Prices versus Quantities Reconsidered

Nicholas Brozović
University of Illinois at Urbana-Champaign
Department of Agricultural and Consumer Economics
326 Mumford Hall, MC-710
1301 West Gregory Drive
Urbana, IL 61801
nbroz@uiuc.edu

David L. Sunding
University of California, Berkeley
Department of Agricultural and Resource Economics
sunding@are.berkeley.edu

David Zilberman
University of California, Berkeley
Department of Agricultural and Resource Economics
zilber@are.berkeley.edu

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Denver, Colorado, August 1-4, 2004

ABSTRACT

In comparing second-best prices and quantities, studies assume that quantities bind with probability one. We present a more general and realistic model of second-best regulation where quantity instruments can bind with probability less than one. This additional flexibility of quantity instruments makes them much more efficient than previously realized.

INTRODUCTION

Our profession's fondness for the use of economic incentives notwithstanding, regulation via direct control is commonplace in the real world. Diverse consumer and producer activities, from speeds on public thoroughfares to the emission of industrial pollutants from smokestacks, are controlled using maximum allowable limits. Moreover, even though there is considerable heterogeneity amongst regulated individuals and industries, most standards are applied uniformly and consistently. Such uniform standards, which are second-best policies under heterogeneity, may not be binding on all members of the regulated group. To continue with our previous examples, so-called 'Sunday drivers' choose to drive far below the speed limit, while many industrial producers emit pollutants at rates less than their mandated limits. Hence, many quantity regulations in place today stipulate an upper bound on an agent's activity, rather than an absolute level for it. Almost all existing economic studies fail to capture this potential diversity of behavioral response to quantity regulation. Instead, it is implicitly assumed that quantity instruments represent centralized control by diktat: a regulator announces a level of activity, and this level is then dutifully attained, without regard for prevailing economic incentives.

This paper presents a theoretical analysis of quantity regulation under heterogeneity where variable responses can occur at the individual level. The main results are derived under simple assumptions that can be solved graphically. We show that quantity regulation, when

viewed as an upper bound rather than an absolute level, offers additional flexibility to both the regulator and the regulated group. We develop general conditions – much broader than those suggested by previous studies – under which quantity regulation is preferred to price regulation.

In Section 2 of this paper, we review the existing literature on regulation and instrument choice under heterogeneity. The following section presents the general model used. In Section 4, we consider conditions under which quantity regulation can lead to variable individual response. Analytical rankings for quadratic functional approximations are derived in Section 4. These expressions are used to explore the conditions under which each policy instrument is optimal. Finally, the last section provides a brief discussion and conclusion.

2. REGULATION UNDER UNCERTAINTY AND HETEROGENEITY

Optimal regulation under conditions of heterogeneity requires contingent schemes that differentiate each type of economic agent and each state of nature (Arrow, 1969, Tietenberg, 1974, Baumol and Oates, 1988). With a delegation approach, the regulator imposes a payment function that induces optimal agent behavior (Roberts and Spence, 1976). Alternatively, a regulator may use a revelation mechanism in a regulatory scheme that requires truth-telling on the part of heterogeneous agents (Kwerel, 1977, Dasgupta *et al.*, 1980). However, informational requirements and political considerations often disallow contingent regulation. Indeed, even though most regulatory problems are characterized by significant uncertainty or heterogeneity, most current state and federal regulations are uniform rather than contingent in nature (Russell *et al.*, 1986).

In a second-best setting, uniform price and quantity instruments will no longer have the same distributional and welfare effects. The conditions under which one type of instrument is

superior to the other have been the subject of considerable debate over the last three decades. Weitzman (1974) was the first to derive analytical expressions for the welfare ranking of uniform price and quantity instruments under uncertainty. His original result – that the optimal second-best policy depends only on the relative slopes of the marginal cost and marginal benefit functions – is so powerful and elegant that it is taught in virtually all undergraduate and graduate economics curricula. Weitzman's model used quadratic approximations to the marginal cost and benefit functions, assumed no correlation between cost and benefit uncertainty, and implied uniform supply response by producers to unit increases in taxation. Over the last three decades, Weitzman's results have been extended and discussed in countless papers (Cropper and Oates, 1992, provide a convenient review). Major assumptions have been relaxed in theoretical studies (e.g. Roberts and Spence, 1976, Laffont, 1977, Malcomson, 1978, Stavins, 1996), but empirical applications have also been undertaken (e.g. Nichols, 1984, Kolstad, 1986). However, these studies all implicitly assume that optimal uniform standards will bind under all states of nature.

We are aware of only two papers in this large literature that consider the possibility that uniform standards may behave as upper bounds rather than binding in all states of nature. Hochman and Zilberman (1978) analyze how taxes and standards affect aggregate output, input use and pollution in industries characterized by distributions of fixed proportions production and pollution functions. With a fixed proportions production function, quantity regulation always takes the form of an upper bound. Hochman and Zilberman show that the profitability of heterogeneous microunits will depend on the type of policy instrument employed. However, because a fixed proportions production function is used, adjustment to regulation occurs only at the extensive margin. Each microunit has a binary response to regulation: it may choose to operate, or to shut down. Individual production units cannot adjust at the intensive margin in

response to regulation. Wu and Babcock (2001) consider firm adjustment at both intensive and extensive margins for the problem of second-best regulation under heterogeneity. Although they do recognize the possibility of upper bound-type quantity regulation under heterogeneity, they neither derive analytical results for the optimality of corner solutions nor recognize the potential occurrence of multiple local optima under quantity regulation.

Finally, several authors (e.g. Besanko, 1987, Helfand, 1991) have noted that direct controls may be implemented as a wide variety of regulations: as levels of input use, output, or pollutant emissions, as maximum pollution-output ratios, or as mandates on the use of pollution control technology. While all of these options represent direct control, each will have slightly different output, welfare and distributional effects. However, once again, these studies assume that standards represent absolute levels of agent behavior rather than upper bounds. Conversely, the present study explores conditions under which overcompliance with uniform standards is not only rational, but may also be a desirable feature of regulation.

In the above discussion, we have used the terms 'uncertainty' and 'heterogeneity' as though they were interchangeable. As noted above, some portions of the existing literature study regulation with heterogeneity, whereas other portions are concerned with regulation under uncertainty. Before proceeding, we will comment on the equivalence of heterogeneity and uncertainty for the particular set of issues discussed in this paper. With uncertainty, only a single state of nature will ultimately occur, but *a priori* we can only make probabilistic statements about the expected outcome. Under heterogeneity, all possible states of nature occur simultaneously and the average effect is observed. Second-best regulation with *ex ante* uncertainty and second-best regulation with spatial heterogeneity are formally equivalent. In both cases, the regulator must choose a single policy instrument that is only optimal in an

average sense. In an uncertain setting, the actual outcome will not in general be equal to the expected outcome. Conversely, under a heterogeneous system, the outcome is deterministic and uniquely defined. Stylistically, regulation under uncertainty often assumes a single firm or an industry composed of homogeneous firms. In modeling heterogeneity, the regulated industry is taken to be composed of firms with a distribution of production or pollution technologies. In this paper, we couch our model in terms of a heterogeneous, polluting industry. However, the same analytical framework applies equally well to the case of producer or consumer regulation under uncertainty. Note also that in Weitzman's original article, as in much of the following work, the problem of second-best regulation under uncertainty is expressed as a tradeoff between a social benefit function and a production cost function. Our own preference is to present a model in terms of production and damage functions, but the two approaches are completely interchangeable.

