Optimal Regulation with Exemptions and Corrective Taxes

Louis Kaplow*

Abstract

Regulation produces enormous benefits and costs, both of which are greatly influenced by myriad exemptions and preferences for small firms that contribute a significant minority of output in many sectors. These firms may generate a disproportionate share of harm due to their being exempt and because exemption induces additional harmful activity to be channeled their way. This article analyzes optimal regulatory exemptions where firms have different productivities that are unobservable to the regulator, regulated and unregulated output each cause harm although at different levels, and regulation and the exemption level affect entry and the output choices of regulated and unregulated firms. It also analyzes the optimal use of output taxation alongside regulation — that is, optimal regulation with taxation, in contrast to the traditional comparison of regulation versus taxation. In many settings, optimal schemes involve subtle effects and have counterintuitive features: for example, incentives of firms to drop output to become exempt can be too weak as well as too strong, and optimal output taxes may equal zero despite the presence of externalities. When all instruments under examination are admitted, a planner can achieve the first best, and in this regime optimal regulation is voluntary.

JEL Classes: D61, D62, H23, J88, K20, K23, K32, K42, L51, Q58

Keywords: regulation, exemption, corrective taxation, cost-benefit analysis, externalities, small business

*Harvard University and National Bureau of Economic Research. I am grateful to Nathan Hendren, Steven Shavell, Andrei Shleifer, Joel Slemrod, Robert Stavins, and participants in workshops at Chicago, Michigan, Northwestern, and Stanford for discussion and comments, David Choi, Andrea Lowe, and Nick Warther for research assistance, and the John M. Olin Center for Law, Economics, and Business at Harvard University for financial support.

1. Introduction

Regulation is ubiquitous in developed economies, and both its benefits and costs are immense. Accordingly, cost-benefit analysis has been the subject of substantial study with subtle refinement, but some key issues in regulatory design have received little attention.

One is exemptions or other forms of preferential treatment of small business, often rationalized due to economies of scale in regulatory compliance.[1] Although this topic may seem of secondary importance, we should keep in mind that small business production in many sectors contributes a substantial, if minority share of output. Small business may also generate a disproportionate share of harm precisely because it is exempt from regulation and, moreover, this induces additional harmful activity to be channeled its way.[2] Furthermore, small businesses are not exempt from just one or two regulations but from myriads of them — regarding the environment, workplace safety, hiring, employee benefits, information disclosure, and much more. The aggregate effect can cause both greater harm and larger distortions in production. Additionally, in many developed economies, small business receives tax exemptions (often from the VAT and, in the U.S., from a new $2000-per-employee health mandate penalty) and other benefits (government contract preferences, subsidized loans) that may significantly exacerbate these effects.[3]

A further complication suggested in the foregoing is that regulation in general and exemptions in particular affect firms' output decisions, changing their marginal costs of operation, imposing fixed costs, and creating potentially perverse incentives to remain exempt.[4] Therefore, a proper analysis of regulation and of exemptions must take into account entry/exit decisions and effects on the output of regulated and unregulated firms, each of which causes different levels of harm.[5]

Finally, in light of these activity-level effects and related distortions, it is natural to consider taxation as well: not the traditional comparison — regulation versus taxation — but

---

[1]For information on the small business sector, exemption from regulation and taxation, and the magnitude (if any) of scale economies in regulatory compliance, see Becker et al. (2012), Bradford (2004), Brock and Evans (1986), Brown, Hamilton, and Medoff (1990), Crain (2005), Dey and Sullivan (2012), Hurst and Pugsley (2011), IMF (2007), and Pierce (1998).

[2]Yet another problem is that ex post fines and tort liability may be less effective against small firms because they tend to be judgment proof and are also less susceptible to reputational sanctions (hence the notion of "fly by nights"), which in turn gives such firms a socially inefficient competitive advantage. Ringleb and Wiggins (1990) find a large increase in small companies in hazardous sectors subject to liability. Ironically, these features are a standard justification for regulation (Shavell 1993) even though, in practice, firms for which this consideration is most forceful are often exempt.

[3]Capital market imperfections or other factors that may justify certain forms of preferential treatment for small business are set to the side in this investigation.

[4]For example, Becker and Henderson (2000) present evidence suggesting that shifts of production to new small-scale plants (subject to a de facto policy of rare inspections) contributed to air quality degradation. Gao, Wu, and Zimmerman (2009) present evidence suggesting that firms remained small to avoid more stringent securities law reporting requirements, and Holder, Karim, and Robin (2013) indicate that such firms' reporting quality suffered relative to that of all nonexempt firms and to firms only moderately larger than the exemption threshold.

[5]Regarding the extensive margin, Snyder, Miller, and Stavins (2003) find that the main channel by which the regulation of chlorine manufacturing reduced pollution was by inducing exit by firms using the dirtier technology.

rather the possibility of combining regulation with taxation. In particular, even when ideal corrective taxation is infeasible because harm (or certain proxies, like emissions) is not observable, some attributes like output or employment might be observed.[6] Indeed, exemptions and related preferences are usually predicated on the observability of some such measures (e.g., output, revenue, number of employees). Taxation that is geared to observable quantities can offer a useful supplement to regulation for a number of reasons. First, unregulated firms cause harm and thus produce too much, which also includes inefficient entry. Second, it is often forgotten that regulated firms typically cause harm and consequently produce socially excessive quantities: most regulation does not in fact eliminate all harm, and it is familiar that optimally stringent regulation usually will not do so due to rising marginal costs and diminishing marginal benefits of control. Third, exemptions, even if they are second-best efficient, may create distortions that we wish to correct.

Section 2 introduces a model of regulation that applies to firms with different productivities, which differences are unobservable to the government. Regulation consists of imposing a supplemental production technology that entails both fixed and variable costs of compliance, and this technology reduces but does not eliminate the external harm caused by firms' output.[7] The section characterizes the first best, analyzes firms' behavior under regulation and no regulation, and then compares the two taking into account output effects, including the decision whether to produce at all. Last, the section examines an output tax and compares it to regulation. The analysis throughout this section serves mainly as background for that which follows, but there are some results of interest. For example, technically inefficient regulation — regulation whose compliance costs exceed harm reduction for all firm types at any level of output — can dominate no regulation (because of favorable output effects), but it in turn is dominated by output taxation. Technically efficient regulation, however, can dominate output taxation because the latter is based merely on output and not on how the output was produced and thus how much harm it generates, a matter governed by regulation.[8]

Section 3 introduces an exemption under which firms are subject to regulation if and only if their output exceeds a quantity threshold.[9] A number of cases arise. In one, no firm chooses

---

[6]As Glaeser and Shleifer (2001) emphasize, this may not always be true, but as mentioned in the text that follows, when it is not, it seems difficult to fashion exemptions based on such quantities.

[7]It is immaterial whether the harm is caused by the production process or output itself. For some purposes, however, it is important whether the harm is a conventional externality (what is modeled here), a so-called internality, or one that is deemed in need of regulation (rather than relying solely on optimal contracting between firms and either employees or customers) due to various information problems. In the latter cases, as a very rough first cut, one might view the degree of individuals' discounting or other underestimation of harm as corresponding to the magnitude of an externality. See Gruber and Köszegi (2001).

[8]In any event, as a practical matter governments often employ regulation even when tax schemes may be superior, so it is important to determine how such regulation is best designed.

[9]Brock and Evans (1985, 1986) differs in many respects, including that their "regulations" are output taxes and that the government is assumed to be able to observe firms' types, which greatly changes the analysis (and, if they had not restricted available instruments, the regulator could in essence have dictated efficient behavior). Keen and Mintz (2004) and Dharmapala, Slemrod, and Wilson (2011) model tax exemptions, where the relevant questions (focused on the deadweight loss of taxation) are largely different. Dharmapala, Slemrod, and Wilson (2011) employs a different model of entry and allows for general equilibrium effects on industry price, which would introduce additional dimensions but are set aside to avoid undue complexity.

an output above the exempt level yet regulation binds in the sense that the most efficient firms would otherwise have produced higher outputs but now produce less in order to be exempt. Such schemes can dominate no regulation because unregulated firms' output is socially excessive due to the external harm they cause. In another, more interesting case, some firms (the most efficient) have outputs above the exemption threshold, thereby subjecting themselves to regulation; other firms (of intermediate efficiency) have outputs clustered at the exemption threshold; and still other firms (of lower efficiency) produce output below the threshold. Determination of the optimal exemption threshold is complex. A central reason is that raising the exemption, which causes some regulated firms to jump down (discretely reduce their output) to the exempt level, has ambiguous effects on social welfare. Although their output is now unregulated and thus more harmful, their quantity of output is also lower than was their regulated quantity level, and we must keep in mind that regulated output is also harmful (although less so per unit of output). The optimal exemption can be zero (tantamount to simple regulation without an exemption), and this may be so regardless of how high are the costs of regulatory compliance (again, due to output effects). Relatedly, higher fixed or marginal costs of regulatory compliance do not necessarily favor a higher exemption level. In all, cost-benefit analysis of a simple regulation with an exemption — which is the sort of regulation often employed — is much less straightforward than usually imagined, on account of output effects and the fact that the outputs of both regulated and unregulated firms cause harm, although to different degrees.

Section 4 introduces output taxation alongside of regulation with an exemption.[10] It first examines taxation of exempt output (only), motivated by the fact that this output is more harmful and by the concern that firms jumping down to bunch at the exempt level of output may reduce productive efficiency and increase harm. It turns out that there are many cases to consider, some raising a number of subtleties. When the optimal regime involves some firms producing above-exempt quantities, some bunching at the exemption, and some producing less, the optimal tax on exempt output is strictly below the harm caused by such output (because the incentive of regulated firms to jump down is otherwise *too small*), and it is even possible that this tax is optimally set equal to zero. Second, a tax on (only) regulated output is considered, motivated in part by administrative considerations (regulated firms are already subject to inspection) and the fact that, as mentioned, regulated output often causes harm, even though less than that caused by unregulated output. Once again, when the optimal regime involves masses of firms in all three output categories (above the exemption, at the exemption, and below the exemption), the optimal tax is strictly less than the harm caused by such output (because the incentive of regulated firms to jump down is now *too large*), and it is possible that this tax is optimally set equal to zero.

Finally, section 4 analyzes regulation with an exemption when there are separate taxes on unregulated and regulated output. With these instruments, a planner can achieve the first best in

[10]For prior work on tax-like instruments, regulation, and administrative costs (typically, only two of the three, and without exemptions), see Glaeser and Shleifer (2001, 2003), Polinsky and Shavell (1982, 1992), Shavell (1993), and Shleifer (2012). Christiansen and Smith (2012) consider regulation and taxes when different units of consumption cause different external harm and when some consumption escapes taxation; Eskeland (1994) supplements a common abatement requirement with an output tax; Montero (2005) assumes that technology choice is observable but output is not; and Spulber (1985) compares effluent taxes and permits, which achieve the first best, to output taxes (when harm is caused by an input) and regulation that takes the form of a per-firm effluent ceiling.

a decentralized manner by setting each tax equal to the corresponding external harm and making the decision whether a firm is to be subject to regulation voluntary. In the most interesting case, very efficient firms choose to be subject to regulation and produce high quantities, moderately efficient firms choose not to be subject to regulation and produce low quantities (with no firms producing in a quantity range between the quantities of these two groups), and relatively inefficient firms do not operate. If a mandatory regulation is to be employed, there is no welfare loss as long as there is an exemption and the threshold is set in this quantity gap where no firms produce. If it is set above that range, there will be firms that optimally produce quantities below the exemption level but should be — and wish to be — subject to regulation (that is, to employ the costly technology but be subject to the lower output tax). If it is set below that range, there will be firms that optimally produce quantities above the exemption level but wish to be — and should be — exempt from regulation (that is, not to employ the costly technology but be subject to the higher output tax).

A concluding section discusses some respects in which the present investigation casts a different light on how to think about the regulation of harmful activity viewed more broadly. The central analysis here focuses on optimal regulation supposing that it is of the form often employed: command and control regulation, allowing also for exemptions. Proper cost-benefit analysis differs importantly from customary methods. Moreover, when regulation is supplemented by simple forms of output taxation, new and in some instances qualitatively different results emerge. This latter analysis illuminates the relationship between the social planner's problem when only conventional instruments are available and that when the only constraints are due to the underlying technology and available information, as under a pure mechanism design approach.[11]

2. Preliminary Analysis

2.1. Model

In the absence of regulation, a firm of type γ produces output q at cost $\gamma c(q)$, where $c(\cdot)' > 0$, $c(\cdot)'' > 0$, $c(0) = 0$, and $c'(0) > 0$. Firms' types are distributed according to the positive density function $g(\gamma)$ on the interval $[\gamma^\circ, \infty)$, where $\gamma^\circ > 0$.[12] The government observes firms' outputs but not their types. Consumers buy output at the constant price p.

Each unit of output causes external harm of h^i, with $h^N > h^R > 0$; $\Delta h \equiv h^N - h^R$, where the superscripts N and R denote regimes with no regulation and with regulation, respectively. As

[11]On regulation more generally (the typical application involving monopoly pricing), see Baron (1989), Baron and Myerson (1982), and Laffont and Tirole (1993); see also Baron (1985) and Laffont (1994), extending the analysis to a monopolist that pollutes, and Dasgupta, Hammond, and Maskin (1980), examining in the pollution context the Groves-Clarke-Vickrey mechanism and also instruments with no communication from firms.

[12]This formulation is similar to Lucas's (1978) model where heterogeneity in managerial talent, which is subject to diminishing returns, underlies the size distribution of firms. In many of the cases examined below, it will be assumed that γ° is sufficiently small that the firm of this type earns positive profits; the pertinent conditions in such cases will be obvious. In any event, in regimes in which no firms operate, social welfare will be zero, and comparisons with other regimes would be simplified accordingly. Furthermore, it will be supposed throughout that c' rises at a sufficient rate that $q^N(\gamma^\circ)$ (determined by expression 2) is finite.

mentioned in the introduction, the assumption that $h^R > 0$ is often realistic and also tends to hold when regulatory stringency is chosen optimally. It will be obvious which results below would differ, and how, for the case in which $h^R = 0$.

