On Optimal Taxation with Costly Administration

By WALTER PERRIN HELLER AND KARL SHELL*

In adopting a set of taxes, governments are influenced by the relative costs of administering and enforcing each kind of tax. Similarly, costs of compliance and transactions for households and firms depend on the mix of taxes used. Yet, in previous studies of optimal taxation (e.g., Peter Diamond and James Mirrlees), administration and transaction costs play no role in the determination of government tax policy. We offer here a first attempt at incorporating these important costs[1] into a formal model of optimal taxation.[2] In what follows, the class of tax instruments to be used is endogenously determined by an explicit model of an economy with an administrative sector, rather than being exogenously given as in previous studies.

We build on the equilibrium-with-transaction-cost literature (Duncan Foley, Frank Hahn 1971) to extend the theory of optimal taxation to account for costly transactions. We introduce a new concept, the administrative feasibility set, which describes the costs of government tax administration. Because of limited space, we content ourselves with describing some basic elements of the model and then employing the model to consider whether maximization of social welfare requires production efficiency, i.e., whether production efficiency is "desirable."[3] When transactions are costly, pure production efficiency is optimal only in very special and unlikely circumstances. We next extend our notion of efficiency to accommodate the tradeoff between pure production efficiency and efficiency in transaction. We then go on to prove a theorem establishing the desirability of efficiency of production-*cum*-transaction plans, but strong conditions are required to achieve this result. Contrary to the conclusions of the previous literature—which omitted the role of government administration costs—our examples and theorems suggest that optimality of production-*cum*-transaction efficiency is very much in doubt for real-world economies.

I. The Model

Households. Household h is endowed with a vector of commodities, $\omega^h > 0$, including labor and human capital services. Households can sell only on the wholesale market and can buy only on the retail market. For household h, the vector of

* University of Pennsylvania. Research support from the National Science Foundation and the Fels Center of Government is gratefully acknowledged. We thank Tony Atkinson, Bent Hansen, Mark R. Killingsworth, David Starrett, members of the Stanford University Institute for Mathematical Studies in the Social Sciences summer seminar and members of the University of Pennsylvania seminar on the endogenous theory of policy for helpful suggestions. Proofs of theorems appear in a Mathematical Appendix which is available from the authors on request.

[1] These include costs of enforcing tax laws and costs of complying with tax laws as well as costs of information processing and information transmission.

[2] This is not to say that the role which administrative costs should play in a full theory of optimal taxation has been overlooked; just the opposite is the case. Diamond and Mirrlees, for example, state in their Concluding Remarks: "As economists have been aware, the omitted constraints on communication, calculation, and administration of an economy . . . limit the direct applicability of the implications of this theory to policy problems. . . ."

[3] If production efficiency is desirable, then at the optimum there is no tradeoff between equity and production efficiency.

goods purchased (at retail) is $x^h \geq 0$ and the vector of goods sold (at wholesale) is $y^h \geq 0$. The household consumption vector is then $c^h = \omega^h + x^h - y^h \geq 0$. If households face the vector of buying prices p and the vector of selling prices q, then the budget constraint for household h is: $px^h \leq qy^h$ + (profits distributed to h) − (net direct taxes levied on h). The household is assumed to maximize its strictly quasi-concave utility function, $u^h(c^h)$, subject to its budget constraint.

Production and marketing. Firms purchase wholesale inputs from households and purchase wholesale and retail inputs from other firms. Firms sell retail outputs to households and sell retail and wholesale outputs to other firms. Private firm f ($f = 1, \ldots, F$) maximizes profits at prevailing producer retail prices, $\hat{p}$, and prevailing producer wholesale prices, $\hat{q}$. The firm's production-transaction plan must lie within its given *production-cum-transaction* feasibility set, A^f. Let w^f be firm f's vector of net output on the wholesale market (i.e., output sold wholesale minus inputs purchased on the wholesale market). Similarly, let r^f be the vector of net output on retail markets. Then the production-transaction plan of firm f is feasible if (w^f, r^f) belongs to the set A^f. The firm's profits are thus $\pi^f = \hat{p}r^f + \hat{q}w^f$. For each of the F private firms we assume that A^f is a closed, convex set, so that competition is sustainable. Let A^{F+1} represent the government production-transaction opportunity set. We can think of the government production-transaction sector as "firm" $F+1$, but the government need not maximize profits.