3. THE MODEL

An industry is a unit continuum of firms, each with a production technology and a pollution technology. The production technology captures the net surplus of a given level of production activity by each firm. For example, the production technology could represent the value of the marginal product of input use net of all input costs. Conversely, the pollution technology captures the net costs of each firm's activity that are not internalized in the firm's decision-making process. In an agricultural setting, the pollution technology could involve fertilizer runoff into rivers, pesticides leaching into groundwater, or downstream siltation caused by soil erosion. We assume that production and pollution technologies are independent attributes of each firm. For example, the net marginal product of fertilizer use may represent the production technology and the marginal social cost of fertilizer runoff may represent the

pollution technology. Alternatively, production technology may be a function of capital vintage, whereas pollution technology may be related to geographical location.

Possible production technologies are contained within the set *I*. For simplicity, assume that there are only two production technologies, so that $I = \{L, H\}$ and the industry comprises low (*L*) and high (*H*) productivity firms in proportions θ and $(1-\theta)$ respectively. Similarly, pollution technologies are contained within the set *J*. There are two possible pollution technologies, $J = \{C, D\}$, so that the industry contains clean (*C*) and dirty (*D*) firms in proportions η and $(1-\eta)$ respectively. There are thus four types of firms in the industry, contained in the set $I \times J = \{LC, LD, HC, HD\}$. The choice of discrete, rather than continuous, distributions of production and pollution technologies allows both tractable analysis and graphical solution. In particular, the assumption of two discrete production technologies allows the support of the net surplus function to be partitioned into two sets, in each of which the net surplus function is concave. Our major results may be extended to the case of continuous production and pollution technologies, but at the expense of clarity.

Each firm with production technology $i \in I$ uses a scalar input, $x_i \in \mathbb{R}_+$, in the production of a numeraire good. Each firm's quasi-rents net of input prices are given by the production function $f_i(x_i)$. The industry faces a horizontal output demand curve. While limiting, this assumption is commonplace in the literature on regulation under uncertainty and heterogeneity (e.g. Weitzman, 1974, Helfand, 1991). Production functions are continuously differentiable and satisfy $f_i(x_i) > 0, f_i'(x_i) > 0, f_i''(x_i) < 0$ and $f_i(0) = 0$. Moreover, we assume

that the marginal productivity of *H*-type firms is higher than that for *L*-type firms for all output levels, so that $f_L'(x) \leq f_H'(x)$ for all positive values of *x*.

Input use by each firm causes a negative externality. For a firm with production technology *i* and pollution technology $j \in J$, the damage caused by using x_i units of input is given by $g_j(x_i)$, where we assume that $g_j(x_i) > 0, g_j'(x_i) > 0, g_j''(x_i) > 0$ and $g_j(0) = 0$. We also assume that $g_C'(x) \leq g_D'(x)$ for all positive values of *x*, and that $f_L'(0) > g_D'(0)$, so that it is socially desirable for both low and high productivity firms to operate at some scale. This assumption implies that no firms within the industry will shut down as a result of regulation, and adjustment will take place at the intensive margin only.

Net surplus is given by the sum of quasi-rents and damages caused by production,

$$E_i\left[f_i(x_i) - E_j\left[g_j(x_i)\right]\right] \tag{1}$$

where $E_i[\cdot]$ is the expectation operator taken over realizations of the production technology and $E_j[\cdot]$ is the expectation operator taken over realizations of the pollution technology.

In the absence of any regulation, each firm will maximize quasi-rents by using an input level x_i^c such that $f_i'(x_i^c) = 0$, so that $x_L^c \leq x_H^c$. This means that without regulation, high productivity firms will, on average, produce more pollution than low productivity firms. Because $g_j''(x_i) > 0$, high productivity firms will also produce more pollution per unit of input than low productivity firms. Note, however, that clean, high productivity firms may produce less total pollution and less pollution per unit of input than dirty, low productivity firms. Which firm type produces more pollution per unit of output will depend on the functional form of the production function.

We assume that the regulator is constrained to use a single, uniform, instrument. Under such conditions, the welfare effects of second-best price and quantity instruments will, in general, differ.

3.1. *Uniform price controls*

The regulator seeks a uniform price instrument, t^*, which will maximize net surplus, defined by

$$t^* = \arg\max_{t \in \mathbf{R}_+} E_i\left[f_i\left(x_i\left(t\right)\right) - E_j\left[g_j\left(x_i\left(t\right)\right)\right]\right] \tag{2}$$

where $x_i\left(t\right)$ is the input use decision by a firm of type *i* facing price instrument *t*. The price instrument represents the per-unit tax on the input x_i. We assume that tax revenues are recycled in a non-distorting fashion and thus net out of equation (2).

The first order condition corresponding to equation (2) is

$$E_i\left[f_i'\left(x_i\left(t^*\right)\right)\cdot\frac{\partial x_i}{\partial t}\left(t^*\right) - E_j\left[g_j'\left(x_i\left(t^*\right)\right)\cdot\frac{\partial x_i}{\partial t}\left(t^*\right)\right]\right] = 0 \tag{3}$$

where $\partial x_i\left(t\right)/\partial t$ is the marginal response of a firm of productivity type *i* to a unit increase in the price instrument from its current level, *t*. Each firm must take the price instrument into account in its own production decision, and will choose its input use so that $f_i'\left(x_i\left(t\right)\right) = t$. Thus, each firm equates the value of its marginal product and the per-unit tax. Under the price instrument, more efficient firms will continue to use more inputs than less efficient firms, so that $x_L\left(t\right) \le x_H\left(t\right)$. As before, high productivity firms will on average produce more pollution per firm and more pollution per unit of input than low productivity firms. Once again, the relative magnitudes of pollution per unit of output by each firm type will depend on the functional form of the production function. Note also that if the regulator chooses a system of input-use quotas that are transferable on a one-to-one basis then input use, output and pollution for each firm type,

as well as net surplus, will be identical to that for the optimal uniform price instrument. Because of this, transferable quota schemes will not be considered further in this paper.

3.2. *Uniform quantity instruments as absolute levels*

If the regulator decides to address the production externality using a prescribed absolute level of input use, then the choice of an optimal uniform quantity is given by

$$\bar{X} = \arg\max_{X \in \mathbf{R}_+} E_i\left[f_i(X) - E_j\left[g_j(X) \right] \right] \tag{4}$$

Equation (4) is the usual formulation of the choice of uniform quantity instrument under uncertainty or heterogeneity (e.g. Weitzman, 1974, Laffont, 1976, Stavins, 1996). It is implicitly assumed that the optimal instrument $\bar{X}$ will bind on both high and low productivity firms. The formulation of the quantity regulation problem as a choice of absolute and enforced input or output levels, as expressed in equation (4), is the approach taken in the vast majority of existing studies. The first order condition corresponding to equation (4) is

$$E_i\left[f_i'(\bar{X}) \right] = E_j\left[g_j'(\bar{X}) \right] \tag{5}$$

Thus, the optimal uniform quantity instrument will equate the expected value of the marginal product of input use with the expected marginal damage. A comparison of equations (3) and (5) shows that optimal uniform price and quantity instruments will generally result in different input use decisions.

3.3. *Uniform quantity instruments as upper bounds*

If there is enough heterogeneity between firm types, the assumption that the optimal quantity instrument binds on all types of firm need no longer hold. In this case, the choice of the optimal uniform quantity instrument will be given by

$$\tilde{X} = \arg\max_{X \in \mathbf{R}_+} E_i\left[f_i\left(x_i(X)\right) - E_j\left[g_j\left(x_i(X)\right) \right] \right] \tag{6}$$

where $x_i(X)$ is the input use decision by a firm with production technology *i* facing quantity instrument *X*. Now, from the original assumptions about the two different firm types, we know that any optimal quantity instrument will always bind on high productivity firms. Thus, the optimal input use decision of each firm will be given by

$$x_L(\tilde{X}) = \begin{cases} x_L^c & \text{if } f_L'(\tilde{X}) \le 0 \\ \tilde{X} & \text{if } f_L'(\tilde{X}) > 0 \end{cases} \qquad (7)$$
$$x_H(\tilde{X}) = \tilde{X}$$

Using these definitions and equation (6), the first order condition for optimality of the uniform quantity instrument falls into one of two cases.