Regulation also involves compliance costs. Specifically, regulated firms (of every type) incur a positive fixed cost K and a positive marginal cost of k per unit of output. The government is assumed to bear no administrative costs. An equivalent interpretation is that the government does in fact bear such costs (fixed and/or variable), but it charges each firm a fee equal to these costs, and this fee is included in firms' regulatory compliance costs. (In section 3, the model will be extended to allow for an exemption from regulation for any firm with output $q \leq q^E$. In addition, various forms of output taxes will be introduced later in this section and in section 4.[13])

Finally, as a benchmark, let us state the first best. First, taking as given whether a firm of type γ is subject to regulation, its (conditionally) optimal output is that which equates its marginal cost to $p - h^i$, although if its marginal cost exceeds this level at $q = 0$, it optimally does not produce in regime i. Second, a firm of type γ should be regulated if and only if its contribution to welfare (profits minus the external harm it causes) is higher in that regime.[14] Four types of optima can arise: no firms produce (suppose $h^R > p$); all firms that produce are unregulated (suppose $k \geq \Delta h$ and h^N is small); all firms that produce are regulated (suppose k and K are near zero, $h^N > p$, and h^R is small); and only some firms that produce are regulated (in which case those regulated will be all types γ above some γ^*, on account of scale economies).[15]

2.2. No Regulation

In the regime with no regulation, a firm of type γ chooses $q^N(\gamma)$ to maximize profits:

$$(1)\ \pi^N(q^N(\gamma),\gamma) = pq^N(\gamma) - \gamma c(q^N(\gamma)).$$

The first-order condition for an interior solution, if one exists, is simply

$$(2)\ \gamma c'(q^N(\gamma)) = p,$$

which, if we differentiate with respect to γ and rearrange terms, indicates that

[13]Taxes directly on external harm are taken to be infeasible, which may be motivated by the unobservability of such harm. The case with multiple inputs, some more closely related to external harm than output is, would be intermediate, with differential input taxes better controlling external harm but creating input distortions (unless the input is a perfect proxy for external harm). See Plott (1966) and Spulber (1985).

[14]In a regime in which a firm does not produce, its contribution to welfare is zero, so we need not separately require that welfare be nonnegative.

[15]Compare Proposition 6(e) and 6(f).

$$(3)\ \frac{dq^N(\gamma)}{d\gamma} = -\frac{c'(q^N(\gamma))}{\gamma c''(q^N(\gamma))} < 0.$$

As we would expect, the unregulated profit-maximizing quantity is falling in γ. Moreover, it is apparent that there will exist some γ, denoted γ^N, such that $q^N(\gamma^N) = 0$. (From expression (2), $\gamma^N = p/c'(0)$.) Firms with $\gamma \geq \gamma^N$ will not produce in the regime with no regulation.

Social welfare in this regime is given by

$$(4)\ W^N = \int_{\gamma^\circ}^{\gamma^N} \left[pq^N(\gamma) - \gamma c(q^N(\gamma)) - h^N q^N(\gamma) \right] g(\gamma) d\gamma.$$

That is, firms that produce, those with $\gamma \in [\gamma^\circ, \gamma^N)$, generate benefits from their output (here, just the price times quantity), production costs (the net of these first two terms constituting firms' profits), and external harm.

2.3. Regulation

In the regime with regulation, a firm of type γ chooses $q^R(\gamma)$ to maximize profits:

$$(5)\ \pi^R(q^R(\gamma), \gamma) = pq^R(\gamma) - \gamma c(q^R(\gamma)) - K - kq^R(\gamma).$$

The first-order condition is

$$(6)\ \gamma c'(q^R(\gamma)) = p - k.$$

A number of observations should be made. First, as will be elaborated below, the marginal cost of regulatory compliance, *k*, has an effect on quantity akin to that of a linear tax on output.[16] Second, this first-order condition indicates a firm's optimal choice of $q^R(\gamma)$ taking as given that it chooses to operate. Because we now have the fixed cost K, expression (6) is a necessary but not sufficient condition for $q^R(\gamma) > 0$ to maximize profits. To explore this further, we can again differentiate the first-order condition with respect to γ and rearrange terms to learn (essentially as before) that

$$(7)\ \frac{dq^R(\gamma)}{d\gamma} = -\frac{c'(q^R(\gamma))}{\gamma c''(q^R(\gamma))} < 0.$$

[16] Although *k* has an output effect like that of an output tax, it is hardly the same because *k* involves a real resource cost rather than a transfer and hence has different implications when examining social welfare (and not just firms' behavior).

The regulated quantity — again, conditional on a firm's producing positive output — is falling in γ (and, for a given quantity, at the same rate as without regulation). Furthermore, for $\gamma = (p-k)/c'(0)$, we know that the optimal quantity, given operation, is zero. For γ that is only slightly lower (for a type of firm only infinitesimally more efficient), the optimal quantity barely exceeds zero, so revenue minus variable costs (production costs and marginal regulatory compliance costs) will be barely positive and therefore insufficient to exceed the fixed compliance cost K. Hence, the γ^R below which firms earn positive profits with regulation is strictly less than $(p-k)/c'(0)$ (which in turn is strictly below γ^N). Assume that γ° is sufficiently low and that K is not too large such that $\gamma^\circ < \gamma^R$, i.e., that some firms produce in the presence of regulation (see note 12).

Social welfare in the regime with regulation is given by

$$(8)\quad W^R = \int_{\gamma^\circ}^{\gamma^R} \left[pq^R(\gamma) - \gamma c(q^R(\gamma)) - K - kq^R(\gamma) - h^R q^R(\gamma) \right] g(\gamma) d\gamma .$$

This expression differs from expression (4) in a number of respects: Firms that produce are now those with $\gamma \in [\gamma^\circ, \gamma^R)$, which is a narrower interval (on account of both the marginal cost effect and the fixed cost effect). For firms that do produce, costs are higher due to the additional costs (fixed and marginal) imposed by regulatory compliance. Finally, for output that is produced, harm per unit, now h^R, is lower.

2.4. Regulation versus No Regulation

Regulation is optimal if and only if the value of W^R given by expression (8) exceeds the value of W^N given by expression (4). Subtracting expression (4) from expression (8), this condition is

$$(9)\quad -\int_{\gamma^R}^{\gamma^N} \left[pq^N(\gamma) - \gamma c(q^N(\gamma)) - h^N q^N(\gamma) \right] g(\gamma) d\gamma$$

$$+\int_{\gamma^\circ}^{\gamma^R} \left[p\left(q^R(\gamma) - q^N(\gamma)\right) - \gamma\left(c(q^R(\gamma)) - c(q^N(\gamma))\right) \right.$$

$$\left. -\left(K + kq^R(\gamma)\right) + \left(h^N q^N(\gamma) - h^R q^R(\gamma)\right) \right] g(\gamma) d\gamma \;>\; 0.$$

The first integral indicates the social welfare loss (which can be negative, i.e., a gain) from firms that no longer operate on account of regulation. The first two terms in the integrand, revenue minus production costs, are what their profits would have been (see expression 1) and

hence are positive, indicating a welfare loss to that extent. However, these firms also no longer generate the external harm, $h^N q^N$. Clearly, if h^N is sufficiently large, the first integral is negative, so the aggregate effect on social welfare on account of induced exit is positive. (Indeed, h^N could be large enough that even the most efficient firms in this range, those of type γ^R, generate more harm than good.)

The second integral indicates the effects of regulation on firms that operate under both regimes. The first line of the integrand shows the revenue difference and the production cost difference on account of output reduction (recall that $q^R(\gamma) < q^N(\gamma)$ due to the marginal regulatory compliance cost k). For each type of firm in this range, revenue and costs fall, but we know that the net effect on social welfare must be negative because the additional production by the unregulated firms is profitable in that regime. The next line of the integrand shows, to begin, the additional welfare loss on account of the cost of regulatory compliance. Not surprisingly, looking just at the effects of regulation on output and costs, regulation reduces welfare. Finally, we have the difference in external harm. Regulation reduces this cost, for each firm type, for two reasons: $h^R < h^N$, i.e., harm per unit of output is lower; and $q^R(\gamma) < q^N(\gamma)$, i.e., output is also lower.

Expression (9) is just the cost-benefit assessment of a very simple regulation in a basic setting, yet it is more complex than is ordinarily appreciated due to output effects (both exit and quantity reduction) and also the fact that even regulated firms generate some external harm. A few generalizations are possible. Most obviously, the desirability of regulation rises unambiguously with the magnitude of h^N and falls with h^R.

Interestingly, however, a requirement that $\Delta h > k$ (that regulation reduces harm per unit of output by more than the marginal cost of regulatory compliance, even ignoring the positive fixed cost K) is not a necessary condition for regulation to raise social welfare. Indeed, it is not even necessary that $\Delta h > 0$. That is, a regulation that imposes both marginal and fixed costs and also fails in reducing harm one iota can raise social welfare. Consider, for example, a case in which $\Delta h = 0$ but h^N is sufficiently large that even the most efficient firm type, γ°, causes more harm than good. Moreover, suppose that K is sufficiently high that $\gamma^R = \gamma^\circ$. Regulation is desirable because it shuts down the industry, even though, conditional on operation, regulation imposes more cost on any firm than would be gained by the reduction in harm caused by the output that would be produced by that firm. Of course, products or production methods are sometimes banned. More broadly, because the marginal cost of regulatory compliance, k, acts in some respects as an output tax, causing $q^R(\gamma)$ to be below $q^N(\gamma)$, regulation contributes something to social welfare precisely because of its (marginal) costs, because we are assuming that $h^R > 0$, a benefit that supplements any reduction in harm for a given level of output (having a magnitude of Δh, which we ordinarily suppose to be positive when regulation is efficient). This point is in addition to induced exit, i.e., the fact that both marginal and fixed costs (k and K) lead less productive firms (which produce less social surplus per unit of output and thus of external harm) to exit ($\gamma^R < \gamma^N$).

Accordingly, we can state:

Proposition 1, comparing Regulation and No Regulation:

a. Regulation (versus no regulation) raises social welfare if and only if inequality (9) holds.

b. A higher h^N and a lower h^R favor regulation.

c. Regulation can raise social welfare even if $\Delta h \leq 0$ — and thus, a fortiori, even if $\Delta h \leq k$, implying that, conditional on a given level of output, regulation is strictly inefficient.

2.5. Output Taxation

Suppose now that the government has available an additional, alternative instrument, a linear output tax t. At present, we will only compare an output tax to a regime with no regulation and to a regime with regulation; regimes that mix taxes and regulation (a central focus of this article) are analyzed in section 4. Because output is taken to be observable, such an instrument is natural to consider. To facilitate a brief analysis, assume that there are no government administrative or firm compliance costs associated with the tax and that the shadow value of government revenue is one (so that the transfer of revenue in itself is socially neutral).

Because the analysis of such a regime is simple and familiar, a sketch will suffice. Comparisons for now are to the regime with no regulation. Firms' profits differ from expression (1) because we now must also subtract $tq^T(\gamma)$ (each appearance of q now bears the superscript T). The firm's first-order condition, corresponding to expression (2), has $p - t$ on the right side instead of just p. (Comparing this condition to (6) reinforces the prior statement that the marginal cost of regulatory compliance, k, acts like an output tax with regard to regulated firms' output choices.) Firms' optimal choices of $q^T(\gamma)$ vary with γ as before (expression 3). Finally, the expression for social welfare (4) is unchanged except that the upper limit of integration is γ^T (the type whose first-order condition implies a quantity of zero) and the quantities are $q^T(\gamma)$:

$$(10)\ W^T = \int_{\gamma^\circ}^{\gamma^T} \left[pq^T(\gamma) - \gamma c(q^T(\gamma)) - h^N q^T(\gamma) \right] g(\gamma) d\gamma .$$

Note that the external harm per unit of output remains h^N, as in the case with no regulation, a point that will be significant when we compare this output tax to regulation.

Maximizing W^T obviously involves setting $t = h^N$, so that firms of each type (that choose to operate) equate marginal cost to price minus marginal harm, which is equivalent to their equating the social marginal cost (the marginal production cost plus the externality cost) to price, which indicates consumers' marginal benefit. Firms induced to exit are unable to produce any quantity that has a social marginal cost less than price. Finally, because this problem nests that with no regulation (which corresponds to $t = 0$), optimal taxation dominates a regime without any regulation or taxation.

Next, compare the optimal output tax regime to the regime with regulation (and, as mentioned, no taxation). The analysis, which is more involved than one may have expected, appears in the appendix. The primary complication is that either γ^R or γ^T could be larger: that is,

the efficiency of the firm just indifferent to operation could be higher under either regime. The reason is that regulation, on one hand, imposes the marginal cost k and the fixed cost K, both of which reduce γ^R below γ^N, whereas the output tax imposes what is effectively a marginal cost of t. Now, if $t \leq k$, the output tax is less costly to firms, so we will have $\gamma^T > \gamma^R$. (This inequality is strict, even when $t = k$, due to the fixed cost K.) But if t exceeds k by a sufficient amount, then $\gamma^T < \gamma^R$. The expressions for the social welfare comparison differ qualitatively in these two cases and thus must be stated separately (even though in both instances we are subtracting expression 10 from expression 8). The main results can be summarized as follows:

Proposition 2, comparing Regulation and Output Taxation:

- *a. Regulation (versus output taxation) raises social welfare if and only if the applicable inequality (A1 or A2) holds.*
- *b. A higher h^N and a lower h^R favor regulation.*
- *c. Higher fixed or marginal costs of regulatory compliance do not necessarily disfavor regulation.*
- *d. Regulation is necessarily dominated by output taxation when $\Delta h \leq k$.*
- *e. In addition, output taxation dominates no regulation, and the optimal output tax is such that $t = h^N$.*

Regarding Proposition 2(a), keep in mind that, although output taxation induces optimal output for each type of firm given the technology it employs, it does not reduce harm per unit of output, which regulation does. If regulation has high costs and harm is negligible even without regulation, output taxation produces greater social welfare, but if regulation has negligible costs and Δh is large, regulation produces greater welfare. Proposition 2(b) is straightforward. As before, a higher h^N favors regulation because any level of output under taxation (whether attributable to firms that do not operate under regulation or to firms that do) causes more harm.[17] And a lower h^R favors regulation because harm is lower with regard to output produced by regulated firms that operate. The explanation for Proposition 2(c) is now familiar: under regulation, higher compliance costs have output effects that may raise social welfare by more than the costs themselves;[18] therefore, because social welfare under regulation can rise whereas that under output taxation is unaffected, higher costs can favor regulation. Proposition 2(d), however, stands in contrast to Proposition 1(c): Technically inefficient regulation cannot dominate output taxation because in that case the only benefit of regulation is its negative output effect, which output taxation produces without incurring compliance costs.[19]

---

[17]Raising h^N raises the optimal output tax t, but the effect on social welfare in the output tax regime is not, at the margin, affected (by the envelope theorem, assuming that t was initially set optimally).

[18]For example, raising k from an initial value of zero might raise social welfare under regulation on account of the output effect, and this gain would be large if h^R was large. It might appear that raising K unambiguously reduces social welfare under the regulation regime, favoring output taxation. However, in addition to K appearing in the second integrand of (A1) and (A2), it also influences γ^R. Specifically, increasing K induces exit. Under regulation, the firm just indifferent to operating earns zero profits and thus its exit raises welfare by $h^R q^R(\gamma^R)$. When h^R and $g(\gamma^R)$ are sufficiently large, this effect will exceed the welfare loss due to inframarginal regulated firms bearing a higher fixed cost.