The government. Before taxation policy can be studied, the universe of available tax instruments must be specified. We consider a universe which includes all those instruments featured in the previous optimal taxation literature. Some new instruments are also included.

The government has four basic tax-subsidy instruments at its disposal. In general, each tax used and the rates at which the tax is set will affect the government's administrative cost. The basic instruments considered are (1) commodity taxes on retail markets, (2) commodity taxes on wholesale markets, (3) profits taxes, including firm-specific profits taxation rates and firm-specific licensing fees and licensing subsidies, and (4) lump-sum taxes and subsidies for households. Since labor is one of our commodities, we include proportional income taxes in our list of instruments.

Why do we include lump-sum taxes in this list? After all, it is frequently argued that lump-sum taxes are so costly that they are rarely used by governments despite their advantages. Some kinds of lump-sum taxes (e.g., a tax based on age) may not be costly to administer but the use of the kinds of lump-sum taxes which are likely to be interesting for optimal distribution may require information which is very costly to acquire (e.g., "ability" levels of all individuals). Our point is that whether or not the government will employ lump-sum taxes is a matter which ought to be determined within the model rather than by *a priori* specification.

The government administrative feasibility set. How does one model costs of administering the tax system? Following what we take to be the spirit of the transaction-cost literature, we introduce the concept of the administrative feasibility set, G, which directly and indirectly relates g, the vector of real resources used up in administration, to the government tax instruments employed.

Let $\tilde{x}$ be the vector of vectors of household purchases, $\tilde{x} = \{x^h\}_1^H$, and similarly, let $\tilde{y}$ be the vector of vectors of household sales, $\tilde{y} = \{y^h\}_1^H$. If there are M commodities and H households, then $\tilde{x}$ and $\tilde{y}$ are vectors of dimension MH. Let m be the

H-dimensional vector of lump-sum subsidies to households; n the vector of lump-sum taxes. Let $\beta^{1f} \leq 1$ be the rate of profit taxation for firm f while β^{2f} $(-\infty < \beta^{2f} < +\infty)$ is the licensing fee to firm f. Therefore after-tax profit, α^f, is related to pretax profit, π^f, by $\alpha^f = (1-\beta^{1f})(\pi^f - \beta^{2f}) \geq 0$. The requirement that α^f be nonnegative is made to incorporate limited liability of stockholders into the model and is a constraint on the government in selecting β^{2f}. Let β denote the $2F$-dimensional vector $\{\beta^{1f}, \beta^{2f}\}_1^F$.

An administrative plan of the government is described by the vector $(\bar{x}, \bar{y}, m, n, \beta, g)$. The set G is then the set of all *feasible* government administrative plans. Administrative costs related to lump-sum taxes and subsidies and profits taxes are thus accounted for directly through the entry of m, n, and β in the administrative technology. The costs of commodity taxes, which drive a wedge between consumer and producer prices, are indirectly accounted for through the consumer purchase and sales vectors, $\bar{x}$ and $\bar{y}$. Presumably, the "farther" $\bar{x}$ and $\bar{y}$ are from their *laissez-faire* values, the higher is the cost of administering the underlying commodity taxes (i.e., the higher is g). To relate administrative costs directly to commodity taxes would mean including prices in the vectors belonging to G. We have not yet explored this possibility, nor are we confident that such a formulation would more accurately model the administrative costs of commodity taxation.