Case 1. $\tilde{X}$ binds on both types. If $f_L'(\tilde{X}) > 0$, the quantity instrument will bind on both high and low productivity firms. The relevant first order condition is equation (5), so that $\tilde{X} = \bar{X}$ as defined in equation (4).

Case 2. $\bar{X}$ binds only on efficient types. If $f_L'(\tilde{X}) \le 0$, the quantity instrument $\tilde{X}$ will bind only on high productivity firms. The relevant first order condition is then given by

$$f_H'(\tilde{X}) = E_j\left[g_j'(\tilde{X})\right] \qquad (8)$$

In this case, the regulator targets the quantity regulation solely on the subset of high productivity firms, allowing low productivity firms to operate unconstrained.

The difference between uniform quantity instruments as absolute levels and as upper bounds is easily demonstrated graphically. Figure 1 represents the regulator's problem, showing the tradeoff between social costs and private benefits. The marginal values of the low and high productivity technologies are shown by $f_L'(x)$ and $f_H'(x)$ respectively. The expected marginal

product is given by $E_i\left[f_i'(x)\right]$. Three possible distributions of pollution technologies, denoted A, B, and C, are represented in Figure 1. The expected marginal damage under the dirtiest pollution technology, A, is given by $E_A\left[g_A'(X)\right]$ for output level *X*. The cleanest technology is technology C; B represents an intermediate level.

It is clear from Figure 1 that under pollution technology A, the only quantity instrument that satisfies first order condition (5) is an absolute level of input use given by $\overline{X}_A$. Under pollution technology C, although $\overline{X}_C$ satisfies first order condition (5), low productivity firms will never operate where their marginal product is negative. Instead, they will use $\hat{X}$ units of input under a quantity regulation of $\overline{X}_C$ (Figure 1). But in this case, net surplus can be strictly improved by increasing the quantity regulation to $\tilde{X}_C$, thus targeting the regulation to high productivity types. Hence, pollution technology C corresponds to Case 2 and first order condition (8) above. Finally, under pollution technology B, the regulator has a choice of two quantity instruments that satisfy local optimality conditions (Figure 1). The quantity instrument $\overline{X}_B$ satisfies condition (5) and is an absolute level-type regulation that binds on all firm types. Quantity instrument $\tilde{X}_B$ satisfies (8), binding on high productivity types but allowing low productivity types to operate at the unconstrained input use level $\hat{X}$. Thus, the net surplus function has two local maxima. However, the assumption of only two production technologies allows the support of the net surplus function to be partitioned into two sets, with a discontinuity at $\hat{X}$. Each portion of the net surplus function then contains a unique, well-behaved maximum. If multiple local maxima exist, the choice of a global maximum will depend on the functional forms of the production and damage functions. Similarly, if production technologies are a distribution rather than a binary set, it is no longer possible to partition the net surplus function

into well-behaved subsets, although a unique global optimum that satisfies the relevant first order conditions will still exist.

If only Case 2 above holds, the regulator will equate the value of the marginal product of input use for high productivity firms with the expected marginal damage of those firms. Low productivity firms will be allowed to operate unconstrained. What are the advantages of this kind of quantity regulation? The benefits of targeting regulations specifically for the high productivity firms may more than compensate for the additional damages caused by allowing low productivity firms to operate unconstrained. Upper bound-type regulation will yield higher input use and higher outputs than absolute level-type regulation for both high and low productivity firms. Additionally, because low productivity firms are unconstrained and high productivity firms face no tax burden under upper bound-type regulation, there may be a broad base of industry support for such regulation.

Policymakers are particularly interested in how industry-wide input use, output, and total pollution change under the alternative policy instruments described above. A comparison of optimality conditions (3), (5) and (8) suggests that, in general, optimal uniform price and quantity instruments will lead to different total input use amounts, and hence different aggregate output, pollution and net surplus.

4. FEASIBILITY OF UPPER BOUNDS AS REGULATORY INSTRUMENTS

We seek to characterize the relative importance of upper bound-type quantity regulations as compared to absolute level-type quantity regulations. Clearly, if upper bounds are only theoretically feasible under an extremely restricted set of conditions, their practical implications

are limited. However, without specific functional forms, analytical comparisons of the features of each policy instrument are not possible.

To proceed, we make the same assumptions about functional forms as Weitzman (1974), namely that quadratic approximations to the production and damage functions are adequate. Thus, the production and damage functions have the following forms:

$$f_i(X) \approx f_i'(X)X + \frac{f_i''(X)}{2}X^2$$
$$g_j(X) \approx g_j'(X)X + \frac{g_j''(X)}{2}X^2 \tag{9}$$

By following Weitzman's functional forms, we can analyze the potential importance of upper-bound type quantity regulation in a framework familiar to readers and directly comparable to previous work. Additionally, as in Weitzman's study, we assume that:

$$f_L''(x) = f_H''(x) = f''; \quad g_C''(x) = g_D''(x) = g'' \tag{10}$$

Under these assumptions, firm type *i*'s marginal response to a unit increase in the price instrument, $\partial x_i(t)/\partial t$, is constant and given by $\partial x_i(t)/\partial t = 1/f''$, so that equation (3) can be simplified to give

$$t^* = E_i\left[E_j\left[g_j'\left(x_i\left(t^*\right)\right)\right]\right] \tag{11}$$

Optimality conditions (5) and (8) for uniform quantity instruments are unchanged from our previous analysis. With these assumptions about the functional forms of the production and damage functions, it is possible to obtain expressions for the conditions under which multiple solutions to the quantity regulation problem, as well as regulation via upper bounds, are feasible.

In order to simplify the analytical results presented, we introduce several parameters. The absolute level-type quantity regulation, $\bar{X}$ (Figure 1), forms a convenient baseline for

comparison. Thus, define $\gamma = f_H'\left(\bar{X}\right) - f_L'\left(\bar{X}\right)$, so that γ is the difference in marginal products between high and low productivity firms at the optimum quantity instrument $\bar{X}$, as given by equation (5). Recalling that the proportions of low (*L*) and high (*H*) productivity firms are θ and $(1-\theta)$ respectively, the variance of $f_i'\left(\bar{X}\right)$ is defined as $\sigma^2 = E_i\left[f_i'\left(\bar{X}\right)^2\right] - \left(E_i\left[f_i'\left(\bar{X}\right)\right]\right)^2 = \theta(1-\theta)\gamma^2$. The product $\theta(1-\theta)$ is a measure of the skewness of the distribution of production technologies within the industry. It attains a maximum value when there are equal proportions of low and high productivity firms. Thus, the parameter γ may be thought of as the standard deviation of the marginal productivity at $\bar{X}$ multiplied by the skewness of production technologies within the industry. The parameter $\tilde{\gamma}$ is the difference in marginal products between the two types of firm divided by the marginal product of high productivity firms and is defined as $\tilde{\gamma} = \left(f_H'\left(\bar{X}\right) - f_L'\left(\bar{X}\right)\right) \Big/ f_H'\left(\bar{X}\right) = \gamma \Big/ f_H'\left(\bar{X}\right)$. Finally, the parameter β is defined as the ratio of the elasticity of the expected marginal productivity with respect to input use to the elasticity of expected marginal pollution damage with respect to input use. Thus, noting that by definition $E_i\left[f_i'\left(\bar{X}\right)\right] = E_j\left[g_j'\left(\bar{X}\right)\right]$, β is given by the expression $\beta = -\left(\bar{X} \cdot f''\left(\bar{X}\right) \Big/ E_i\left[f_i'\left(\bar{X}\right)\right]\right) \Big/ \left(\bar{X} \cdot g''\left(\bar{X}\right) \Big/ E_j\left[g_j'\left(\bar{X}\right)\right]\right) = -f''/g''$, which is the negative of the ratio of the marginal product to the marginal pollution damage. The parameter β is always positive, and it captures the relative tradeoff between increased production and increased pollution as input use increases. A value of β of more (less) than unity implies that as a firm's input use increases, its marginal product will decrease by more (less) than the marginal pollution damage increases.

