** < **za/m**. Using A2 and this fact it is easily computed that the term added to the left hand side exceeds that added to the right. Thus, the desired inequality is true for **J'-1** firms if it is true for **J'** firms. This completes the proof for risk neutral firms.

For risk averse firms note that if the expected discovery of under-reporting when purchases are made from **m** firms is identical to that when purchases are made from fewer than **m** firms, then the former will be preferred. This is because the distribution of discovered under-reporting with fewer than **m** firms can be obtained as a mean preserving spread of the distribution with **m** firms. In fact, the mean discovery of under-reporting with **m** firms is lower than that with fewer firms, stregthening the dominance of purchases from **m** firms.