In order to fix ideas about our description of administrative costs, we can consider as an example the special case of *laissez-faire*. In terms of our nomenclature, *laissez-faire* is feasible only if the vector $(\bar{x}^0, \bar{y}^0, 0, 0, 0, 0)$ is an element in the set G, where $\bar{x}^0$ and $\bar{y}^0$ are *laissez-faire* equilibrium allocation vectors. That is, *laissez-faire* is feasible only if zero government administration costs are incurred when all tax and subsidy rates are zero. This would not be the case if the administrative costs of government enforcement of private contracts are included in G.

Our modeling of costly administration in terms of the set G is, of course, crucial. The usefulness of the subsequent analysis probably hinges on whether or not the administrative feasibility set is an appropriate way of describing the costs of alternative taxation policies. It is, therefore, worth noting some features of this new ingredient in the general equilibrium model of taxation.

First, the administrative feasibility set is expressed in "reduced-form." The production feasibility set of the classical equilibrium model is independent of tastes and endowments of households. In our model, however, the specification of the set G depends on the economy to which it is applied. Thus, one would expect that the administrative costs of achieving an egalitarian consumer allocation for an economy with small differences in initial household endowment are less than such costs for an economy with great disparities in initial endowments. That suggests that the set G is dependent upon the basic parameters of the economy (e.g., the pretax distribution of wealth). This is an important point since cross-section and time-series estimates of the administrative cost set cannot be made until the model is extended to relate changes in the set G to changes in fundamental economic parameters. Nonetheless, the model as it now stands is useful in assessing questions about optimal tax policy in a given economy as long as the set G is known to the policymaker.

Second, the set G is valid only for a given price convention. For instance, if prices are constrained to sum to unity, then administrative feasibility would be described in terms of an administrative set, G_1. If instead, prices are constrained to sum to

two, then another set, G_2, would be applicable. Remember that direct household taxes and subsidies, m and n, and license fees, β^2, are in units-of-account (say, dollars). If (m, n, β^2) were doubled and all prices doubled, *ceteris paribus*, should not costs of administration, g, be unaffected? This avoidance of "*numeraire* illusion" is easy to model, and we should emphasize that it is not a statement about the form of any given administrative set but rather how one set G_1 is related to another set G_2.

Finally, for simplicity, it is assumed that resources devoted to enforcement accomplish perfect compliance, in that all households and firms exactly fulfill their obligations. We feel that our model can be extended to handle imperfect enforcement and compliance; indeed, a substantial public finance literature analyzes what might be called the "optimal degree of enforcement." For example, what level of resources should the government devote to tax audit? The dual strategic question for the taxpayer is: What degree of cheating should be risked? In the present analysis such questions are ignored, for at this stage the extension of the model to cover such considerations would only complicate matters unnecessarily.

Social goals. We assume that the government seeks to set taxation policies and operate government production so as to maximize a strictly increasing, individualistic Bergson welfare function $W(u^1, \ldots, u^h, \ldots, u^H)$. The government is constrained by the behavior of households (maximization of utilities subject to budget constraints), the behavior of firms (maximization of profits subject to production-transaction feasibility), feasibility of government production and transaction, and administrative feasibility of government taxation policy. Of course, it is required that materials balance in the economy, which implies (by Walras' Law) that the government satisfies its budget constraint.

In positing a government which maximizes W, we—and previous authors on optimal taxation—have made a sharp and unrealistic distinction between public agents (who are assumed to seek to maximize social welfare) and private agents (who are assumed to maximize their own utilities). We are aware that in a fundamental sense all agents are private agents. Government bureaucrats are red-blooded people with individual goals in possible conflict with social policy. It is our long-run objective to incorporate such considerations into the analysis, but we will find it convenient at present to follow the custom of making this somewhat artificial distinction between public and private agents.