Using these parameters, we can derive conditions for the feasibility of multiple solutions to problem (6), and conditions under which a set input use level or an upper bound are the only possible uniform quantity regulations.

Lemma 1. In the range of $\tilde{\gamma}$ given by $\frac{1+\beta}{1+(1+\theta)\beta} \leq \tilde{\gamma} < 1$, multiple solutions to problem (6) exist. If $\tilde{\gamma} < \frac{1+\beta}{1+(1+\theta)\beta}$, there is only one solution to problem (6) and it involves an absolute level of input use that binds on both high and low productivity firms. If $\tilde{\gamma} \geq 1$, there is a unique solution to problem (6), but in this case it corresponds to an upper bound-type regulation.

Corollary 1. *As β increases (decreases), the range of $\tilde{\gamma}$ over which multiple solutions are possible also increases (decreases).*

Corollary 2. *As θ, the proportion of low productivity firms, increases (decreases), the range of $\tilde{\gamma}$ over which multiple solutions are possible also increases (decreases).*

The results contained within Lemma 1 and Corollaries 1 and 2 are conveniently represented graphically (Figure 2). Each of the three panels of Figure 2 corresponds to a different proportion of low productivity firms, given by θ values of 0.8, 0.5 and 0.2. Feasible solutions for the uniform quantity regulation problem are shown for each of these values of θ in the parameter space of $\tilde{\gamma}^2$ and $\log_{10}\beta$. The range of $\log_{10}\beta$ corresponds to β ranging from approximately one-fifth to five. This covers a broad range of relative elasticities of the marginal product and marginal pollution damage functions.

It is intuitively clear that as the variance in production technology increases, upper bounds are more likely to be feasible (Figure 2). If the variance in production technology

becomes large enough, upper bounds become the only possible quantity instrument. Corollary 1 states that as the marginal productivity becomes relatively more elastic, the feasibility range for upper bounds becomes larger. As β increases, the unconstrained input use level for low productivity firms will decrease relative to $\bar{X}$, making optimal absolute level-type regulation less binding on low productivity firms. Finally, Corollary 2 states that as the proportion of low productivity firms increases, upper bound regulation becomes more feasible. As θ increases, the expected marginal product moves towards the marginal product of low productivity firms, decreasing $\bar{X}$, the optimal absolute level regulation. Although this increases the likelihood that quantity regulation binds on low productivity types, it also increases the separation between $E_i\left[f_i'\left(\bar{X}\right)\right]$ and the high productivity firms, leading to an overall increase in the likelihood that an upper bound is feasible.

Under quadratic damage and production functions and with assumption (10) and the relevant optimality conditions, it is possible to characterize the relative input use, output and pollution of the various possible policies.

Proposition 1. *First-best contingent regulation and second-best price regulation and absolute level-type quantity regulation will each have the same input use. If feasible, upper bound-type regulation will have a higher input use. The uniform price instrument will have higher aggregate output and damage than under the first-best policy, and the uniform quantity instrument will have lower aggregate output and damage than the first-best. Where feasible, upper bound-type regulation will have higher aggregate output and more damage than first-best and absolute level regulation, but this may be more or less than the output and damage with the uniform price instrument.*

Note that if low and high productivity firms respond differently to a unit increase in the price instrument, then the slopes of their marginal benefit functions will differ. In this case, first best, price and quantity instruments will result in different aggregate outputs and Proposition 1 will no longer hold.

5. RANKING OF SECOND-BEST POLICIES

In the preceding sections, we have presented three solution concepts to the problem of second-best regulation under heterogeneity. Two of these, a uniform price instrument, and a quantity instrument that defines an absolute level of input use, have been analyzed extensively in previous theoretical and empirical studies. The third, an upper bound-type quantity regulation that is targeted to high productivity firms but allows low productivity firms to operate unconstrained, has not been considered elsewhere. By definition, each of these second-best instruments involves a welfare loss relative to the first-best regulation. However, as uniform policies are pervasive in the real world, we are particularly interested in the performance of these second-best policies relative to each other. Below, we develop comparative measures for pairs of these policies and analyze the conditions under which each instrument type will be preferred to the others.

5.1. *Prices vs. Absolute Levels*

The choice of price versus absolute level-type quantity instrument is the problem originally analyzed by Weitzman (1974). As shown in Proposition 1, a uniform price instrument leads to a mean-preserving spread of input use relative to absolute level-type regulation, with $x_L\left(t^*\right) < \overline{X} < x_H\left(t^*\right)$. The difference in net surplus between a uniform price instrument and an absolute level-type instrument following equations (11) and (5) respectively is given by

$$w^p - w^s = \theta E_j \left[\int_{x_L(t^*)}^{\bar{X}} \left\{ f_L'(z) - g_j'(z) \right\} dz \right] + (1-\theta) E_j \left[\int_{\bar{X}}^{x_H(t^*)} \left\{ f_H'(z) - g_j'(z) \right\} dz \right] \quad (12)$$

Since the expectation operator is a linear operator, it can be passed through the integral signs to give

$$w^p - w^s = \theta \int_{x_L(t^*)}^{\bar{X}} \left\{ f_L'(z) - E_j\left[g_j'(z) \right] \right\} dz + (1-\theta) \int_{\bar{X}}^{x_H(t^*)} \left\{ f_H'(z) - E_j\left[g_j'(z) \right] \right\} dz \quad (13)$$

As noted by many previous authors from Weitzman (1974) onward, only one type of heterogeneity (in this case the expected marginal product) plays a part in determining the difference in net surplus between price and absolute level-type quantity instruments. The magnitude of the difference between realizations of the damage function is irrelevant here, and only becomes important if there is a correlation between pollution and production technologies at a firm level (Weitzman, 1974, Stavins, 1996).

Under our assumption of quadratic production and damage functions, the difference in net surplus is given by (Figure 3A)

$$w^p - w^s = \frac{\sigma^2 \left(f''^2 - g''^2 \right)}{-2 f''^2 \left(f'' - g'' \right)} = \frac{\sigma^2 \left(f'' + g'' \right)}{-2 f''^2} \quad (14)$$

This is an exact analogue to the well-known expression derived by Weitzman (1974) that states the comparative advantage of a price instrument over an absolute level-type quantity instrument in terms of the relative slopes of the production and damage functions. Given the model presented here, if the marginal production function has a higher slope magnitude than the marginal damage function, a uniform price instrument will yield a higher net surplus. An equivalent condition for superiority of a uniform price instrument over an absolute level-type quantity instrument is that the parameter β is greater than unity.