[19]To confirm this conclusion, note first that, when $\Delta h \leq k$, any unit of output produced by a regulated firm of any type would contribute more to social welfare if the firm were instead subject to an output tax(because the greater harm is less than the cost savings. (As noted above, this is strictly so even if $\Delta h = k$ because the positive fixed cost K is also avoided.) Second, any unit of output that contributes positively to social welfare in the output tax regime will be produced.

3. Regulation with an Exemption

3.1. Model and Firms' Behavior

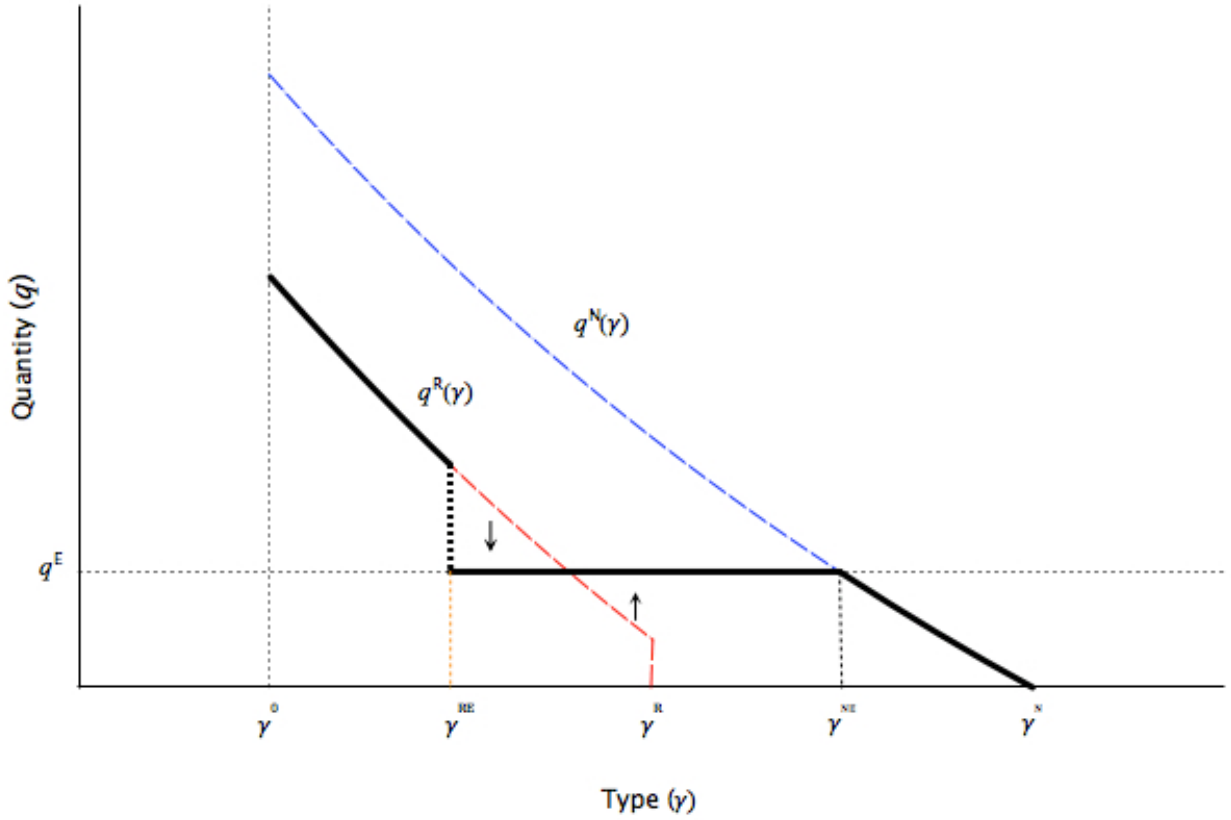
In an exemption regime, firms are subject to regulation if and only if $q > q^E$. This regime nests simple regulation, when $q^E = 0$, and no regulation, when $q^E \geq q^N(\gamma^\circ)$.

To understand the effects of such an exemption regime on firms' behavior, begin with $q^E = q^N(\gamma^\circ)$ — i.e., where the most efficient firm, with the highest output, is exempt when its output is at its unregulated profit-maximizing level. Next, contemplate gradually reducing q^E to zero. At first, when we reduce q^E to just below $q^N(\gamma^\circ)$, the effect is that firms in a neighborhood of γ° will reduce their output to q^E: specifically, all firms of types γ such that $q^N(\gamma) > q^E$. To maintain their former output, $q^N(\gamma)$, or indeed any output above q^E, means that they are subject to regulation and thus incur the fixed compliance cost K and also the marginal compliance cost k on each unit of such output, a strictly positive total. Because $q^N(\gamma)$ was their profit-maximizing quantity without regulation, a slight reduction in q from that level (remaining free from regulation) reduces profits negligibly. Therefore, they indeed choose q^E. Firms with higher γ's, particularly, all those with $q^N(\gamma) \leq q^E$, have no reason to change their output. Define γ^{NE} such that $q^N(\gamma^{NE}) = q^E$, indicating the type of firm whose profit-maximizing output under no regulation just equals the exemption level. All firms with $\gamma \in [\gamma^\circ, \gamma^{NE}]$ produce q^E, and those with $\gamma \in (\gamma^{NE}, \gamma^N)$ produce $q^N(\gamma)$, which is below q^E. Regulation with an exemption that barely binds causes output suppression by the most efficient firms, who cluster at the exemption threshold, and has no effect on the less efficient firms, who produce below the threshold. In this situation, no firm is actually subject to the regulation, but, as explained, regulation still affects behavior.

As q^E is reduced further, there will come a point at which the most efficient firms, those with sufficiently low γ's, no longer wish to produce q^E and instead choose $q^R(\gamma) > q^E$, subjecting themselves to regulation.[20] This happens as q^E falls below the level at which $\pi^R(q^R(\gamma^\circ), \gamma^\circ) = \pi^N(q^E, \gamma^\circ)$, that is, where the most efficient type is just indifferent between choosing the higher output, $q^R(\gamma^\circ)$, which maximizes profit under regulation, and the lower output, q^E, which generates the highest possible profit while remaining exempt from regulation. When q^E is below this level, there will be a range of firms, $\gamma \in [\gamma^\circ, \gamma^{RE})$, that produce $q^R(\gamma) > q^E$, where γ^{RE} is defined such that $\pi^R(q^R(\gamma^{RE}), \gamma^{RE}) = \pi^N(q^E, \gamma^{RE})$. Firms with $\gamma \in [\gamma^{RE}, \gamma^{NE}]$ produce q^E, and firms with $\gamma \in (\gamma^{NE}, \gamma^N)$ produce $q^N(\gamma)$, which is below q^E, as before. See Figure 1.

[20] In principle, this point may never come, which would be true if regulatory costs were sufficiently high that even the most efficient firm, type γ°, cannot profitably produce any output under regulation; as note 12 states, this possibility is assumed not to prevail.

Figure 1: Regulation with Exemption (Three Regions)



Firm types (γ) are depicted on the horizontal axis, with the most efficient type (γ°) toward the left and the least efficient of relevance (γ^N) toward the right. Quantity is on the vertical axis. The outer (northeast) dashed curve depicts quantity choices (q^N) of firms under no regulation, and the inner (southwest) dashed curve shows quantity choices (q^R) under regulation. Note in the latter case that, at γ^R, quantity drops discontinuously to zero due to the fixed cost K.

The bold curve shows quantity choices under regulation with an exemption at q^E, as described just above: The most efficient firms, those with $\gamma \in [\gamma^\circ, \gamma^{RE})$, choose the quantities they would have under pure regulation; they are unaffected by the exemption. Firms of intermediate efficiency, $\gamma \in [\gamma^{RE}, \gamma^{NE}]$, cluster at q^E. Less efficient firms, $\gamma \in (\gamma^{NE}, \gamma^N)$, choose the (unconstrained) quantities that they would have under no regulation. The discontinuity at γ^{RE} reflects the fixed cost (K) and the marginal costs of regulation (k, applied to $q^R(\gamma^{RE})$). In addition, as can be seen, there is an interval of firms to the left of the intersection of the q^R curve and q^E — those with γ toward the left of $[\gamma^{RE}, \gamma^{NE}]$ — that are induced by the exemption to drop their output discretely down to the exempt level, q^E (i.e., $q^R(\gamma) > q^E$). In contrast, all firms to the right of that intersection that are in operation produce more due to the introduction of the exemption: As the figure is drawn, for those with γ toward the middle and right of $[\gamma^{RE}, \gamma^{NE}]$, respectively, some produce more (q^E) than their positive quantities $q^R(\gamma)$ under pure regulation, and some produce q^E

whereas they would not have operated under pure regulation.[21] Finally, firms with $\gamma \in (\gamma^{NE}, \gamma^{N})$ produce the unconstrained $q^{N}(\gamma)$ instead of nothing.

As we continue to reduce q^{E}, this depiction continues to hold — and the magnitudes of γ^{RE} and γ^{NE} rise (move to the right) — until $q^{E} = 0$. At that point, $\gamma^{NE} = \gamma^{N}$. In other words, since it is impossible to produce positive output and remain exempt, the rightmost region vanishes. Similarly, $\gamma^{RE} = \gamma^{R}$. (When $q^{E} = 0$, all the firms clustered at the exempt level of output are firms that do not produce.) As stated above, this case corresponds to pure regulation.

3.2. Optimal Exemption Level

First, it is straightforward to demonstrate that the optimal exemption level q^{E} is strictly below $q^{N}(\gamma^{\circ})$, which is to say that regulation with an optimal exemption produces greater social welfare than a regime with no regulation. To see this point, recall from subsection 3.1 that the only effect of reducing q^{E} slightly from a starting point of $q^{N}(\gamma^{\circ})$ is to induce firms with γ in the (positive) neighborhood of γ° to reduce their output slightly, from $q^{N}(\gamma)$, which is barely above this q^{E}, to q^{E}. (These firms remain exempt from regulation.) Because $q^{N}(\gamma)$ maximized profits, the slight reduction in output has no first-order effect on profits and thus on social welfare, except through the change in external harm. Moreover, because $h^{N} > 0$, this output reduction generates a first-order welfare gain. This analysis provides an initial insight into exemptions (and reinforces a lesson from the earlier analysis of pure regulation): output effects matter and, in this instance with regard to firms not subject to regulation, some reduction in output necessarily increases social welfare due to the presence of (uninternalized) externalities. Here, this result holds even though the imposition of this regulatory scheme does not actually subject any firm to regulation.

Next, suppose that we continue to reduce q^{E} but that we stay in the scenario in which there are only two regions: more efficient firms clustering at q^{E}, and less efficient firms producing lower levels of output, as they would if there were no regulation. The marginal loss of profits (due to the quantity reduction) by the ever increasing group of firms in the first region (γ^{NE} is rising, so $[\gamma^{\circ}, \gamma^{NE})$ is widening) is now first-order and eventually may (but need not) exceed the welfare gain from the marginal reduction of external harm. Accordingly, there might exist an interior optimum in this scenario. When this is also a global optimum, which is possible,[22] we would have a situation in which the optimal regulatory scheme (relative to a regime of no regulation) raises social welfare entirely due to the output reduction by firms that reduce output and thereby cluster at the exemption threshold. Note that, in this instance, firms' jumping down to avoid regulation is the source of the regulation's benefit, not an inefficient side-effect of exemption.

---

[21] To confirm that it is possible that some firms in the middle group would be ones that would not operate in the absence of an exemption, suppose that q^{E} is near zero and consider firms with γ just below γ^{NE}.

[22] Note that, by considering a K that is sufficiently large, we will continue to have only two regions as we keep reducing q^{E} until we reach a level as low as we like, including $q^{E} = 0$. Moreover, by considering an appropriate level of h^{N}, the optimum could be at any q^{E} that binds, including $q^{E} = 0$: if h^{N} is sufficiently small, reducing q^{E} will soon switch to lowering social welfare, whereas if h^{N} is sufficiently large, reducing q^{E} will continue to raise welfare.

For most of the remainder of the analysis, attention will be confined to the more interesting scenario in which there are three (nonempty) regions, as depicted in Figure 1. Social welfare for this case (denoted by the superscript *R/E*) is given by

$$(11)\ W^{R/E}(q^E) = \int_{\gamma^\circ}^{\gamma^{RE}} \left[pq^R(\gamma) - \gamma c(q^R(\gamma)) - K - kq^R(\gamma) - h^R q^R(\gamma) \right] g(\gamma) d\gamma$$

$$+ \int_{\gamma^{RE}}^{\gamma^{NE}} \left[pq^E - \gamma c(q^E) - h^N q^E \right] g(\gamma) d\gamma$$

$$+ \int_{\gamma^{NE}}^{\gamma^{N}} \left[pq^N(\gamma) - \gamma c(q^N(\gamma)) - h^N q^N(\gamma) \right] g(\gamma) d\gamma .$$

The first integral in expression (11) indicates the net contribution to social welfare from the most efficient firms; as explained, they choose quantities $q^R(\gamma) > q^E$ and thus are subject to regulation. The integrand in this term is identical to that in expression (8) for social welfare when all operating firms are subject to regulation. The second integral is the contribution from firms of intermediate efficiency that, as explained, choose to produce at the exemption threshold, q^E. Because they are accordingly exempt, the integrand is the same as that in expression (4) for the case with no regulation, except that the quantities here are q^E rather than $q^N(\gamma)$ because these are firms for which $q^N(\gamma) > q^E$, which quantity choices would subject them to regulation. The third term is for firms that are less efficient, but not so much so that (unregulated) operation is unprofitable. Because $q^N(\gamma) < q^E$ for them in any event, their situation is precisely as in the unregulated world, and the integrand in this term is identical to that in expression (4).

Supposing that we remain in a range that is consistent with this scenario — that is, $q^E > 0$, but q^E is not so high as to eliminate the region in which at least some firms, the most efficient, choose $q^R(\gamma) > q^E$ — a necessary condition for an optimal q^E is that $dW^{R/E}(q^E)/dq^E = 0$.[23] Examining expression (11) for $W^{R/E}(q^E)$, we can see that there are two types of effects from a marginal increase in q^E. Most obviously, the value of the integrand in the second term changes, reflecting that firms clustered at q^E will now raise their output accordingly.