II. Should Production be Efficient or Inefficient?

Diamond and Mirrlees establish the desirability of production efficiency in a model with costless transaction, costless administration of commodity taxes (the only "feasible" instruments) and constant returns-to-scale technologies. Production efficiency is not necessarily optimal when decreasing returns and consequently positive profits are allowed. Partha Dasgupta and Joseph Stiglitz allow for decreasing returns and show that when firm-specific profits taxes are feasible and the private production sector is already operating efficiently, then it is optimal for the government production sector to operate so as to attain aggregate production efficiency. If there are very large administrative set-up costs in moving to a system of firm-specific taxes, then these instruments will not be selected, and one would expect that production-*cum*-transaction inefficiency may be optimal. In what follows, we generalize and formalize this intuition: Absence of administrative set-up costs is crucial to our efficiency theorems. We argue that these set-up costs (or "spikes")

are fundamental to tax administration and are likely to be pervasive in practice. The assumptions we need for our first efficiency theorem follow.

Assumptions. (1) A^f is closed, convex and permits free disposal for $f=1, \ldots, F, F+1$. Convexity of the F private sets allows for decentralization with a price system. Convexity of the government set, A^{F+1}, is assumed because it is known in simpler models (Hahn 1973) that nonconvexity of the public production set can by itself lead to failure of the efficiency theorem.

(2) The production-transaction possibility frontiers (the upper boundaries of the sets $\{A^f\}_1^{F+1}$) are assumed to be differentiable and the possibility frontiers are assumed to exhibit maximality. By this, we mean that every point on the possibility frontier is efficient, i.e., there are no horizontal or vertical flats.

(3) The administrative feasibility set, G, is closed, exhibits maximality of its possibility frontier and exhibits free disposal in the limited sense that if $(\bar{x}, \bar{y}, m, n, \beta, g) \in G$ then $(\bar{x}, \bar{y}, m, n, \beta, g^\dagger) \in G$ for any $g^\dagger \geq g$.

(4) The economy exhibits *weak sensitivity* of gross profits to profits taxes. We must rule out seemingly unusual cases where the indirect effects of profits taxation swamp the direct effects. Weak sensitivity obtains when marginal costs of administering license fees and subsidies are zero or small or when the administrative sector is small by comparison with the total level of economic activity.

(5) A constrained welfare maximum exists.

(6) At an optimum, wholesale and retail consumer prices differ for at least one commodity. This (seemingly weaker) hypothesis replaces the Diamond-Mirrlees assumption of a produced consumer good or nonproduced production input. Alternatively, one could retain the Diamond-Mirrlees assumption.

Efficiency Theorem1. If assumptions (1)–(6) are satisfied, then all optima lie on the possibility frontier of the aggregate production-*cum*-transaction set $A = \sum_1^{F+1} A^f$, i.e., efficiency of production-*cum*-transaction plans is optimal. Furthermore, if the sets A^f can be decomposed into separate transaction and production sets, then pure production efficiency is optimal.

Efficiency Theorem 1 requires further interpretation. First, why is it that we speak of *production-cum-transaction* efficiency rather than pure production efficiency? This is because the model allows for tradeoffs between pure production efficiency and pure transaction efficiency. This tradeoff arises at the firm level and consequently also appears at the aggregate level. For example, if production activities are not separable from transaction activities, firm 1 may be slightly less efficient in production than firm 2, while firm 1 is substantially more efficient in transacting, so that 1 rather than 2 should be in operation at the welfare optimum.

Suppose that assumptions (1)–(6) hold and the private production sector is already operating efficiently. Then, the government production-transaction manager should operate to ensure overall production-transaction efficiency. The reason for this is that costs g are continuous in the allocation vector $\bar{x}$ under our assumptions. That is, small changes in $\bar{x}$ can be accomplished by small changes in g. Therefore, the resources needed to facilitate a (sufficiently small) Pareto-improvement in the consumer allocation vector are always available if the economy is operating in the interior of the set A.

Further suppose that firm-specific prices are also possible but that it is always more costly to administer a system of firm-specific prices than any system in which producer prices are uniform. Then, we have established the desirability of efficiency *within* the private sector.[4] This as-

[4] See the Appendix available from the authors.

sumption, however, is not realistic, and consequently our result on efficiency within the private sector is of limited applicability as can be seen from the case of intermediate-goods taxation.