5.2. *Absolute Levels vs. Upper Bounds*

Comparisons of the two kinds of quantity instrument, absolute levels and upper bounds, are only meaningful if both are feasible, as defined by Lemma 1 (see Figure 3B). If this is the case, then the difference in net surplus between the two instruments is given by

$$w^s - w^u = \theta E_j \int_{\bar{X}}^{\hat{X}} \left\{ f_L'(z) - g_j'(z) \right\} dz + (1-\theta) E_j \int_{\bar{X}}^{\tilde{X}} \left\{ f_H'(z) - g_j'(z) \right\} dz \qquad (15)$$

where for ease of notation, $\hat{X} = x_L^c$, the input use for low productivity firms in the absence of regulation. Once again, linearity of the expectation operator can be used to simplify equation (15) to give

$$w^s - w^u = \int_{\bar{X}}^{\hat{X}} E_i \left[f_i'(z) - E_j \left[g_j'(z) \right] \right] dz + (1-\theta) \int_{\hat{X}}^{\tilde{X}} \left\{ f_H'(z) - E_j \left[g_j'(z) \right] \right\} dz \qquad (16)$$

The first term on the right hand side of equation (16) is always negative and represents the loss in surplus from allowing both the unconstrained operation of low productivity firms and higher input uses (calculated over the range from $\bar{X}$ to $\hat{X}$) for high productivity firms. The second term on the right hand side of equation (16) is positive and represents the gains from targeting an upper bound-type regulation to the subset of high productivity firms. For quadratic production and damage functions, the difference in net surplus becomes

$$w^s - w^u = \frac{\sigma^2 \dfrac{1-\tilde{\gamma}}{\tilde{\gamma}} \left\{ \dfrac{1-\tilde{\gamma}}{(1-\theta)\tilde{\gamma}} \left(f'' - g'' \right)^2 + 2f'' \left(f'' - g'' \right) - \dfrac{\theta \tilde{\gamma}}{1-\tilde{\gamma}} f''^2 \right\}}{-2f''^2 \left(f'' - g'' \right)} \qquad (17)$$

5.3. *Prices vs. Upper Bounds*

If both types of quantity instrument are feasible then the difference in net surplus between the price instrument and the upper bound is simply the sum of the differences between a price

instrument and an absolute level and an absolute level and an upper bound, namely the sum of equations (14) and (17):

$$w^p - w^u = \left(w^p - w^s\right) + \left(w^s - w^u\right)$$
$$= \frac{\sigma^2\left\{\left(f''^2 - g''^2\right) + \frac{1-\tilde{\gamma}}{\tilde{\gamma}}\left\{\frac{1-\tilde{\gamma}}{(1-\theta)\tilde{\gamma}}\left(f'' - g''\right)^2 + 2f''\left(f'' - g''\right) - \frac{\theta\tilde{\gamma}}{1-\tilde{\gamma}}f''^2\right\}\right\}}{-2f''^2\left(f'' - g''\right)} \quad (18)$$

Alternatively, if $\tilde{\gamma} \geq 1$, an upper bound is the only feasible quantity-type regulation. In this case, the appropriate difference in net surplus is given by

$$w^p - w^u = \theta \int_{x_L(t^*)}^{\hat{X}} \left\{f_L'(z) - E_j\left[g_j'(z)\right]\right\} dz + (1-\theta) \int_{\tilde{X}}^{x_H(t^*)} \left\{f_H'(z) - E_j\left[g_j'(z)\right]\right\} dz \quad (19)$$

The first term on the right hand side of equation (19) is the difference in net surplus between a price instrument and unconstrained operation for low productivity firms under an upper bound. This term may be positive or negative (Figure 3C), as the gains from increasing aggregate output for low productivity firms may be more or less than the concurrent increase in aggregate pollution. The second term on the right hand side of equation (19) is the difference in surplus between the price and upper bound instruments for the subset of high productivity firms. Because an upper bound is targeted specifically to high productivity firms, this term is always negative. For quadratic production and damage functions, equation (19) becomes

$$w^p - w^u = \frac{\sigma^2\left\{(1-\theta)f''^2 + \frac{\left(1-(1-\theta)\tilde{\gamma}\right)^2}{(1-\theta)\tilde{\gamma}^2}\left(f'' - g''\right)^2 + 2\left(1-(1-\theta)\tilde{\gamma}\right)f''\left(f'' - g''\right) - g''^2\right\}}{-2f''^2\left(f'' - g''\right)} \quad (20)$$

Unlike the simple and elegant comparison between a price instrument and an absolute level-type quantity instrument expressed in equation (14), when an upper bound is possible, analytical expressions for the differences in net surplus become cumbersome.

By factoring the term g''^2 out of expressions (14), (17), (18) and (20), the signs of the welfare differences may be rewritten using the three previously defined parameters, θ, $\tilde{\gamma}$ and β. Recall that θ is the proportion of low productivity firms and $\tilde{\gamma}$ is the normalized product of the skewness of production technology type and the standard deviation of the marginal product. The parameter β is the negative of the ratio of elasticities of the marginal product and marginal damage functions. In summary, the signs of differences in net surplus follow:

Case 1. If $\tilde{\gamma} < \dfrac{1+\beta}{1+(1+\theta)\beta}$, the only feasible instruments are a uniform price instrument or an absolute level-type quantity instrument.

$$sign\left(w^p - w^s\right) = sign\left(\beta^2 - 1\right) \tag{21}$$

Case 2. If $\dfrac{1+\beta}{1+(1+\theta)\beta} \le \tilde{\gamma} < 1$, all three instruments are feasible, so that

$$\begin{aligned} sign\left(w^p - w^s\right) &= sign\left(\beta^2 - 1\right) \\ sign\left(w^s - w^u\right) &= sign\left(\frac{1-\tilde{\gamma}}{\tilde{\gamma}}\left\{\frac{1-\tilde{\gamma}}{(1-\theta)\tilde{\gamma}}(1+\beta)^2 + 2\beta(1+\beta) - \frac{\theta\tilde{\gamma}}{1-\tilde{\gamma}}\beta^2\right\}\right) \\ sign\left(w^p - w^u\right) &= sign\left(\frac{1-\tilde{\gamma}}{\tilde{\gamma}}\left\{\frac{1-\tilde{\gamma}}{(1-\theta)\tilde{\gamma}}(1+\beta)^2 + 2\beta(1+\beta) + \frac{(1-\theta)\tilde{\gamma}}{1-\tilde{\gamma}}\beta^2\right\} - 1\right) \end{aligned} \tag{22}$$

Case 3. If $\tilde{\gamma} \ge 1$, a price instrument and an upper bound-type quantity regulation are feasible, so that

$$sign\left(w^p - w^u\right) = sign\left((1-\theta)\beta^2 + \frac{\left(1-(1-\theta)\tilde{\gamma}\right)^2}{(1-\theta)\tilde{\gamma}^2}(1+\beta)^2 + 2\left(1-(1-\theta)\tilde{\gamma}\right)\beta(1+\beta) - 1\right) \tag{23}$$

The welfare rankings of regulations defined by expressions (21), (22) and (23) are far more complex than Weitzman's original relative slopes criterion, even though the functional forms used in this analysis are identical. However, for upper bound-type regulation to be

meaningful in a policy context, it is not sufficient to show that it is feasible. There must also exist significant portions of the relevant parameter space over which failure to consider upper bound-type quantity regulation will lead to an incorrect choice of second-best policy instrument. Analytical comparison of expressions (21), (22) and (23) is awkward, but the choice of surplus-maximizing regulation may be presented graphically in a straightforward manner. The panels of Figure 4 show the optimal choice of uniform instrument in the parameter space of $\tilde{\gamma}^2$ and $\log_{10}\beta$ for three values of θ (0.2, 0.5 and 0.8). The parameter space of $\tilde{\gamma}^2$ and $\log_{10}\beta$ and the values of θ used are the same as in Figure 2. For values of $\tilde{\gamma}^2$ greater than unity, the only feasible quantity instrument is an upper bound; a dashed line marks the boundary in each panel. The range of $\log_{10}\beta$ used, around –0.7 to 0.7, implies marginal product functions with input use elasticities from one-fifth to five times that of the marginal damage functions in absolute value.

As is immediately obvious from Figure 4, there is a significant portion of the studied parameter space in which Weitzman's original analysis is not valid. There are two separate reasons for this. First, absolute level-type regulation may not be feasible, so that the appropriate comparison is between upper bounds and prices. Second, even when absolute level regulation is feasible, an upper bound may be the preferred instrument. In either case, if the simple 'relative slopes' rule is applied in these regions, the welfare ranking may be incorrect. In particular, the symmetry of Weitzman's original result is lost, and the relative slope of the marginal production and damage functions no longer uniquely determines the preferred instrument. Moreover, as θ changes, the region in which upper bounds are the preferred quantity instrument changes dramatically.