Furthermore, the limits of integration (boundaries between the regions), γ^{RE} and γ^{NE}, each fall. Regarding γ^{RE}, because the exempt quantity is now higher, some firms that had barely preferred to be subject to regulation (the least efficient in that range) will now drop their output to q^E and become exempt. In other words, some of the mass in the first integral will now appear

[23]Most terms in the second-order condition associated with expression (12), below, are of indeterminate sign, so this is not a sufficient condition.

in the second. Note, however, that even though these two integrands are entirely different, the effect on social welfare from this shift is rather simple. Firms of type γ^{RE}, who are the ones that jump down, are those that were just indifferent between producing $q^R(\gamma)$ under a regime of regulation and producing q^E under a regime of no regulation. Accordingly, the sum of all the terms except the last (external harm) in the first integrand equals the sum of the first two terms (all but external harm) in the second integrand. Therefore, the change in social welfare from this change in γ^{RE} will simply equal the difference between these two external effects, weighted by the mass of firms that jump down.

Regarding γ^{NE}, the higher exempt quantity now means that the marginal firm that was at this boundary (a firm whose unconstrained output in an unregulated world, $q^N(\gamma)$, just equalled q^E in any event), will now be producing the same level of output (rather than raising its output as q^E is increased), but that lack of change will put it in the bottom integral rather than the middle one. The movement in this boundary obviously has no effect on behavior or on social welfare. (Note that the value of the integrand in the third integral, at this boundary, where $q^N(\gamma) = q^E$, is the same as the value of the integrand in the second integral.)

In light of the foregoing (which implies that, when one mechanically takes the stated derivative, a substantial majority of the terms cancel), we can write

$$(12)\quad \frac{dW^{R/E}(q^E)}{dq^E} = \left[h^R q^R(\gamma^{RE}) - h^N q^E\right] g(\gamma^{RE}) \left(-\frac{d\gamma^{RE}}{dq^E}\right)$$

$$+ \int_{\gamma^{RE}}^{\gamma^{NE}} \left[p - \gamma c'(q^E) - h^N\right] g(\gamma) d\gamma,$$

where

$$(13)\quad -\frac{d\gamma^{RE}}{dq^E} = \frac{p - \gamma^{RE} c'(q^E)}{c(q^R(\gamma^{RE})) - c(q^E)}.$$

The right side of expression (13) is positive because, with reference to the numerator, the most efficient type of firm in (at the boundary of) the middle region, type γ^{RE}, has a marginal cost below price (it does not expand output above q^E because then it would no longer be exempt from regulation). In the denominator, since $q^R(\gamma^{RE}) > q^E$, the cost difference is positive. Therefore, as stated above, raising q^E reduces γ^{RE}. The minus sign is placed on the left side of expression (13) rather than on the right, and in the large parentheses at the end of the first line of expression (12), so that the first-order condition is easier to interpret.

The first term in expression (12) indicates, as previewed above, the change in social welfare due to regulated firms jumping down to q^E as that exemption threshold is increased.

There is a welfare gain due to the external harm from regulated output no longer arising, and a welfare loss because we now have external harm from unregulated output. The former involves less harm per unit of output, because $h^R < h^N$, but more harm due to the greater quantity, because $q^R(\gamma^{RE}) > q^E$. Accordingly, the net effect on external harm from firms jumping down could be of either sign: the benefit of regulation (lower harm per unit of output) is forgone, but the quantity of output that gives rise to external harm falls.[24] Finally, this term in brackets is weighted by the mass of firms that jump down, indicated by the product of the density of the marginal firm type and the rate of change in the marginal type.

The second term in expression (12) indicates the greater production by firms that cluster at q^E. For firms at the upper limit of this integral, of type γ^{NE}, q^E is their profit-maximizing choice, so price equals marginal cost. Hence, for them, the first two terms in brackets in the integrand together equal zero. Their marginal output reduces social welfare by h^N, the harm per unit of output associated with an unregulated firm's production. For more efficient firms in this region, price exceeds their marginal cost (but they do not raise output because they wish to remain exempt), so raising q^E raises profits and thus social welfare on this account, but there remains the externality, h^N. It is a priori indeterminate whether, even for the most efficient firm in this range, the output increase from raising q^E raises or lowers social welfare, accounting for both profits and external harm. Moreover, even if it raises welfare, it remains indeterminate whether the term as a whole, integrating over all firm types in this region (the least efficient of which reduce welfare when they raise output), is positive or negative.

All together, both terms in our first-order condition for the optimal q^E (within this scenario) are of ambiguous sign. It is apparent, however, that a high level of h^R favors a higher exemption level: the only effect of regulation entailing a higher level of residual harm per unit of output is to raise the social benefit (or reduce the social cost, as the case may be) of firms jumping down to the exempt level of output as q^E is increased. A high level of h^N favors a lower exemption level on two accounts: it makes the jumping-down phenomenon more detrimental, and it also renders more harmful (or less beneficial, perhaps, for the most efficient exempt firms) the increase in output for firms clustered at the exempt level of output.

Next, consider the effect of regulatory compliance costs on the optimal exemption level. Interestingly, these do not appear directly in expression (12), our first-order condition for the optimal q^E. The reason is that raising the exemption level saves all compliance costs for firms that, as a consequence, jump down to the exempt level of output, but in so reducing their output, these firms also forgo profits, an excess of price over marginal production cost, which also contributes to social welfare. Moreover, as noted previously, for the marginal firm (type γ^{RE}), these effects are precisely equal. Hence, perhaps surprisingly, regulatory costs have no direct (mechanical) impact on the marginal welfare effect of raising q^E.

Regulatory compliance costs are nevertheless relevant to the optimal value of q^E because

[24] At a given q^E, $q^R(\gamma^{RE})$ is endogenous, but we know in any event that $q^R(\gamma^{RE}) > q^E$. In contrast, the h^i are exogenous and have no effect on any endogenous variables. Accordingly, we can imagine cases in which Δh is arbitrarily small, in which event the quantity effect dominates, and cases in which Δh is arbitrarily large (and, moreover, in which h^R is arbitrarily small), in which event the harm effect dominates.

q^E influences γ^{RE}, the efficiency of the marginal type of firm (and k influences $q^R(\gamma^{RE})$ for a given γ^{RE}). Specifically, it is straightforward to demonstrate that, as we would expect, a higher fixed cost K and a higher marginal cost k each reduce γ^{RE}; that is, when regulatory compliance costs are greater, the firm just indifferent to operating at a high output in the regulated regime, versus producing q^E while being exempt from regulation, is one with greater efficiency. Moreover, it can be demonstrated that this more efficient firm is one with a greater quantity under regulation.[25] Therefore, for a given q^E, more costly regulation is associated with a greater value for the bracketed portion of the first term in expression (12), making a higher exemption more desirable on that account: because the quantity drop when the marginal firm jumps down is from a higher initial level, the savings in harm from regulated output is greater. Whether the optimal q^E rises, however, is a more complicated question because γ^{RE} appears elsewhere in both terms of expression (12). As it turns out, each of these additional effects is of indeterminate sign, so it is a priori indeterminate whether higher regulatory costs favor a higher exemption level.[26]

Finally, having previously demonstrated that the optimal q^E is strictly below $q^N(\gamma^\circ)$, consider now whether the optimal q^E is strictly greater than zero. Given the fixed cost K of regulatory compliance, it may seem that this would be so, but the foregoing analysis suggests that the matter is more complicated. To begin, review expression (11) for $W^{R/E}(q^E)$. At $q^E = 0$, there is no production in the middle and right regions: firms clustering at q^E produce nothing, and less efficient firms (which had output below q^E when there were three operative regions) do not produce either. Therefore, if we raise q^E slightly, starting from zero, there are two effects: firms just at (below) γ^{RE} (which, in this instance, equals γ^R since, after all, an exemption regime with an exemption of zero is identical to regulation without any exemption) jump down to the now-positive q^E, and the more efficient firms among those that were not producing will now enter. (These are firms in the interval $\gamma \in [\gamma^R, \gamma^N)$. The left endpoint was just explained; for the right endpoint, recall that all firms with $\gamma < \gamma^N$ produce a positive quantity when not subject to regulation, and since the marginal contribution of quantity to profit is at its maximum when $q^N = 0$, marginally increasing quantity from that level as the exemption is increased will indeed be profitable, there being no fixed costs for unregulated firms.) Accordingly, we wish to evaluate expression (12), for $dW^{R/E}(q^E)/dq^E$, at $q^E = 0$ when the limits of integration are as just described.

---

[25]For a higher K, this is obvious, on account of γ^{RE} falling. For a higher k, there is a countervailing effect on $q^R(\gamma^{RE})$ due to the higher marginal cost. However, the indifference condition that defines γ^{RE} requires higher profits when regulated (because the fall in γ^{RE} implies higher unregulated profits at q^E), which necessarily implies lower overall marginal costs under regulation and hence a higher q^R. (Actually, a lower overall marginal cost is required even to achieve the same level of profit under regulation because raising k shifts up marginal cost by a constant amount at all levels of output.)

[26]To elaborate briefly, because the first bracketed term in expression (12) is of indeterminate sign, effects of changing γ^{RE} (which is lowered by raising K or k) due to the latter two components of the first term will be indeterminate. Moreover, the effect of changing γ^{RE} on each of those factors is of indeterminate sign. Finally, changing γ^{RE} changes the lower limit of integration of the second term, but, as explained previously, the sign of the integrand at that value is also indeterminate.

$$(14)\quad \left.\frac{dW^{R/E}(q^E)}{dq^E}\right|_{q^E=0} = \left[h^R q^R(\gamma^R)\right] g(\gamma^R)\left(-\frac{d\gamma^{RE}}{dq^E}\right)$$

$$+ \int_{\gamma^R}^{\gamma^N} \left[p - \gamma c'(0) - h^N\right] g(\gamma)\, d\gamma.$$

In contrast to expression (12), the first term in expression (14) is unambiguously positive. (The component of the bracketed term in which we had subtracted the exempt quantity level times the unregulated per unit harm now equals zero because we are evaluating the expression at $q^E = 0$.) Therefore, at least initially, the effect of a higher exemption of inducing firms to jump down to the exemption threshold unquestionably raises social welfare: the negative externality associated with regulated output is avoided (and nothing substituted in its place) when firms at the margin drop their output to become exempt. (And, as before, all the other effects on social welfare cancel in light of the marginal firm's indifference condition.) At least in a neighborhood of $q^E = 0$, a net positive effect will continue to prevail. Interestingly, introducing some exemption from regulation is desirable in this regard precisely *because* it induces some firms subject to regulation to become exempt (by dropping their output).

The second term, however, continues to be ambiguous: At the upper limit of integration, the first two terms in the integrand, taken together, equal zero, so the integrand as a whole is negative, whereas at the lower limit of integration the value may be positive or negative. Clearly, if h^N is sufficiently large, we know the second term will be negative, and if it is true that h^N is large and h^R is sufficiently small, then a barely positive exemption is undesirable.[27] (And this is so regardless of the magnitude of the fixed regulatory compliance cost K or the marginal cost k, although larger costs imply a lower γ^R, which adds at the lower end of the interval of integration more efficient firms that, on this account, have higher levels of marginal profit.[28]) By contrast, there clearly exist combinations of h^N, h^R, and the density function $g(\cdot)$ such that a positive exemption is optimal. For example, if h^N is sufficiently small that the integrand in the second term is positive for some range of γ, consider $g(\cdot)$ arbitrarily close to zero for any γ such that the integrand is negative. In that event, the second term is positive and the first term is always positive, so a positive exemption is optimal.

The foregoing analysis can be summarized as follows:

Proposition 3, on Optimal Exemption from Regulation:

---

[27]The analysis in the text only shows that, in the neighborhood of zero, the optimal q^E is zero. However, it is also clear from the earlier analysis of expression (12) that if h^N is sufficiently large, welfare will be falling as the exemption level rises for any q^E, until $q^E = q^R(\gamma^\circ)$. And welfare rather obviously falls beyond that point when h^N is sufficiently large.

[28]On the other hand, a lower γ^R implies a larger $q^R(\gamma^R)$ for the reasons given in note 25, which, ceteris paribus, raises the magnitude of the first term. And the other factors in the first term change as well.

a. Regulation that exempts at least some firms (whose unregulated quantities would exceed the exemption level) dominates no regulation.

b. Regulation with an exemption can dominate no regulation and the exemption level can be optimal even if no firms are subject to regulation — that is, if any firm whose unregulated quantity would exceed the exemption level chooses to reduce its quantity to the exemption level.

c. The optimal exemption from regulation can be positive or zero, and it can be zero regardless of how high are the fixed and marginal costs of regulatory compliance.[29]

d. If it is optimal to set the exemption at an intermediate level — that is, a positive q^E such that, for some $\gamma > \gamma^\circ$, $\pi^R(q^R(\gamma), \gamma) > \pi^N(q^E, \gamma)$, meaning that a mass of firms finds it profitable to produce quantities above the exemption level — then a necessary condition for the optimal exemption level is that expression (12) equals zero.

e. A higher h^N and a lower h^R favor a lower level of the exemption.[30]

f. Higher fixed or marginal costs of regulatory compliance do not necessarily favor a higher exemption level.

4. Regulation with an Exemption and Output Taxation

This section first considers how the analysis changes when one allows for taxation of exempt output, then taxation of regulated output, and finally taxation of both types of output at different rates.

4.1. Taxation of Exempt Output

Suppose that, in addition to regulation with an exemption, as analyzed in section 3, it is also possible to impose a tax on output that is exempt from regulation. One motivation is that, because such output is unregulated, it is more harmful, so a tax on it seems particularly appealing. Moreover, the feasibility of an output-based exemption, q^E, does suppose that the output level of exempt firms is observable, suggesting that such a tax may be feasible.

Accordingly, let us modify section 3's model by allowing a (nonnegative) tax at the rate t^N on all output that is not subject to regulation. As in subsection 2.5, it will be assumed throughout this section that there are no government administrative or firm compliance costs associated with the tax and that, regarding social welfare, tax payments per se are pure transfers.

---

[29]Note that this final point can hold regardless of how small is Δh, including where $\Delta h = 0$. In expression (14), observe that, even in this limiting case, h^N can be arbitrarily large and $g(\gamma^R)$ can be arbitrarily small. Regarding the latter clause in result (c), it is true that, when regulatory compliance costs become sufficiently large, it will no longer be true that γ^R is high enough that any firms choose to produce subject to regulation, but it remains true that $q^E = 0$ can be optimal: we are left with only the latter term in expression (14) but, as explained (see note 27), that term will be negative if h^N is sufficiently large.

[30]A lower h^R only weakly favors a lower level of the exemption because, if the optimal q^E is such that there are only two regions (the earlier case where even the most efficient firms produce at q^E), then changing h^R at the margin has no effect on the optimal q^E.