In our model, intermediate-goods taxation leads to producer prices which differ across firms. With a positive tax on a particular commodity, the firms producing that commodity face a lower price than the firms buying the commodity for use as an input. It may well be administratively less costly to tax all transactions in the same commodity at the same rate regardless of destination (whether to a household or to a firm) than to administer a system in which transactions between firms are untaxed while sales from firms to households are taxed. Therefore, intermediate-goods taxes, being less costly to administer, may well be employed at an optimum, so that inefficiency is optimal. This contrasts with Diamond and Mirrlees, who find that intermediate-goods taxation is undesirable because administration is costless in their model.[5]

It is important to emphasize the crucial role of assumption 3 in establishing Efficiency Theorem 1. The restrictions on the geometry of G are seemingly mild—allowing for decreasing and constant returns and most but not all forms of increasing returns. In particular, maximality of frontiers rules out "spikes" or set-up costs. In Figure 1, we exhibit a set G not satisfying assumption 3. The horizontal axis "represents" the vector $(\bar{x}, \bar{y}, m, n, \beta)$, while g is measured on the vertical axis. G is shaded and the possibility frontier of G is shown by heavy ink. The "spike" of Figure 1 might be thought of as representing the substantial reduction of costs

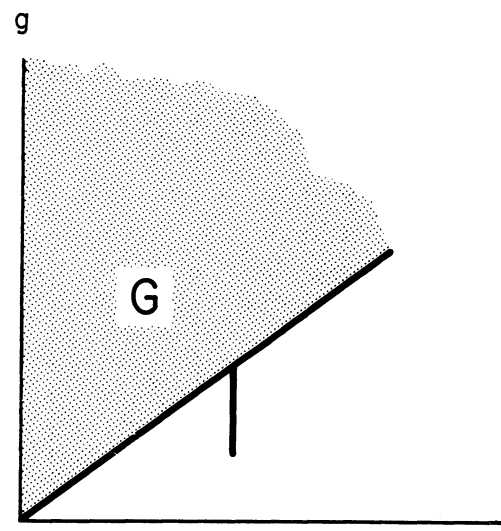

FIGURE 1

due to administering a simple tax program. Consider, for example, a simple across-the-board 5 percent-sales-tax regime which makes some target allocations relatively cheap to attain because of ease of enforcement and computation. The 5 percent system is substantially easier to administer than a system of sales taxes with rates varying over commodities but lying between, say, 4.99 percent and 5.01 percent. Moving to nearby allocations requires a discrete jump in real administrative resources g. These "spikes" or set-up costs are contrary to the crucial assumption of the maximality of the possibility frontier of G and thus provide possible examples of the desirability of production inefficiency.

Mirrlees puts forward another example in which inefficiency is optimal. An inefficient, decreasing-returns firm is owned by deserving households and thus should be operated because of the profits it distributes to those households. Our efficiency theorem denies the desirability of this inefficiency because of the inclusion of licensing subsidies. The Mirrlees inefficiency example, nonetheless, reappears in our model when assumption 3 is relaxed to allow for "spikes" in G. In particular, if

[5] Other examples from the policy literature can be cited. Production inefficiency may be desirable in poor countries because tariffs are much easier to administer than income taxes. Similarly, socialism is sometimes thought to be desirable because income redistribution is much more costly under capitalism, even though private ownership may be productively more efficient.

there is a discontinuous jump in administrative costs when the government moves from a system without license fees or subsidies to one with firm-specific license fees and subsidies, then inefficiency may be desirable.

If assumptions (1)–(6) do not hold, it may be the case that the optimum production-*cum*-transaction plans are not efficient. However, at the optimum, costs versus gains are balanced for each of the tax instruments, so we might expect at first blush that the optimal plan would at least lie on the efficiency frontier of the aggregate production-*cum*-transaction-*cum*-administration set (the set of feasible aggregate production-transaction plans net of administrative costs). This is not necessarily the case as the following simple result makes clear.