Proposition 2. *The greater the proportion of low productivity firms, the larger the portion of the parameter space of* $\tilde{\gamma}$ *and* β *over which upper bound-type quantity regulation is the preferred instrument.*

In particular, for large values of θ, corresponding to a high proportion of low productivity firms, upper bounds may be preferable to a price instrument even if the input elasticity of the marginal product is much larger in absolute value than the input elasticity of marginal pollution damage (Figure 4A). In other words, quantity regulation using upper bounds may be preferred to price regulation even if input use is responsive to price incentives and marginal damage varies little with the scale of input use.

If the majority of firms are of the low productivity type, then uniform price and absolute level quantity instruments will disproportionately penalize high productivity firms in terms of quasi-rents. Thus, using an upper bound-type regulation to target high productivity firms may increase net surplus even though input use by the majority of firms moves further away from optimal levels. Conversely, if most of the firms are of the high productivity type, the gains in moving to an upper bound from either of the other instruments are small (Figure 4). In this case, high productivity firms are already close to their optimal level of input use. Allowing low productivity firms to operate unconstrained decreases net surplus by a larger amount than is gained from targeting the regulation to high productivity firms, even though the proportion of low productivity firms is small.

6. CONCLUSIONS

Economic activity is characterized by heterogeneity (Rosen, 2002). This applies as much to consumers and producers as to the externalities arising from their activities. In the presence of

heterogeneity and uncertainty, optimal regulation requires policies that are contingent on each possible state of nature. Practically, however, regulations tend to be consistent and uniform despite heterogeneity. Under such conditions, it is well known that aggregate welfare will vary with the choice of second-best uniform instrument. Within the economics literature, it has been common practice to portray second-best regulation as a choice between a uniform price instrument and a quantity instrument that enforces an absolute and exact level of behavior. The assumption that quantity controls uniformly constrain heterogeneous agents is at odds with observations of real-world regulations, where quantity controls often take the form of upper bounds on admissible activity. Because upper bound-type regulation admits variable individual response, it is inherently more flexible than absolute level-type regulation. Perhaps surprisingly, there have been no previous theoretical studies of the conditions under which upper bound regulation is either feasible or a welfare-maximizing second-best policy.

In this paper, we present a model for the choice of second-best regulation of a polluting industry under heterogeneity that includes consideration of upper bound-type quantity regulation. Employing the same functional forms as Weitzman (1974) allows the results of that classic study to be nested within a more general framework. The present analysis demonstrates that upper bound-type quantity regulations are both feasible and optimal for a broad range of characteristics of agent heterogeneity. In contrast to Weitzman's simple relative slope criterion for determining the optimal second-best instrument, we find that welfare rankings of price and quantity instruments also depend on the magnitude and distribution of producer heterogeneity.

The option of using an upper bound as a regulatory instrument significantly broadens the conditions under which quantity regulation is preferred to price regulation. In particular, our analysis demonstrates that if there are large differences in production technology between firms

and only a small proportion of firms are highly productive, upper bounds will be preferred almost irrespective of the input elasticities of the marginal production and marginal damage functions. The intuition behind this result is that the gains from targeting quantity regulation to firms that are highly productive may outweigh the additional damages caused by allowing low productivity firms to operate unconstrained.

Finally, there are several additional reasons why an upper bound-type quantity instrument may be favored compared to alternative price or quantity instruments. At least for the functional forms used in this paper, upper bounds will lead to higher aggregate input use, which may be a popular political goal. Arguably, all types of firm will prefer an upper bound-type regulation to other instruments. Under an upper bound, low productivity firms are allowed to operate unconstrained. For this subset of firms, an upper bound is clearly preferred to any alternative regulation, as it is equivalent to no regulation whatsoever. Similarly, for high productivity firms, production under an upper bound may approach that possible under a uniform price instrument, but without any associated tax burden. Thus upper bound-type regulations may have a broader base of political support than alternative instruments.

REFERENCES

ARROW, K. J. (1969), "The organization of economic activity: Issues pertinent to the choice of market versus non-market allocation", *The analysis and evaluation of public expenditure: The PPB System* (Washington DC: Joint Economic Committee of the Congress of the United States).

BESANKO, D. (1987), "Performance versus design standards in the regulation of pollution", *Journal of Public Economics*, **34**, 19-44.

CROPPER, M. L., and OATES, W. E. (1992), Environmental economics: A survey, *Journal of Economic Literature*, **30**, 675-740.

DASGUPTA, P., HAMMOND, P., and MASKIN, E. (1980), "On imperfect information and optimal pollution control", *Review of Economic Studies*, **47**, 857-860.

HELFAND, G. E. (1991), "Standards versus standards: The effects of different pollution restrictions", *American Economic Review*, **81**, 622-634.

HOCHMAN, E., and ZILBERMAN, D. (1978), "Examination of environmental policies using production and pollution microparameter distributions", *Econometrica*, **46**, 739-760.

KOLSTAD, C. D. (1986), "Empirical properties of economic incentives and command-and-control regulations for air pollution control", *Land Economics*, **62**, 250-268.

KWEREL, E. R. (1977), "To tell the truth: Imperfect information and optimal pollution control", *Review of Economic Studies*, **44**, 595-601.

LAFFONT, J. J. (1977), "More on prices *vs*. Quantities", *Review of Economic Studies*, **44**, 177-182.

MALCOMSON, J. M. (1978), "Prices *vs*. Quantities: A critical note on the use of approximations", *Review of Economic Studies*, **45**, 203-210.

NICHOLS, A. L. (1984), *Targeting economic incentives for environmental protection* (Cambridge: MIT Press).

ROBERTS, M. J., and SPENCE, M. (1976), "Effluent charges and licenses under uncertainty", *Journal of Public Economics*, **5**, 193-208.

ROSEN, S. (2002), "Markets and diversity", *American Economic Review*, **92**, 1-15.

RUSSELL, C. S., HARRINGTON, W., and VAUGHN, W. J. (1986), *Enforcing pollution control laws* (Washington DC: Resources for the Future).

STAVINS, R. N. (1996), "Correlated uncertainty and policy instrument choice", *Journal of Environmental Economics and Management*, **30**, 218-232.

TIETENBERG, T. H. (1974), "Derived decision rules for pollution control in a general equilibrium space economy", *Journal of Environmental Economics and Management*, **1**, 3-16.

WEITZMAN, M. L. (1974), "Prices *vs*. Quantities", *Review of Economic Studies*, **41**, 477-491.

WU, J., and BABCOCK, B. A. (2001), "Spatial heterogeneity and the choice of instruments to control nonpoint pollution", *Environmental and Resource Economics*, **18**, 173-192.

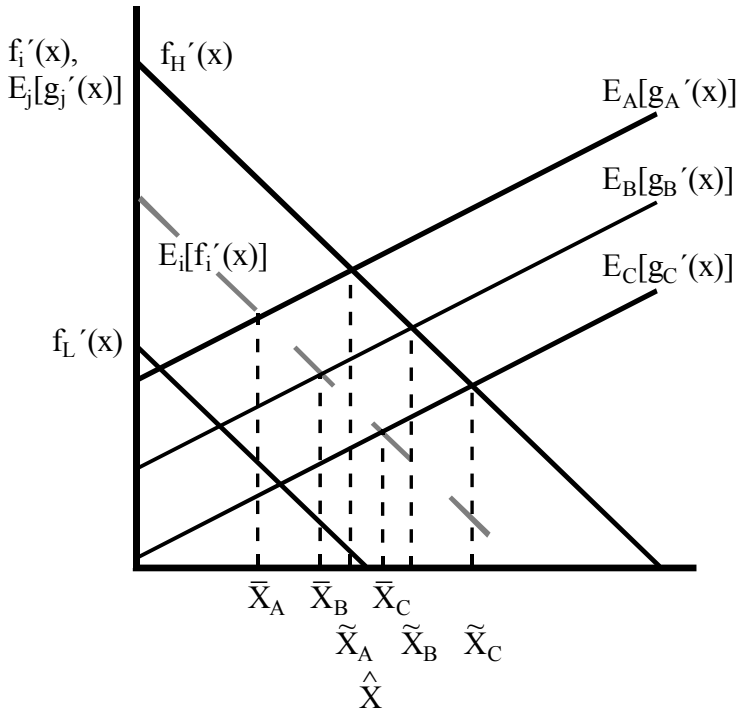


FIGURE 1

The choice of quantity instrument under alternative pollution technologies.