The analysis appears in the appendix. The main results are:

Proposition 4, on Optimal Taxation of Exempt Output (for a given exemption level, q^E):

- *a. It is possible to have an optimum with $\gamma^{RE} < \gamma^\circ$, that is, with no firms producing output in excess of q^E and thus subjecting themselves to regulation. In that case, the optimum has $t^N = h^N$.* [31]
- *b. It is possible to have an optimum with $\gamma^{RE} > \gamma^\circ$, that is, with a mass of efficient firms producing output in excess of q^E. In that case, the optimum has $t^N \in [0, h^N)$, and a necessary condition for the optimal tax (if it is interior) is that expression (A3) equals zero. That is, the optimal tax on exempt output does not fully internalize the externality, and this optimal tax might equal zero.*
- *c. If there is a missing middle region (a necessary condition for which is $t^N > k$), no firms produce q^E, and, for K sufficiently small, some types of firms wish to be subject to regulation despite producing output below q^E (and, if voluntary regulation were not permitted, some would jump up, producing output just above q^E, in order to be subject to regulation).*

For Proposition 4(a), if indeed all output is unregulated, it is hardly surprising that the optimum sets $t^N = h^N$. Proposition 4(b) indicates that this is not true when we also have regulated firms. The intuition is that, as we fully internalize the externality regarding the output of unregulated firms, there is no marginal social gain from output reduction by them, whereas the induced jumping up to a higher, regulated level of output reduces social welfare because the harm caused by regulated firms (even though it is lower per unit of output) is external, whereas the harm caused by unregulated output is completely internalized. (To be sure, harm is harm with regard to social welfare, but there are differences regarding revenue, production costs, regulatory compliance costs, and tax payments that change when firms jump up. The net of all this is just *uninternalized* externalities, which in the present scenario are only $h^N - t^N$ per unit of output by unregulated firms while for regulated firms per-unit harm is still h^R.) This conclusion runs against conventional wisdom that tends to view the jumping down induced by exemptions as detrimental: If indeed that were always so, then the jumping (back) up caused by raising the tax on unregulated output would always be beneficial, but we can see that this need not be so. To complete our discussion of Proposition 4(b), note that, even when $t^N = 0$, it is possible that jumping down can be sufficiently detrimental that no tax on unregulated output is desirable (indeed, if allowed, a marginal subsidy could be optimal); for details, see the appendix. The analysis underlying Proposition 4(c) is also in the appendix. In brief, it concerns a qualitatively different two-region case characterized by a "missing middle" in which no firms cluster at the exemption level: As one raises t^N, γ^{NE} falls because firms' aggregate marginal cost of producing unregulated output rises, and γ^{RE} rises because jumping down to be unregulated becomes less attractive; if these boundary types meet and cross, which is possible, no firms produce q^E.

[31] It is also possible to have an optimum with $\gamma^{RE} = \gamma^\circ$, which likewise implies that there is no mass of firms producing output in excess of q^E. (The most efficient firm is indifferent to producing q^E and a higher, regulated level of output. The convention has been that it produces the former, exempt quantity, but in any event firms of this type have no mass.) In that case, the optimum has $t^N \in [0, h^N]$. That is, in this intermediate scenario, it is possible that t^N takes a value at one of the endpoints or anywhere in between.

Consider next how the introduction of a tax on unregulated output, t^N, changes the optimal exemption from regulation, q^E. First, examine the two-region scenario in which there is an exemption that is low enough to bind on some firms (forcing them to reduce output) but not so low as to induce any firms to produce output above the exempt level and thereby subject themselves to regulation. The results in Proposition 3(a) and 3(b) are that such a regime dominates no regulation and can be optimal overall. Once we introduce a tax on exempt output, however, the latter is no longer true. The advantage revealed in section 3 of a regulation whose only effect is to induce some firms to reduce (unregulated) output was to diminish the uninternalized externality, h^N. Proposition 4(a) now informs us that, when a tax on unregulated output is introduced into such a regime, it is optimal to set $t^N = h^N$. Once that is done, there is no longer any uninternalized externality, which implies that, conditional on not being subject to regulation, firms' quantities are chosen optimally. Hence, it is not desirable to impose regulation with an exemption, q^E, whose only effect is to suppress quantity. A pure output tax — which is what a tax on all unregulated output in a world with a nonbinding exemption amounts to — would be superior.

Corollary 1: When it is possible to impose a tax on unregulated output, it cannot be optimal to employ regulation with a binding exemption under which no firms produce output above the exempt level (subjecting themselves to regulation). Such a regime is dominated by one with $t^N = h^N$ and a nonbinding exemption ($q^E \geq q^N(\gamma^\circ)$), which is tantamount to no regulation combined with a pure output tax.

To further assess how the availability of a tax on unregulated output affects the optimal exemption level, let us reexamine the intermediate, three-region scenario. To begin, it is clear that such an intermediate scheme may well be optimal despite the introduction of this tax instrument. Consider the case in which k, K, and h^R are each close to zero: that is, in which regulation eliminates virtually all harm at negligible cost. Obviously, regulation (with a binding exemption) will be optimal.[32]

How does introduction of a tax on unregulated output — and, in particular, one set optimally (which, recall, involves $t^N < h^N$ in this scenario) — affect q^E, the optimal exemption level? If we restate the pertinent derivative (expression 12) for this case, the only nominal change is that the first bracketed term substitutes $h^N - t^N$ for h^N, as just explained. This modification favors a higher q^E. The intuition is that, without taxation, raising q^E caused firms to jump down to the exempt level of output, and this had a cost, the magnitude of which was given by the exempt output times the level of external harm per unit of such output; this latter component is now reduced to the uninternalized portion of that harm. Nevertheless, the impact of the tax on the optimal exemption is ambiguous because there are many other effects as well: Because a positive tax implies a higher initial level of γ^{RE}, as explained above, $q^R(\gamma^{RE})$ is lower (the marginal firm is a less efficient one), so the social gain when marginal firms jump down to q^E that is attributable to the reduction in the externality on their regulated output is smaller. In addition, the density is evaluated at a different γ^{RE}, the expression for and value of $d\gamma^{RE}/dq^E$

[32]Furthermore, it is possible that social welfare is rising as we increase q^E from zero: recall the discussion of expression (14) and note that the fact that unregulated output is now subject to a positive tax is immaterial.

change, and both limits of integration for the second (ambiguous) term in expression (12) change (they move closer together, as mentioned). Accordingly, it is not possible to offer a simple characterization of how the introduction of a tax on unregulated output affects the optimal exemption level in this scenario.

4.2. Taxation of Regulated Output

Suppose now that, instead of taxation of unregulated output, regulation with an exemption can be supplemented by (only) a tax, t^R, on regulated output, that is, on all output of firms that are subject to regulation. A natural rationale is that, with regard to firms already subject to regulation, it may be particularly inexpensive to impose such a tax. In any event, for present purposes, it will be assumed that there are no government administrative or firm compliance costs associated with introducing this instrument and that tax payments are pure transfers. In other respects as well, the model is unchanged.

The analysis appears in the appendix. The main results are:

Proposition 5, on Optimal Taxation of Regulated Output (for a given exemption level, q^E):

a. It is possible to have an optimum with $\gamma^{RE} \leq \gamma^\circ$, that is, with no firms producing output in excess of q^E and thus subjecting themselves to regulation. In that case, the optimum has t^R sufficiently high to induce this result, but the particular level of t^R is inconsequential.

b. It is possible to have an optimum with $\gamma^{RE} > \gamma^\circ$, that is, with a mass of efficient firms producing output in excess of q^E. In that case, when $q^E > 0$, the optimum has $t^R \in [0, h^R)$, and a necessary condition for the optimal tax (if it is interior) is that expression (A6) equals zero. That is, the optimal tax on regulated output does not fully internalize the externality, and this optimal tax might equal zero.

c. If $q^E = 0$, the optimum has $t^R = h^R$.

The intuition behind the optimal levels for t^R are straightforward with regard to Propositions 5(a) and 5(c). For Proposition 5(b), the reasoning is analogous to that for Proposition 4(b). The difference is that, in the present case, in which we tax only regulated output (rather than only unregulated output), the incentive for firms to jump down is more readily too large rather than too small. Again, this pushes against full internalization, and if this force is sufficiently strong, it may be optimal to set $t^R = 0$.

Next, examine how the introduction of a tax on regulated output, t^R, changes the optimal exemption from regulation, q^E. If, without such a tax, the optimal q^E is such that there are only two regions — that is, no firms produce regulated output — then a positive tax on regulated output is of no consequence. (It is possible that the availability of such a tax, by improving welfare in the three-region scenario, would make the highest achievable welfare in that setting surpass the maximum achievable welfare when there are only two regions, in which event the previous global optimum would no longer be the global optimum. But this global optimum also

may remain so.[33])

Consider next how the optimal q^E is affected in the three-region scenario. The most straightforward consequence of the tax on regulated output is to reduce the element of social gain from a higher q^E that arises due to the fact that firms jumping down no longer produce regulated output. From the analysis just above and that in subsection 4.1 pertaining to unregulated output, we know that this component now reflects only the uninternalized external harm per unit of regulated output, $h^R - t^R$, and not the full h^R. But, just as when we were considering the analogous question in subsection 4.1, introducing t^R will also change γ^{RE}, $g(\gamma^{RE})$, $q^R(\gamma^{RE})$, and $d\gamma^{RE}/dq^E$, so no simple characterization can be offered.

Finally, observe that the case mentioned in Proposition 5(c) is one that may involve a fully optimal scheme. It was already recalled that $q^E = 0$ can be optimal in the absence of taxation on regulated output. When such a tax is positive, the welfare effect of raising the exemption from zero (see expression 14) changes in a number of ways. Most directly, h^R is replaced by $h^R - t^R$: there is a smaller benefit of raising q^E from zero because the welfare gain from marginal firms that jump down, due to their no longer producing harmful regulated output, is diminished. Of course, as noted just above, the other terms change as well, so the effect could be in either direction. Nevertheless, as explained when discussing expression (14), since h^N can be arbitrarily large, the optimal exemption can be zero even when regulatory compliance costs are high. In sum, it remains possible that $q^E = 0$ is optimal even when there is available a tax on regulated output, and, as already stated, in that event this tax should fully internalize the externality. What we have, in essence, is a regime of simple regulation (no exemption) *plus* a pure output tax (here, all output is regulated output).

Corollary 2: When it is possible to impose a tax on regulated output, it can be optimal to employ regulation with no exemption ($q^E = 0$), and in such a case the optimal tax fully internalizes the externality ($t^R = h^R$). This regime is equivalent to simple regulation (no exemption) combined with a pure output tax.

4.3. Taxation of Exempt and Regulated Output, at Different Rates

This subsection combines the foregoing instruments. That is, we have a regulation, an exemption, q^E, and also two tax instruments: a tax on unregulated output, t^N, and a tax on regulated output, t^R.[34] In all other respects, the model is unchanged. This case is rather

[33]Consider the case in which h^R is near zero, in which case a tax on regulated output can do little to raise welfare; regulation is extremely expensive and output is very valuable, making a substantial exemption optimal; but, as shown in Propositions 3(a) and 3(b), we nevertheless want the exemption to have some binding force (but parameters are such that not much force is optimal because h^N is also low).

[34]An omitted case combines regulation, an exemption, and the pure output tax from section 2.5: a common tax t on all output. Not surprisingly, when one solves this case, behavior and welfare effects are essentially a combination of those with only t^N and only t^R. The presence of the tax has an ambiguous effect on the optimal q^E, and the optimal t can only be shown to lie in the interval $[0, h^N)$. A uniform output tax may have appeal on administrative grounds if goods are already taxed for purposes of raising revenue, such as under a VAT, and different rates can be applied at only modest additional administrative cost. Note further that, if this is so and it is also true that there would be little cost to imposing an output tax or subsidy on firms that are regulated in any event, then the combination of these two taxes allows for

straightforward to analyze, especially by comparison to the preceding ones.

Of particular interest, it is possible to implement the first-best allocation, so the analysis will concentrate on how this is done.[35] Suppose that we have a planner who wishes to maximize social welfare and, moreover, is able to observe each firm's type and command its behavior: whether it is subject to regulation (using the more expensive but less harmful production technology) and its level of output.

Starting with the latter, the socially optimal output, conditional on whether a firm is regulated, maximizes revenue net of production costs, regulatory compliance costs (if applicable), and harm. As is familiar, the pertinent first-order condition is precisely the same as what the firm's would be if it were to bear the full social cost of the external harm. This internalization can be accomplished by setting $t^N = h^N$ and $t^R = h^R$. Moreover, because the maximand is strictly concave, the solution is unique, which implies that output decisions can be decentralized with these output tax instruments. As is familiar, although the optimal output for unregulated and regulated firms (continuing to take that choice as given) does depend on firms' types (their productive efficiency), the planner does not need to rely on this knowledge when it sets output taxes that fully internalize the externality. Finally, note that a firm's optimal output may be zero (a corner solution) with or without regulation, but firms also make this choice — whether or not to produce a positive quantity — in a socially optimal manner when they are subject to these taxes.

The remaining question is which firms should be subject to regulation — that is, which types should produce using the more expensive technology that reduces external harm per unit of output. Our planner wants a firm of type γ to be regulated if and only if

$$(15)\quad pq^R(\gamma) - \gamma c(q^R(\gamma)) - K - kq^R(\gamma) - h^R q^R(\gamma)$$

$$> pq^N(\gamma) - \gamma c(q^N(\gamma)) - h^N q^N(\gamma),$$

where the $q^i(\gamma)$ refer to the quantities that firms would choose when subject to the output taxes t^i. Once again, for any γ, the planner's calculus is the same as the firm's: a profit-maximizing firm would wish to be subject to regulation if and only if inequality (15) holds because, once we set $t^N = h^N$ and $t^R = h^R$, its profits under each regime are the same as its net contribution to social welfare. Hence, the decision of which technology to employ can be left to the firms themselves. The principle is the same as with the quantity decision: once there is full internalization of externalities, firms' decisions — here, whether to be regulated rather than how much to produce — are socially optimal. An implication is that, like with the quantity decision, the planner's knowledge of firms' types, γ, is not needed to implement a decentralized scheme.

---

unrestricted differential taxation of the sort examined in this section.