Proposition. Consider the basic model above except that assumptions (1)–(6) need not necessarily hold. For the economy to exhibit production-*cum*-transaction-*cum*-administration efficiency, it is necessary that both production-*cum*-transaction efficiency and administrative efficiency obtain.

The proposition follows directly from the separability of production-*cum*-transaction plans from administrative plans. So far, there is no tradeoff between production-transaction efficiency and administrative efficiency, although there may be a tradeoff between production-transaction-administrative efficiency and distortions in demand caused by commodity taxation. We can go further and alter the underlying assumptions to consider cases in which production-*cum*-transaction efficiency is not a property of the optimum, but in which production-*cum*-transaction-*cum*-administrative efficiency is a property of the optimum.

One important case is where administrative costs depend directly on production-transaction plans. In discussing Efficiency Theorem 1, we argued that if the economy is operating in the interior of the set A, a feasible Pareto-improving change can be made in consumer allocations because of the continuity of administrative costs with respect to consumer allocations. If, however, it is administratively costly to instruct the government production-transaction sector to seek overall production-transaction efficiency, then the desirability of production-transaction efficiency would be in doubt. This is because the administrative costs in moving to the frontier of A could outweigh the production gains in the move.

Consider another example in which production-transaction plans directly affect administrative costs. Say that firm I is slightly more efficient in production and transaction than firm II, but assume that it is very much more costly for the government to tax firm I than firm II. In this case, it may be desirable to shut down I while encouraging II to operate. In these examples, we are faced with a tradeoff between resources made available by a particular production-transaction plan and the administrative resources used up in accomplishing the particular plan. The next theorem shows that, at the social optimum, this tradeoff is taken into account.

Efficiency Theorem 2. Let assumptions (1)–(3) and (5)–(6) hold except that government administrative costs may also depend on the production-transaction plans of individual firms. Replace assumption (4) with the simpler assumption that marginal costs of administering licensing fees and subsidies are zero or negligible. Under these conditions, even though production-*cum*-transaction efficiency may not be a property of the optimum, efficiency of production-*cum*-transaction-*cum*-administrative plans is optimal.

III. Concluding Remarks

The above analysis is couched in terms

of a "reduced form" administrative feasibility set, G, where the real resources required for administration are directly related to the target levels of the overall allocation vector and thus administrative costs are only indirectly related to the set of commodity taxation instruments. Such direct effects should be incorporated; we plan to do so. We also plan to delve more deeply into the mechanism of taxation. Our thoughts on how to accomplish this are motivated by the following observation. Under a 5 percent sales tax regime, the government acts to encourage the opening of markets in which the 5 percent tax is collected and acts to discourage markets in which the tax is not collected or collected at a rate other than 5 percent. Thus, administrative costs can be related to the costs of government actions in closing down some markets and opening others. Further substantial progress in this area may have to await the development of a general equilibrium theory of markets in which the costs of exchange depend on the mode of exchange.

REFERENCES

P. Dasgupta and J. E. Stiglitz, "On Optimal Taxation and Public Production," *Rev. Econ. Studies,* Jan. 1972, *39,* 87–103.

P. A. Diamond and J. A. Mirrlees, "Optimal Taxation and Public Production," *Amer. Econ. Rev.,* March and June 1971, *61,* 8–27 and 261–278.

D. K. Foley, "Economic Equilibrium with Costly Marketing," *J. of Econ. Theory,* Sept. 1970, *2,* 276–291.

F. H. Hahn, "Equilibrium with Transactions Costs," *Econometrica,* May 1971, *39,* 417–440.

———, "On Optimum Taxation," *J. of Econ. Theory,* Feb. 1973, *6,* 96–106.

J. A. Mirrlees, "On Producer Taxation," *Rev. Econ. Stud.,* Jan. 1972, *39,* 105–111.