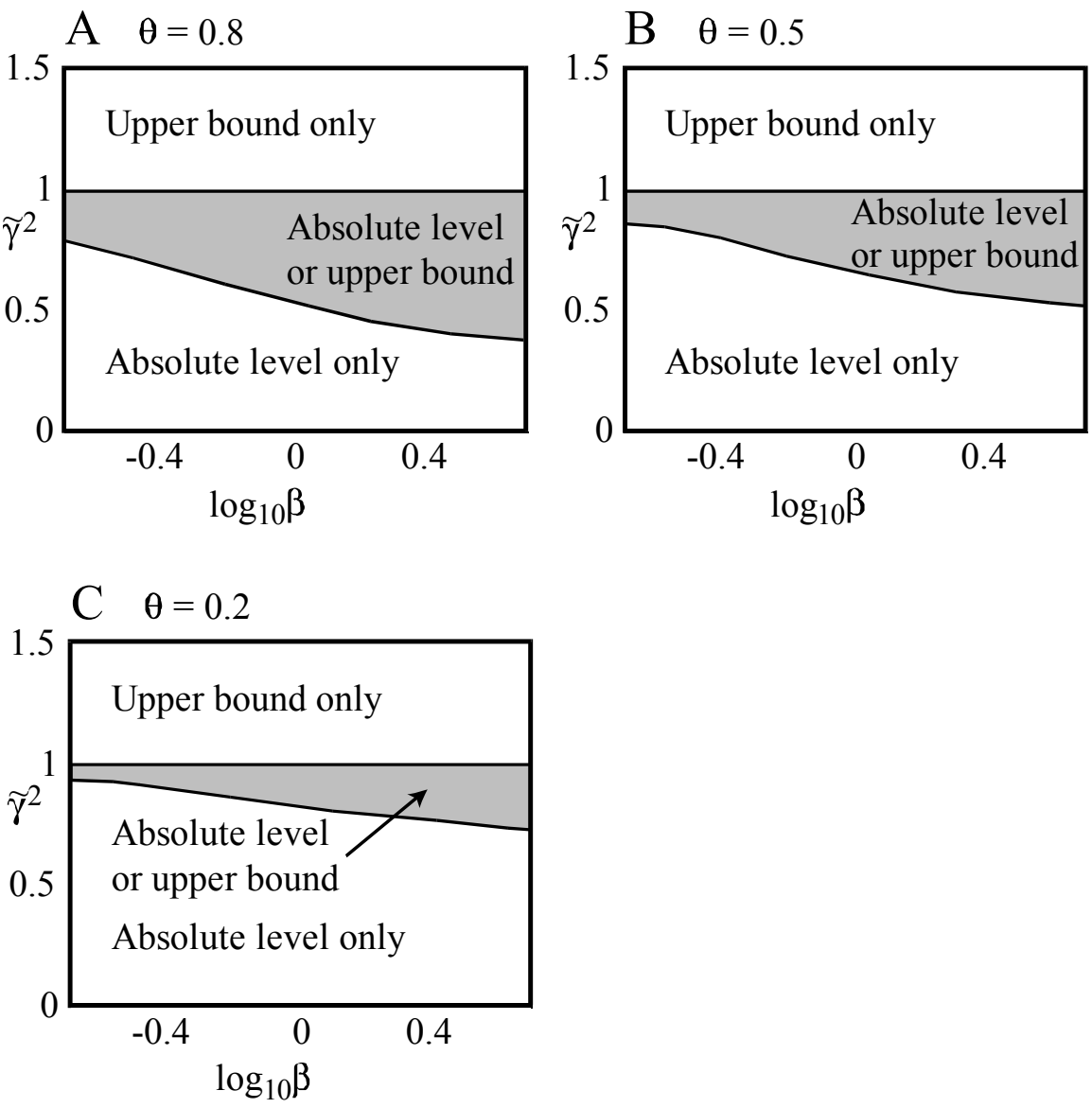
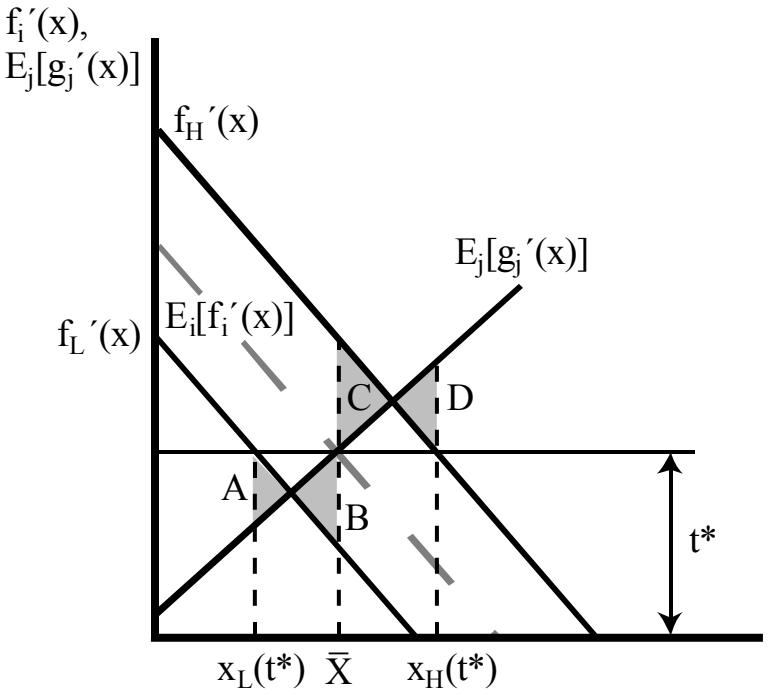


FIGURE 2

Feasibility of absolute level-type and upper bound-type quantity instruments.

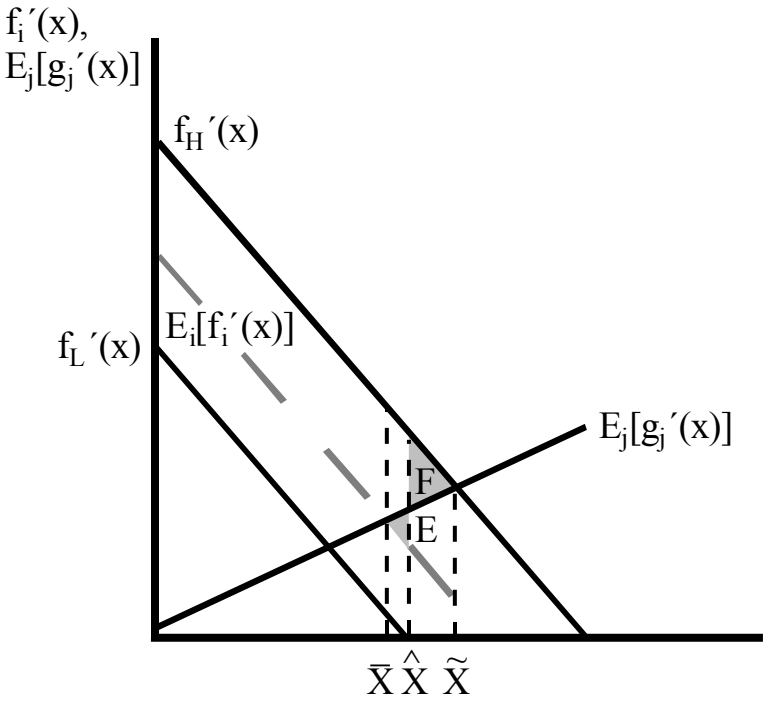


A. Prices vs. Absolute Levels

FIGURE 3A

The choice of prices *vs*. absolute levels under heterogeneity. Notation in this figure follows that in the body of the text. Capital letters refer to the area of adjacent gray triangles. Analytical expressions for these areas are derived in the Appendix. In this panel, upper bound-type quantity regulation is not feasible. The comparative advantage of a uniform price instrument over an absolute level-type quantity instrument is given by