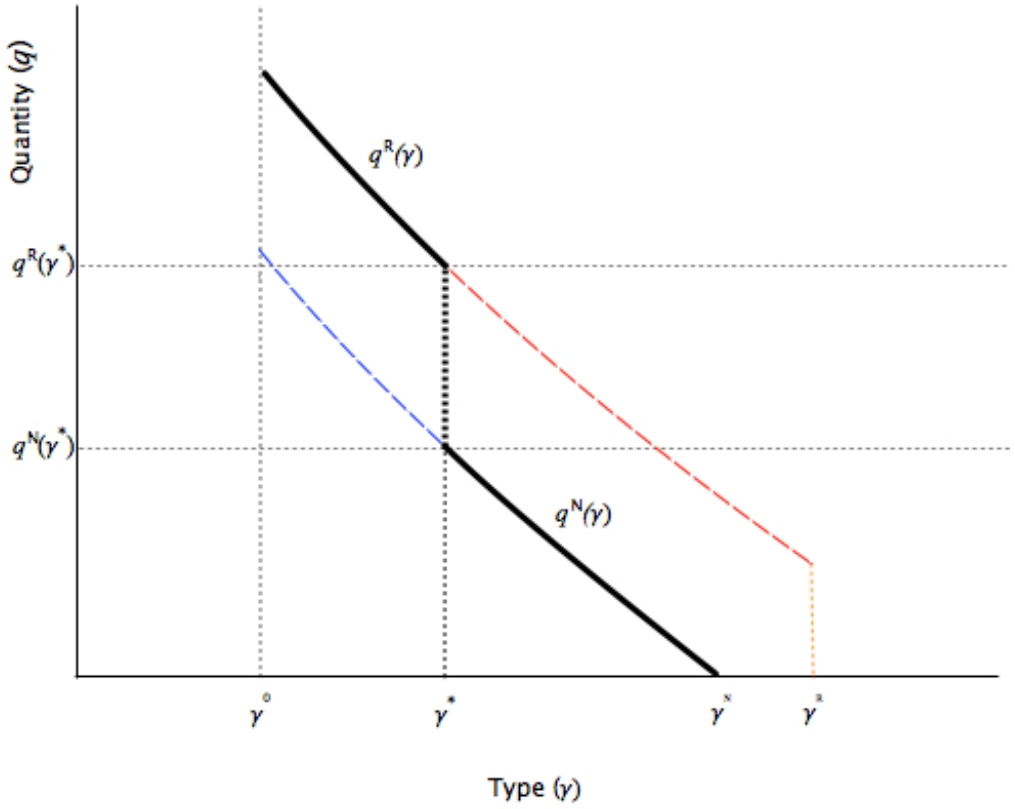
[35] Under the interpretation of the model in which k and K include government administrative costs that arise due to efforts to observe firms and enforce compliance, one would not characterize the result as first best in the traditional sense, but the analysis that follows is obviously unaffected by such matters of interpretation.

Let us now characterize firms' behavior under this optimal regime, with fully internalizing output taxes and what amounts to voluntary regulation. Recall from section 2 that a necessary condition for regulation to be technically efficient is that $k < \Delta h$. (If $k \geq \Delta h$, each unit of output under regulation has an aggregate cost — including the external harm — at least as high as under no regulation, and we also have the positive fixed cost K. Hence, regulation is dominated for any γ, and no firm would choose regulation.) Focusing, then, on technically efficient regulation, and taking as given our (optimal) output taxes, it follows that, for any γ and any quantity choice, a firm's all-inclusive marginal cost under regulation (its marginal production cost plus its marginal compliance cost plus its marginal tax payment) is below its inclusive marginal cost with no regulation. (The condition $k < \Delta h$ means that $k < h^N - h^R$, which here implies that $k < t^N - t^R$; hence, $k + t^R < t^N$.) Therefore, for any γ, we have $q^R(\gamma) > q^N(\gamma)$ — unless $q^R(\gamma) = 0$, as elaborated in the note just below.

Consider next which firms will wish to be (and socially should be) subject to regulation. Because of our fixed cost, K, regulation will be relatively preferred by more efficient (lower γ) firms. In particular, there exists a γ^* such that a firm prefers to be regulated if and only if $\gamma \in [\gamma^\circ, \gamma^*)$. Firms with $\gamma \in [\gamma^*, \gamma^N)$ choose no regulation. Suppose initially that neither of these regions is empty. Note that, as γ falls from γ^* to a value slightly lower, quantity jumps up, because $q^R(\gamma) > q^N(\gamma)$ for all γ (assuming $q^R(\gamma) > 0$[36]), including γ^*. Therefore, we have a quantity interval, $(q^N(\gamma^*), q^R(\gamma^*)]$, in which no firms produce. Likewise, it is clear that there is no quantity (other than zero) at which firms bunch: any positive quantity is produced by at most one type of firm. See Figure 2.

[36]The text corresponds to the case depicted in Figure 2. If the fixed cost K were greater, γ^R would be further to the left, and possibly to the left of γ^N. This poses no complication, however, for it is obvious that γ^* will nevertheless be strictly to the left of γ^R.

Figure 2: Differential Output Taxation of Regulated and Unregulated Output



Furthermore, it is possible that one or both of these two quantity regions is empty. If K is sufficiently small, γ^* may exceed γ^N, which is to say that any firm that chooses to operate will subject itself to regulation — which, of course, is optimal in this case, because the regulation is sufficiently cost-effective that any firm that operates should employ the regulation's technology. (One might think of this as a case in which the optimal exemption, q^E, is zero, but, as explained, no exemption policy is required since firms' voluntary choices whether to be regulated are optimal. This case is tantamount to simple regulation plus pure output taxation.) Furthermore, because the fixed cost is positive, the lowest positive quantity produced is bounded away from zero ($q^R(\gamma^*) > 0$): our quantity gap, therefore, has a lower bound of zero (i.e., $q^N(\gamma^*) = 0$).

If K is sufficiently large, such that the most efficient regulated firm, producing $q^R(\gamma^\circ)$, would earn profits that do not exceed those at $q^N(\gamma^\circ)$ when the firm is unregulated, then under the optimal scheme no firms will be subject to regulation. (That is, we would have $\gamma^* \leq \gamma^\circ$. This case is tantamount to no regulation plus pure output taxation.) This situation, of course, is optimal because regulation, in light of the fixed cost of compliance, is inefficient for all firm types. When this case arises, there is no gap in the range of produced quantities. In addition, note that this case includes a degenerate subcase in which no firms operate. (That is, we may also have $\gamma^N \leq \gamma^\circ$, which is to say that even the most efficient firm has a marginal production cost that, when combined with the tax on unregulated output that is set equal to the external harm of such output, is not below the price p even at zero output.) This subcase arises when unregulated output is sufficiently harmful that even the most efficient unregulated production is

not cost-justified (and, moreover, regulation is too costly to make regulated production cost-effective).

Finally, reflect briefly on employing an exemption, q^E, in this setting. We already know that firms' voluntary choices whether to subject themselves to regulation are optimal. If an exemption is mandatory — which is to say, firms producing above q^E *must* be regulated and firms producing at or below q^E *may not* be regulated (use the harm-reducing technology and thereby be permitted to pay the lower output tax) — then the exemption might reduce social welfare. If the exemption is set within the quantity range in which no firms produce, between $q^N(\gamma^*)$ and $q^R(\gamma^*)$ in Figure 2, it will have no effect because firms would have behaved consistently with that exemption regime in any event. But if the exemption is binding — if it is set above or below that range — it strictly reduces welfare. In the former case (when it is set too high, above $q^R(\gamma^*)$ in the Figure), there are firms that would wish to be regulated but, at their ideal level of regulated output, they will not be. (They would be forced to choose an inefficiently high output while being subject to regulation or a lower output while being unregulated.) In the latter case (when it is set too low, below $q^N(\gamma^*)$ in the Figure), there are firms that wish to be exempt but, at their ideal level of output, they are regulated. Granting them a waiver from regulation would raise social welfare.

In light of the foregoing, we can state:

Proposition 6, on Optimal Regulation with Differential Taxation of Regulated and Unregulated Output:

- *a. The optimal tax on regulated output fully internalizes the externality produced by such output: $t^R = h^R$.*
- *b. The optimal tax on unregulated output fully internalizes the externality produced by such output: $t^N = h^N$.*
- *c. When the taxes on regulated and unregulated output are both (optimally) set as in (a) and (b), optimal regulation is voluntary: that is, firms may freely choose whether to be subject to regulation.*
- *d. The optimal scheme implements the first best: that is, the choice of technology (production with or without regulation) and the level of output (given that choice) are the same as what would be selected by a social-welfare-maximizing planner who could observe each firm's type and command all aspects of its behavior.*
- *e. Under the optimal scheme, it is possible that both regulated and unregulated firms will operate, that all operating firms will be regulated, that all operating firms will be unregulated, and that no firms will operate.*
- *f. Under the optimal scheme, if both regulated and unregulated firms are in operation, there is a range of output, $(q^N(\gamma^*), q^R(\gamma^*)]$, that no firm chooses to produce, where γ^* is the firm type that is just indifferent as to whether to be regulated. All firms producing more output — firms with $\gamma \in [\gamma^\circ, \gamma^*)$ — are regulated; all firms producing less — firms with $\gamma \in [\gamma^*, \gamma^N)$ — are unregulated.*
- *g. When the optimal scheme involves both regulated and unregulated firms in operation, if a mandatory regime with an exemption, q^E, is employed — meaning that a firm of type γ that wishes to produce $q^R(\gamma) \leq q^E$ may not voluntarily subject itself to regulation and one that wishes to produce $q^N(\gamma) > q^E$ cannot opt out of*

regulation — then the optimal exemption q^E must be (anywhere) in the interval $(q^N(\gamma^*), q^R(\gamma^*)]$ in which no firms would voluntarily choose to produce.[37]

5. Conclusion

This article analyzes two largely neglected features of regulation: the use of exemptions and the fact that both regulated and unregulated output involve external harm, which makes the effects of regulation, exemptions, and taxes on all firms' output a first-order concern. As a background to the main analysis, the optimal choices between regulation and no regulation and between regulation and pure output taxation are characterized. Regulation can be superior to no regulation even when regulation is technically inefficient (costs exceed benefits for all firm types at all levels of output) because its output-suppressing effects may raise social welfare. Output taxation unsurprisingly dominates both no regulation and technically inefficient regulation, but the contest with technically efficient regulation is more complicated. For example, higher fixed and marginal costs of regulatory compliance do not necessarily favor output taxation.

The first part of the core analysis examines the optimal level of regulatory exemptions when there is no output taxation, which is typical in practice. Regulation with an exemption can dominate no regulation and may be optimal (compared to other exemption levels) even if no firm produces regulated output, again due to the scheme's effect on output. The optimal exemption can equal zero regardless of the magnitude of the fixed and marginal costs of regulatory compliance. Characterization of an optimal exemption that is intermediate involves two complications: firms jumping down to become exempt, as the exemption is increased, may raise or lower social welfare (they produce output that is more harmful per unit but lower in quantity, and, recall, regulated output is also harmful), and quantity increases by less efficient types of firms clustered at the now-higher exemption reduce welfare, but quantity increases by more efficient types may raise welfare. In addition, higher fixed or marginal costs of regulatory compliance have an ambiguous effect on the optimal exemption level.

The second core section combines regulation (with an exemption) and output taxation. Allowing a tax on (only) exempt output need not raise welfare — its optimal level can equal zero — because such a tax induces the marginal type of firm to jump up to become regulated, which actually can be inefficient because its higher level of output may cause sufficiently greater harm (even though harm per unit of output is lower). Relatedly, in a range of settings, the optimal tax on exempt output is below the level of external harm caused by such output. Allowing instead a tax on (only) regulated output need not raise welfare and, in a range of settings, the optimal tax is below external harm for such output. Here, the reasoning reverses: the incentive of firms to jump down is otherwise socially excessive.

Finally, when exempt and regulated outputs can each be taxed, and at different rates, the results are qualitatively different. Now, optimal taxes fully internalize the respective externalities and the optimal scheme makes firms' decisions whether to be subject to regulation

[37]Or at the lower boundary of that interval.

(employing the more expensive, less harmful technology) entirely voluntary. This scheme implements the first best: what a planner would choose if it could observe all firms' types and dictate their actions. In such a regime, for the case in which some types of firms are regulated and others are not, there is an intermediate range of output in which no firms produce. If a mandatory regulatory regime is employed, the exemption is optimally set (anywhere) in that output range: if it is set higher, some firms would like to be regulated (and would raise social welfare if they were permitted to be), but their optimal output falls short of the exemption level; if the exemption is set lower, some firms would like to be exempt (and would raise welfare if they were), but their optimal output exceeds the exemption level.

Much thinking about regulation, including about the use of exemptions, is incomplete as a consequence of ignoring that firms are not charged for the harm they cause. Unregulated firms obviously cause harm, which is the motivation for regulation in the first instance. And regulated firms usually cause harm as well: indeed, optimal regulation often has this feature for the familiar reason that the marginal cost of control tends to be rising and the marginal benefit falling, so controlling harm becomes ever more costly (and ultimately may not be cost-effective) as regulation becomes more stringent. Because this harm is not priced either — it is free to regulated firms — they produce too much. Moreover, rectifying both omissions greatly complicates the analysis of regulation. As mentioned, higher costs of regulatory compliance do not unambiguously disfavor regulation (versus no regulation), and they do not unambiguously favor a higher exemption (under regulation) — both as a consequence of output effects. These features raise serious questions about cost-benefit analysis that implicitly takes firms' outputs (including decisions whether to operate at all) as given. Also, within an exemption regime, as firms move across the boundary from regulated to unregulated, we have important effects on both sides of the divide that the firms do not take into account, which means that there exist settings in which adjusting an instrument induces inefficient jumping down (to exemption) as well as others in which it induces inefficient jumping up (to regulation).

If we were to properly price output for both regulated and unregulated firms — neither of which is part of standard practice — then these complications disappear, and the decision about which firms should be regulated becomes quite simple. Many of the analytics change and some are even reversed. For example, without taxes, aggregate private marginal costs at any given level of output are higher under regulation by k, but with output taxes that are set optimally, aggregate private marginal costs are lower under regulation by $\Delta h - k$ (which is positive whenever regulation is technically efficient). This difference means that firms that operate under both regimes produce higher, not lower, quantities under regulation. Relatedly, the most efficient firms prefer to be regulated rather than unregulated. Even more, the regulator no longer needs to decide which firms should be subject to regulation (or whether to mandate any regulation at all): Firms bear all social costs either way, so they can be left to decide for themselves.

It is interesting to compare these lessons with the familiar point that ideal corrective taxes are a panacea with regard to externalities. As the analysis in section 2 explores, we are not assuming here that harm itself is observable and thus that taxes can equal harm per se. Rather, only output is observable — but, importantly, so is a firm's technology choice, which tells the

regulator the association between output and harm. Although differences in firms' costs are unobservable, a conventional challenge to regulation, once we allow both types of output taxes, this limitation does not constrain what a social planner can achieve. In this setting, the planner does not need to know even the distribution of firms' types or, really, anything about firms' cost functions. This casts regulatory cost-benefit analysis in a different light. Because optimal regulation is voluntary, cost-benefit analysis becomes moot, or worse, if there is to be a binding rule that forces firms of some types to employ the wrong technology. Of course, one must keep in mind that these results depend on the observability of the technology and of output — assumptions that seem roughly to describe some important regulatory settings but not nearly all of them.[38]

[38]Of course, to impose regulation, it must be observable whether firms comply, even if this entails a nontrivial cost. In this regard, recall that the present model is consistent with the assumption that enforcement is costly, with those costs being charged to regulated firms. Moreover, it is common to make exemptions a function of output, implicitly taken to be observable at essentially no cost, and developed economies already employ a VAT or other forms of taxation on output, and sometimes at differential rates.

References

David P. Baron, Regulation of Prices and Pollution under Incomplete Abatement, *Journal of Public Economics* 28: 211–31 (1985).

_____, Design of Regulatory Mechanisms and Institutions, in Richard Schmalensee and Robert D. Willig, eds., *Handbook of Industrial Organization*, vol. 2 (1989): North-Holland, Amsterdam.

_____ and Roger B. Myerson, Regulating a Monopolist with Unknown Costs, *Econometrica* 50: 911–30 (1982).

Randy Becker and Vernon Henderson, Effects of Air Quality Regulations on Polluting Industries, *Journal of Political Economy* 108: 379–421 (2000).

_____, Carl Pasurka Jr., and Ronald J. Shadbegian, Do Environmental Regulations Disproportionately Affect Small Businesses? Evidence from the Pollution Abatement Costs and Expenditures Survey, EPA National Center for Environmental Economics Working Paper 12–06 (2012).

C. Stephen Bradford, Does Size Matter? An Economic Analysis of Small Business Exemptions from Regulation, *Journal of Small & Emerging Business Law* 8: 1–37 (2004).

William A. Brock and David S. Evans, The Economics of Regulatory Tiering, *Rand Journal of Economics* 16: 398–409 (1985).

_____ and _____, *The Economics of Small Businesses: Their Role and Regulation in the U.S. Economy* (1986): Holmes & Meier, New York.

Charles Brown, James Hamilton, and James Medoff, *Employers Large and Small* (1990): Harvard University Press, Cambridge, MA.

Vidar Christiansen and Stephen Smith, Externality-Correcting Taxes and Regulation, *Scandinavian Journal of Economics* 114: 358–83 (2012).

W. Mark Crain, *The Impact of Regulatory Costs on Small Firms* (2005): SBA Office of Advocacy.

Partha Dasgupta, Peter Hammond, and Eric Maskin, On Imperfect Information and Optimal Pollution Control, *Review of Economic Studies* 47: 857–60 (1980).

R. Mithu Dey and Mary W. Sullivan, Was Dodd-Frank Justified in Granting Internal Control Audit Exemption to Small Firms?, *Managerial Accounting Journal* 27: 666–92 (2012).

Dhammika Dharmapala, Joel Slemrod, and John Douglas Wilson, Tax Policy and the Missing Middle: Optimal Tax Remittance with Firm-level Administrative Costs, *Journal of Public Economics* 95: 1036–47 (2011).

Gunnar S. Eskeland, A Presumptive Pigouvian Tax: Complementing Regulation to Mimic an Emissions Fee, *World Bank Economic Review* 8: 373–94 (1994).

Feng Gao, Joanna Shuang Wu, and Jerold Zimmerman, Unintended Consequences of Granting Small Firms Exemptions from Securities Regulations: Evidence from the Sarbanes-Oxley Act, *Journal of Accounting Research* 47: 459–506 (2009).

Edward L. Glaeser and Andrei Shleifer, A Reason for Quantity Regulation, *AEA Papers and Proceedings* 91(2): 431–35 (2001).

_____ and _____, The Rise of the Regulatory State, *Journal of Economic Literature* 41: 401–25 (2003).

Jonathan Gruber and Botond Köszegi, Is Addiction "Rational?" Theory and Evidence, *Quarterly Journal of Economics* 116: 1261–1303 (2001).

Anthony D. Holder, Khondkar E. Karim, and Ashok Robin, Was Dodd-Frank Justified in Exempting Small Firms from Section 404b Compliance?, *Accounting Horizons* 27: 1–22 (2013).

Erik Hurst and Benjamin Wild Pugsley, What Do Small Businesses Do?, *Brookings Papers on Economic Activity* 73–118 (Fall 2011).

IMF Staff, *Taxation of Small and Medium Enterprises*, Background Paper for the International Tax Dialogue Conference (October 2007).

Michael Keen and Jack Mintz, The Optimal Threshold for a Value-added Tax, *Journal of Public Economics* 88: 559–76 (2004).

Jean-Jacques Laffont, Regulation of Pollution with Asymmetric Information, in Cesare Dosi and Theodore Tomasi, eds., *Nonpoint Source Pollution Regulation: Issues and Analysis* (1994): Kluwer, Dordrecht.

_____ and Jean Tirole, *A Theory of Incentives in Procurement and Regulation* (1993): MIT Press, Cambridge, MA.

Robert E. Lucas, Jr., On the Size Distribution of Firms, *Bell Journal of Economics* 9: 508–23 (1978).

Juan-Pablo Montero, Pollution Markets with Imperfectly Observed Emissions, *Rand Journal of Economics* 36: 645–60 (2005).

Richard J. Pierce, Jr., Small Is Not Beautiful: The Case Against Special Regulatory Treatment of Small Firms, *Administrative Law Review* 50: 537–78 (1998).

Charles R. Plott, Externalities and Corrective Taxes, *Economica* 33: 84–87 (1966).

A. Mitchell Polinsky and Steven Shavell, Pigouvian Taxation with Administrative Costs,

Journal of Public Economics 19: 385–94 (1982).

_____ and _____, Enforcement Costs and the Optimal Magnitude and Probability of Fines, *Journal of Law and Economics* 35: 133-48 (1992).

Al J. Ringleb and Steven N. Wiggins, Liability and Large-Scale, Long-Term Hazards, *Journal of Political Economy* 98: 574–95 (1990).

Steven Shavell, The Optimal Structure of Law Enforcement, *Journal of Law and Economics* 36: 255–87 (1993).

Andrei Shleifer, *The Failure of Judges and the Rise of Regulators* (2012): MIT Press, Cambridge, MA.

Lori D. Snyder, Nolan H. Miller, and Robert N. Stavins, The Effects of Environmental Regulation on Technology Diffusion: The Case of Chlorine Manufacturing, *AEA Papers and Proceedings* 93(2): 431–35 (2003).

Daniel F. Spulber, Effluent Regulation and Long-Run Optimality, *Journal of Environmental Economics and Management* 12: 103–16 (1985).

Appendix

Proof of Proposition 2: As explained in the text, there are two cases to consider. Begin with $\gamma^T > \gamma^R$, that is, when there are low-productivity firms that operate under the (here taken to be optimal) output tax but not under regulation. Welfare is higher under regulation if and only if

$$(A1) \quad -\int_{\gamma^R}^{\gamma^T}\Big[pq^T(\gamma) - \gamma c(q^T(\gamma)) - h^N q^T(\gamma)\Big]g(\gamma)d\gamma$$

$$+\int_{\gamma^\circ}^{\gamma^R}\Big[p\Big(q^R(\gamma) - q^T(\gamma)\Big) - \gamma\Big(c(q^R(\gamma)) - c(q^T(\gamma))\Big)$$

$$-\Big(K + kq^R(\gamma)\Big) + \Big(h^N q^T(\gamma) - h^R q^R(\gamma)\Big)\Big]g(\gamma)d\gamma \; > \; 0.$$

The interpretation is similar to that offered for expression (9), comparing social welfare under regulation with that under no regulation. Nevertheless, the differences are substantial even though the expressions seem nearly identical.

The integrand in the first integral indicates, for each type of firm that operates under taxation but not regulation, its net contribution to social welfare. Because we have an output tax t and, moreover, it is set optimally, equal to h^N, this integrand is necessarily positive (except at the upper limit of integration, where it equals zero). This integral is preceded by a negative sign, so we know that, in this scenario (in which $\gamma^T > \gamma^R$), regulation (rather than output taxation) reduces welfare on account of inducing the exit of firms that contributed more to welfare from their production than they reduced it on account of external harm (which is fully internalized).

For the second integral, consider initially all the terms in the integrand except those relating to external harm. At the upper limit of integration, we compare a regulated firm that earns zero profits (accounting for all costs, including those of regulatory compliance) to a taxed firm that has positive profits. However, as we consider more efficient firms (lower γ's), the comparison is trickier. If $t \leq k$ (recall, a sufficient condition to be in this scenario), then a taxed firm with any γ in the range of integration has greater profits than a regulated firm of that type. But if $t > k$, this could reverse: i.e., sufficiently efficient (low γ) firms may be more profitable under regulation than under taxation.

Similar considerations influence our interpretation of the difference in external harm (the final terms). As when comparing regulation to no regulation, harm per unit of output is lower under regulation than under output taxation. Keep in mind that, although output taxation induces

optimal output for each type of firm given the technology it employs, it does not reduce harm per unit of output, which regulation does. These final terms do not, however, unambiguously favor regulation because (in contrast to the comparison between regulation and no regulation) output could be higher under regulation or under output taxation: the former will prevail if and only if $k < t$, but, since t is set optimally, this arises when $k < h^N$. (All things considered, there is a tradeoff here: when output is indeed higher under regulation, external harm is higher in this respect, but it is also true that harm per unit of output is lower and profits tend to be higher, which raises welfare on account of the other terms in this second integral.)

Let us now more briefly consider the other scenario: $\gamma^T < \gamma^R$. Welfare is higher under regulation if and only if

$$(A2)\quad \int_{\gamma^T}^{\gamma^R}\left[pq^R(\gamma) - \gamma c(q^R(\gamma)) - K - kq^R(\gamma) - h^R q^R(\gamma)\right]g(\gamma)d\gamma$$

$$+\int_{\gamma^\circ}^{\gamma^T}\left[p\left(q^R(\gamma) - q^T(\gamma)\right) - \gamma\left(c(q^R(\gamma)) - c(q^T(\gamma))\right)\right.$$

$$\left.-\left(K + kq^R(\gamma)\right) + \left(h^N q^T(\gamma) - h^R q^R(\gamma)\right)\right]g(\gamma)d\gamma \;>\; 0.$$

The formal differences between expressions (A2) and (A1) are that, regarding the first term, we now have a social welfare difference on account of firms that operate under regulation but not under output taxation (rather than vice versa). Other differences relate to the limits of integration, also reflecting the reversed inequality.

The first integral has ambiguous sign. As with no regulation, at the upper limit (here, γ^R), the integrand is negative, reflecting that the external harm that is nevertheless produced by regulated firms, h^R, is not internalized. (Indeed, not at all, despite the output-tax-like effect of the marginal compliance cost, k: the positive k does indeed reduce output, but k is a real resource cost, so the fact of firms bearing it directly does not indicate that any of h^R is internalized.) For more efficient firms (lower γ, approaching γ^T), the integrand could become positive, but it need not.

The integrand of the second term in expression (A2) is identical to that in the corresponding term in expression (A1), but the interpretation differs because we are now in the scenario in which $\gamma^T < \gamma^R$. As explained previously, this condition requires that t exceed k by a nontrivial amount (in light of the fixed cost K). As a consequence, we know that, for all firm types γ in the range for this integral, profits (even accounting for regulatory compliance costs) are higher under regulation than under output taxation. (At the upper limit of integration, profits

are zero under taxation and positive under regulation, and $k < t$ implies that profits rise faster under regulation than under taxation as γ falls.[39]) Hence, the combination of all terms but those pertaining to external harm is positive. Regarding external harm, however, even though (as always) harm per unit of output is lower under regulation, we now have unambiguously higher output under regulation, making the overall external harm effect ambiguous.

For reasons similar to those adduced under the first scenario, it is again possible that either regulation or output taxation could be superior. Note that, on one hand, the necessary condition for this second scenario — that t (taken to be set optimally) exceeds k by a sufficient amount — guarantees that h^N is not negligible relative to the cost of regulation. Considering cases in which regulation is extremely cheap and highly effective, whereas output taxation merely moderates output (to an optimal extent, given the technology used in the absence of regulation), regulation will be superior when h^N is large. Note also, however, that the condition that k be small relative to t and thus to h^N is of only moderate consolation when Δh is small relative to k, because the regulation is both inefficient with regard to output that is produced and also output is reduced less under regulation than under output taxation (where, moreover, the reduction under the latter is optimal in light of the magnitude of harm).

Taxation of Exempt Output: For firms whose output levels render them exempt, their first-order condition will be as in expression (2) for unregulated firms, except that they now equate their marginal cost, $\gamma c'(q^N(\gamma))$, to $p - t^N$, so their ideal output if unregulated, $q^N(\gamma)$, is lower when $t^N > 0$. In addition, the boundaries between the regions, γ^{NE} and γ^{RE}, change. (For now, we are confining attention to the scenario with three nontrivial regions. Moreover, we are taking q^E as given.) The value of γ^{NE} falls as t^N rises: because $q^N(\gamma)$ falls, the type of firm that maximizes profits if not subject to regulation (but subject to t^N) at an output of q^E will be one that is more efficient. In contrast, the value of γ^{RE} increases as t^N rises: a higher tax makes jumping down to q^E less attractive because profits at q^E fall by $t^N q^E$ and profits at $q^R(\gamma)$ (which exceeds q^E at γ^{RE}) are unaffected. As a consequence, raising t^N causes firms that were at the left end of the middle region (those clustered at q^E) to jump up and become regulated, implying that γ^{RE} is higher, which is to say that the firm just indifferent to becoming regulated will be one with lower regulated profits and thus one that is less efficient.

When there are three nontrivial regions, expression (11) continues to state social welfare, the only differences being, as just stated, that $q^N(\gamma)$, which appears in the third integrand, is lower, and the limits of integration, γ^{NE} and γ^{RE}, change. Accordingly, we can write

[39]Quantity rises at the same rate for a given decrease in γ under both regimes (the modified expression (3) for the output tax regime is the same as the original expression (3), which in turn is the same as expression (7) under regulation).

$$(A3)\ \frac{dW^{R/E,t^N}(q^E,t^N)}{dt^N} = \left[-h^R q^R(\gamma^{RE}) + \left(h^N - t^N\right)q^E\right]g(\gamma^{RE})\frac{d\gamma^{RE}}{dt^N}$$

$$+ \int_{\gamma^{NE}}^{\gamma^N}\left[p - \gamma c'(q^N(\gamma)) - h^N\right]\frac{dq^N(\gamma)}{dt^N}g(\gamma)d\gamma,$$

where

$$(A4)\ \frac{d\gamma^{RE}}{dt^N} = \frac{q^E}{c(q^R(\gamma^{RE})) - c(q^E)},$$

and

$$(A5)\ \frac{dq^N(\gamma)}{dt^N} = -\frac{1}{\gamma c''(q^N(\gamma))}.$$

As when deriving expression (12), mechanically taking the appropriate derivative of expression (11) generates mostly terms that cancel, reflecting firms' profit-maximization decisions.