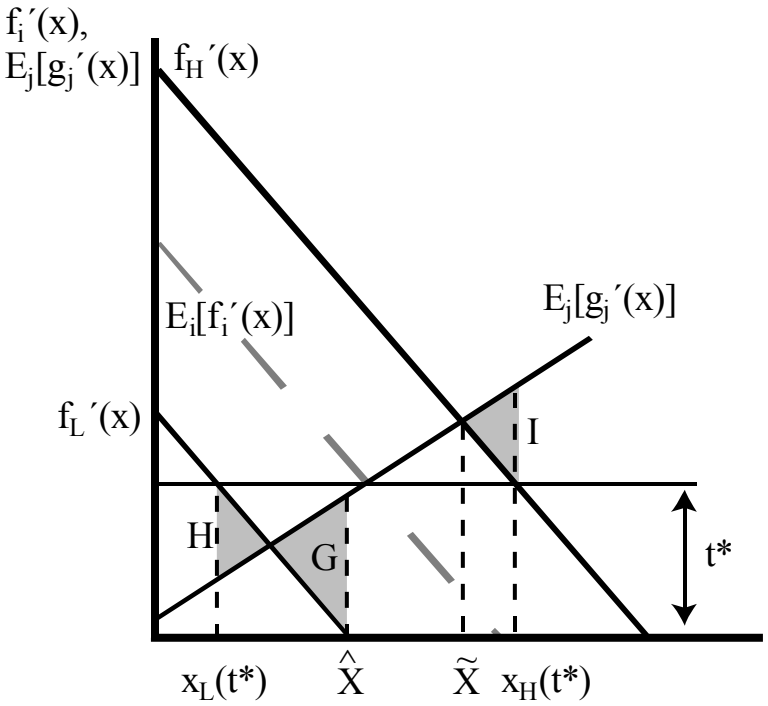
$$w^p - w^q = \theta\left(Area\ B - Area\ A\right) + \left(1-\theta\right)\left(Area\ C - Area\ D\right).$$



B. Absolute Levels vs. Upper Bounds

FIGURE 3B

The choice of absolute levels *vs*. upper bounds under heterogeneity. Notation in this figure follows that in the body of the text. Capital letters refer to the area of adjacent gray triangles. Analytical expressions for these areas are derived in the Appendix. In this panel, multiple solutions to the quantity regulation problem exist. The comparative advantage of an absolute level-type quantity instrument over an upper bound-type quantity instrument is given by $w^q - w^u = Area\ E - (1-\theta)\ Area\ F$.



C. Prices vs. Upper Bounds

FIGURE 3C

The choice of prices *vs*. upper bounds under heterogeneity. Notation in this figure follows that in the body of the text. Capital letters refer to the area of adjacent gray triangles. Analytical expressions for these areas are derived in the Appendix. In this panel, absolute level-type quantity regulation is not feasible. The comparative advantage of a uniform price instrument over an upper bound-type quantity instrument is given by

$$w^p - w^u = \theta\left(Area\ G - Area\ H\right) - \left(1-\theta\right) Area\ I\ .$$

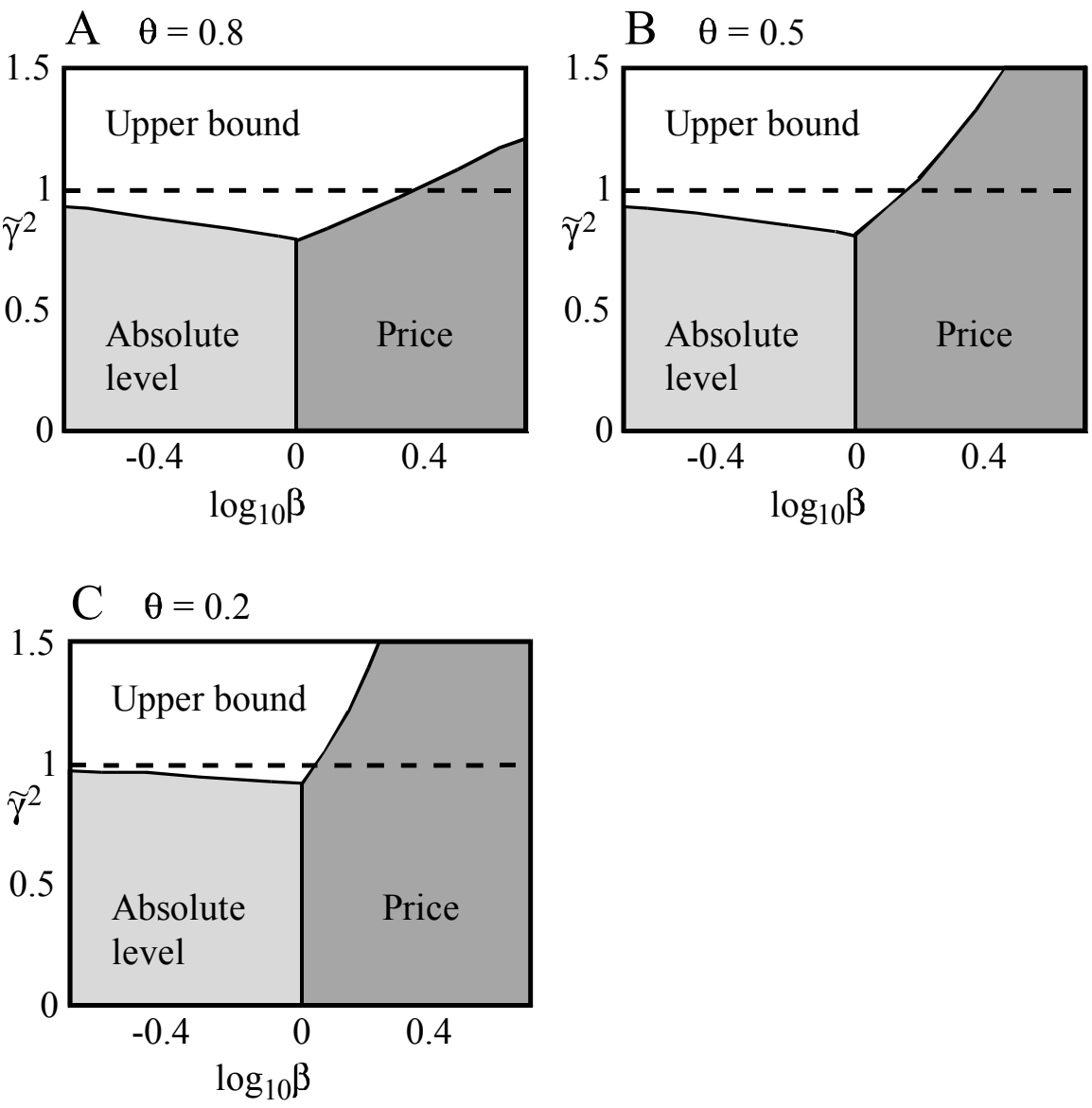


FIGURE 4

The choice of second-best instrument under heterogeneity.