These expressions can readily be understood in terms of the aforementioned effects of raising t^N. The only integrand in expression (11) that changes is the third, and it falls due to the output reduction for unconstrained, unregulated firms; this is reflected in the second line of expression (A3), with the output reduction given by expression (A5).

The reduction in γ^{NE} has no effect on social welfare for the essentially same reason given previously. Because the pertinent marginal type produces q^E initially, and the second and third integrands in expression (11) have the same value when the third integrand is evaluated at q^E, there is no welfare consequence of changing the boundary between these regions.

Finally, we have the increase in γ^{RE}, reflected in the first line of expression (A3), with the magnitude of this increase indicated by expression (A4). These marginal firms raise their quantity to $q^R(\gamma^{RE})$ as they jump up, into the regulated regime, now causing harm per unit of output of h^R. There is also a social gain because they no longer cause harm per unit of output of h^N on their former output of q^E. Note, however, that in contrast to the analysis in section 3 of the pure exemption regime, we now have a social welfare effect per unit of unregulated output of only $h^N - t^N$ instead of h^N. The reason for this reduction is that, as explained earlier, only externalized consequences enter this expression; any social welfare effects borne by the firm are included in the indifference condition that defines the marginal type γ^{RE}. Put more directly, this type of firm's profits when operating at the higher, regulated level of output (revenue minus production costs minus all regulatory costs) just equal its profits when operating at q^E, and the

latter are no longer just revenue minus production costs because now one must also subtract the output tax it pays. By comparison to the calculus for social welfare, the former omits $h^R q^R(\gamma^{RE})$ entirely, whereas the latter omits only $(h^N - t^N)q^E$.

Having explained expression (A3) for the change in social welfare with respect to the tax on unregulated output, t^N, let us now consider what it tells us about the optimal level of this tax. Begin by examining the second line. The bracketed term in the integrand is negative as long as $t^N < h^N$. This result reflects profit-maximization: as mentioned, firms equate marginal cost, $\gamma c'(q^N(\gamma))$, to $p - t^N$, so the integrand is below marginal profits (which equal zero) by $h^N - t^N$. Because $dq^N(\gamma)/dt^N < 0$, the second term as a whole is positive when $t^N < h^N$, reflecting that raising t^N serves to further internalize an externality. Turning to the first term, we have in brackets, just as in section 3, opposing effects, although when we reach the point at which $t^N = h^N$, this first term is unambiguously negative. (In this regard, note from expression (A4) that, as explained previously, $d\gamma^{RE}/dt^N > 0$.) Hence, if this scenario continues to be applicable (on which more in a moment), the optimum necessarily has $t^N < h^N$.

Consider next whether social welfare rises with t^N if we start at $t^N = 0$ (and we continue to assume that there are three regions, so that expression (11) correctly states social welfare). Although the second term in expression (A3) is positive in this case, the first term is ambiguous. Furthermore, it need not be true that at least some internalization raises social welfare for essentially the reason just given: raising t^N causes some firms to jump up, and this effect could dominate (detrimentally) if h^R is sufficiently large (even supposing that it does not exceed h^N), $q^R(\gamma^{RE})$ is large relative to q^E, and the density of firm types is such that many firms would jump up relative to the mass in the rightmost region, (γ^{NE}, γ^N), with relatively inefficient firms.[40]

Let us now relax the assumption that there are three operative regions. Specifically, consider the case, explored in section 3, where q^E binds on the most efficient firms, but it is not low enough that any of them choose to produce higher, regulated output. In that case, the first term in expression (11) for social welfare vanishes, and the lower limit of integration in the second term becomes γ° rather than γ^{RE}. In this scenario, it is apparent that the optimal level of the tax on unregulated output is $t^N = h^N$. For firms choosing quantities below q^E, whose quantity choices are affected by the tax, we wish to fully internalize the externality. Those producing q^E are unaffected in any event. And, finally, because there is no region of firms producing output above q^E and thus subjecting themselves to regulation, we no longer have the jumping-up effect that, as explained, pushes against this result.

All of the foregoing analysis implicitly assumes that changing t^N does not affect which of these two scenarios applies. As explained in section 3, the applicable scenario depends on the relationship between γ^{RE} and γ°: if the former is smaller, we have only two regions, but if it is larger, we have three. Therefore, if we begin with three regions when $t^N = 0$, we will continue to

[40]That this case is possible follows from the facts that all values are finite and no restrictions were placed on $g(\gamma)$, so that we can consider a case in which $g(\gamma^{RE})$ is arbitrarily large and $g(\gamma)$ in the range (γ^{NE}, γ^N) is arbitrarily small. Note further that the possibility that raising the tax rate from zero is welfare reducing implies that, had we not constrained the tax to be nonnegative, a negative tax (an output subsidy) could be optimal.

have three regions because raising t^N increases γ^{RE}. See expression (A4) and recall the prior explanation. However, it is possible that we would have only two regions when $t^N = 0$ but would switch to three regions as t^N is increased. (Consider the case in which, initially, γ^{RE} is barely below γ°.) When there are (and will continue to be) two regions, we wish to raise t^N all the way to h^N, but when there are three regions, we wish to stop short of h^N. Thus, if two regions become three before t^N reaches h^N, the optimal t^N might be characterized by expression (A3) equaling zero (and there being three regions), but it is also possible that welfare would be falling with t^N once the third region emerges, so the optimal t^N would be such that we are at the boundary between these two scenarios, in which case $\gamma^{RE} = \gamma^\circ$.

Finally, consider the possibility that the *middle* region vanishes. Before we introduced output taxation on unregulated output, suppose that we had an intermediate region $[\gamma^{RE}, \gamma^{NE}]$ in which firms produce q^E. However, as explained, as we increase t^N, γ^{RE} rises and γ^{NE} falls, presenting the question whether they ever meet. The answer is that they can.

It is helpful to begin exploring this set of cases by examining what turns out to be a somewhat broader set of possibilities, those that arise when $t^N > k$. Such cases are of interest because technically efficient regulation requires at a minimum that $k < \Delta h$, and this in turn implies (as a necessary condition) that $k < h^N$. Now, since we are explicitly interested in raising t^N until the point at which it equals h^N, it follows that we need to understand settings in which $t^N > k$.

Backing up slightly, consider $t^N = k$. It follows from firms' profit-maximization decisions that any type of firm γ would choose the same quantities, i.e., $q^R(\gamma) = q^N(\gamma)$, *if* it decides to operate under regulation and no regulation, respectively — ignoring for the moment the effect of the exemption, q^E. Because $K > 0$, however, all firms strictly prefer no regulation. Therefore, once we reintroduce our exemption, we have $\gamma^{RE} < \gamma^{NE}$, resulting in our familiar three regions (unless $\gamma^{RE} < \gamma^\circ$, in which case we have only the aforementioned two regions).

When $t^N > k$, however, (unconstrained) profit-maximizing output is higher under regulation, i.e., $q^R(\gamma) > q^N(\gamma)$, for all γ, conditional on operating in the corresponding regimes. In this case, for K sufficiently large, we still have $\gamma^{RE} < \gamma^{NE}$, and the analysis is as before. But if K is in an intermediate range, that inequality fails and, at some critical γ, we have firms jumping down from $q^R(\gamma) > q^E$ to $q^N(\gamma) < q^E$.[41] (More efficient firms, with lower γ, produce strictly more than q^E, and less efficient firms, with higher γ, produce strictly less than q^E.) Finally, for K in a lower range (including values arbitrarily close to zero), there will be some regulated firms with γ's such that $q^R(\gamma) < q^E$, yet they would earn more profits producing subject to regulation than they would if they produce $q^N(\gamma)$ and are not regulated. In other words, they would prefer to subject themselves to regulation. (And if they are not permitted to do so, and their γ is in the relevant range but sufficiently low, they would produce slightly more than q^E in order to be subject to regulation.) Finally, note that the pertinent critical values for K that divide these subcases will all be higher the greater the degree to which t^N exceeds k, ceteris paribus. (Recall

[41] In this case and the next, it is possible that the optimal unregulated output will be zero, which is to say that the γ under consideration might not be less than γ^N.

that, when $t^N = k$, any positive K is sufficient to maintain our original configuration.)

Taxation of Regulated Output: For firms whose output levels exceed q^E, their first-order condition will be as in expression (6) for regulated firms, except that they now equate their marginal cost, $\gamma c'(q^R(\gamma))$, to $p - k - t^R$, so their ideal output if regulated, $q^R(\gamma)$, is lower when $t^R > 0$. In addition, the boundary between the upper and middle region, γ^{RE}, changes. (In contrast to the previous case, γ^{NE} is unaffected because t^R applies neither to firms producing q^E nor to those producing less.) We primarily confine attention to the scenario with three nontrivial regions: when the upper region vanishes (which here will happen when t^R is sufficiently high, even if that region exists when $t^R = 0$), changes in t^R become moot; also, unlike when we were analyzing t^N, we need not be concerned with a vanishing middle region (t^R combines with k to raise the aggregate marginal cost of regulated output above that of unregulated output). Moreover, we initially take q^E as given.

The value of γ^{RE} falls as t^R rises: the tax makes jumping down to q^E more attractive because profits at $q^R(\gamma^{RE})$ (which exceeds q^E) fall by $t^R q^R(\gamma^{RE})$ whereas profits at q^E are unaffected. As a consequence, raising t^R causes firms that were at the right end of the left region to jump down and become exempt, implying that γ^{RE} is lower, which is to say that the firm just indifferent to becoming regulated will be one that is more efficient.

When there are three nontrivial regions, expression (11) continues to state social welfare, the only differences being, as just stated, that $q^R(\gamma)$, which appears in the first integrand, is lower, and the limits of integration that equal γ^{RE} change. Accordingly, we can write

$$(A6)\quad \frac{dW^{R/E,t^R}(q^E,t^R)}{dt^R} = \int_{\gamma^\circ}^{\gamma^{RE}} \left[p - \gamma c'(q^R(\gamma)) - k - h^R\right] \frac{dq^R(\gamma)}{dt^R} g(\gamma)\,d\gamma$$

$$+ \left[\left(h^R - t^R\right) q^R(\gamma^{RE}) - h^N q^E\right] g(\gamma^{RE}) \left(-\frac{d\gamma^{RE}}{dt^R}\right),$$

where

$$(A7)\quad \frac{dq^R(\gamma)}{dt^R} = -\frac{1}{\gamma c''(q^R(\gamma))},$$

and

$$(A8)\quad \left(-\frac{d\gamma^{RE}}{dt^R}\right) = \frac{q^R(\gamma^{RE})}{c(q^R(\gamma^{RE})) - c(q^E)}.$$

The first term in expression (A6) indicates the welfare consequence of the quantity reduction by regulated firms due to the increase in t^R. The bracketed term in the integrand — the welfare impact per unit change in quantity — is negative when $t^R < h^R$ (and it equals zero when they are equal). The now-familiar explanation is that regulated firms' marginal gain from raising quantity differs from this integrand only in substituting t^R for h^R, and profit-maximizing firms set their marginal profit equal to zero. Because $q^R(\gamma)$ falls as t^R rises (see expression A7), the first term is positive when $t^R < h^R$ and zero at $t^R = h^R$.

The second term is the effect of firms jumping down to the exempt level of output as t^R is increased. The bracketed term indicates that they no longer cause the harm associated with regulated output but instead cause the (larger) harm per unit of unregulated output at the output q^E. However, just as with a tax on only unregulated output, with the former we now have only $h^R - t^R$ instead of h^R as the per-unit social harm because the indifference condition for firms of type γ^{RE} differs from the social calculus only in that the former ignores the uninternalized portion of external harm. Note further that this now-reduced first component is positive when $t^R < h^R$ and zero at $t^R = h^R$. Finally, the bracketed portion of the second term is weighted by the density of firms at γ^{RE} and the rate at which γ^{RE} changes with t^R (see expression A8). (As with expressions 12 and 13, it is convenient to employ a minus sign here: raising t^R causes γ^{RE} to fall, which is to say that the marginal firm is a more efficient firm, with those at the margin jumping down as t^R is increased.)

The lessons for the optimal tax on regulated output, t^R, are apparent. First, if this three-region scenario governs, we know that the optimum has $t^R < h^R$: at $t^R = h^R$, the first term in expression (A6) is zero and the second term is negative.[42] Second, if it is optimal to raise t^R to the point that the left region vanishes (no firms produce regulated output), then any further increase in t^R is immaterial. (Note that, at $\gamma^{RE} = \gamma^\circ$, the first term in expression (A6) becomes zero, but if at the level of t^R that just reaches that point, we still have $t^R < h^R$, the second term is still of indeterminate sign, indicating that a social optimum could have this character.)

Third, it need not be true that the optimal t^R is positive because, even at $t^R = 0$, the second term is ambiguous: raising t^R from zero does unambiguously raise social welfare as a consequence of the output reduction from regulated firms, but the induced jumping down to become exempt could produce a net welfare reduction, and one that exceeds the foregoing welfare gain. To illustrate, consider a case in which h^R is very small (so the first term of expression (A6) and the first component of the second term are likewise small), h^N is very large (it can be arbitrarily high, after all), and $g(\gamma^{RE})$ is large. In that instance, the gains from introducing a positive tax on regulated output are insignificant whereas the harm caused by firms jumping down and producing harmful unregulated output is large.

Fourth, suppose that $q^E = 0$ (which, recall from Proposition 3(c), may be optimal in the

[42] If we do set $t^R = h^R$, we restore the unambiguous result that jumping down is detrimental, in accord with conventional wisdom that, as explained in the introduction, implicitly ignores that regulated output often causes external harm. This restoration arises because, when $t^R = h^R$, although there is harm from regulated output, there is no uninternalized harm.

absence of taxation on regulated output). We no longer have three regions, and the appropriate revision to expression (A6) is that, in the bracketed portion of the second term, the latter component equals zero. In this case, it is obvious that the optimum involves $t^R = h^R$: the only force pushing against this equality was the fact that firms jumping down to the exempt level of output cause (uninternalized) external harm of h^N for each unit of exempt output, q^E; but now we have that $q^E = 0$, so this effect vanishes. The only effect of t^R in this scenario is on the output choices of regulated firms (including the choice whether to operate). Their profit-maximization calculus differs from the social calculus only because of the external harm, h^R, so that externality should be internalized fully.