INSTITUTO DE ECONOMÍA

DOCUMENTO de TRABAJO

566
2022

Fare Evasion and Monopoly Regulation

Martín Besfamille, Nicolás Figueroa y León Guzmán

www.economia.uc.cl · ISSN (edición impresa) 0716-7334 · ISSN (edición electrónica) 0717-7593

Fare Evasion and Monopoly Regulation*

M. Besfamille† N. Figueroa‡ L. Guzmán§

March 1, 2022

Abstract

We consider the regulation of a monopoly facing consumers that may evade payments, an important issue in public utilities. To maximize total surplus, the regulator sets the price and socially costly transfers, ensuring that the monopoly breaks-even. With costly effort, the firm can deter evasion. Under unit demand and fixed quality, price is independent of marginal cost, but increasing in the marginal cost of public funds. When quality is endogenous, we find sufficient conditions that imply a non-monotonic relation between price and marginal cost of public funds. We extend the model to consider non-unit demand and moral hazard.

Keywords: Regulation, natural monopoly, evasion and marginal cost of public funds.

JEL Codes: D42, H2, L43 and L51

*We thank L. Basso, E. Bond, C. Conlon, W. Cont, K. Fuyiwara, A. Galichon, A. García, D. Habermacher, J. Kline, A. McLennan, C. Mezzetti, J. P. Montero, C. Müller, P. O'Callaghan, A. Pereira, J. Poblete, D. Ray, H. Silva, J.P. Xandri, and seminar and conference participants at Queensland, Universidad de los Andes, IIPF (2017, 2020) and Mini TOI (2016) for useful comments and suggestions. We also thank P. Donovan for generously sharing with us the entire database of the UBS survey "Prices and Earnings A comparison of purchasing power around the globe". A. Reyes provided superb research assistance. The authors thank financial support from the Complex Engineering Systems Institute (CONICYT-PIA-FB0816) and from CAF-Banco de Desarrollo de América Latina.

†Instituto de Economía, Pontificia Universidad Católica de Chile and CESifo. Email: mbesfamille@uc.cl.

‡Instituto de Economía, Pontificia Universidad Católica de Chile and ISCI. Email: nicolasf@uc.cl.

§Department of Economics, New York University. Email: lgl266@nyu.edu.

1 Introduction

How to finance a natural monopoly that provides a public service, like electricity distribution or transport? This is a classical regulatory problem, whose solution has been discussed under two different institutional settings. In the Ramsey-Boiteux tradition under full information (see Boiteux 1956, Baumol and Bradford 1970), the monopoly must break-even as an exogenous constraint prevents the regulator to subsidize it. Laffont and Tirole (1986, 1993) analyzed optimal regulation under asymmetric information, when transfers to the firm are authorized but are socially costly because they are funded with distortionary taxation. In any case, the optimal regulation is a second-best one: by setting properly the price above the marginal cost, the firm's deficit is financed inducing a minimal loss in consumer surplus. Moreover, the lower the deficit allowed to be run by the firm or the higher the marginal cost of public funds (mcpf),[1] the higher the optimal price should be. In particular, when the mcpf goes to infinity, the regulated price should converge to the monopoly price.[2]

However, fare evasion is a real-world issue that may qualify these conclusions, in particular in transport systems. Albeit this consumer misbehavior is difficult to measure,[3] many investigations demonstrate that fare evasion is a widespread problem. According to Bonfanti and Wagenknecht (2010), a review of 31 transport systems in 18 countries revealed that 4.2% of passengers were fare evaders. More recent studies show that estimations of fare evasion rates are very heterogeneous, varying from 1.3% in London (UK), 15% in Bogotá (Colombia) to 43% in Reggio Emilia (Italy).[4]

Although different factors influence the individual decision to pay the due price,[5] Troncoso and de Grange (2017) and Porath and Galilea (2020) established that evasion crucially depends upon the level of fares. Hence, when designing the optimal regulation of a transport system, regulators should recognize that the relevance of prices as feasible instruments to mitigate firms' financial deficits is attenuated by the existence and extension of fare evasion. In fact, we conjecture that evasion does more than that; It really *dampens* their use to fund transport systems. This assertion is consistent with

[1] The mcpf is the money measure of the welfare cost of raising an additional dollar of tax revenues. See Dahlby (2008) for theoretical considerations and applications of this concept.

[2] These results also hold in a model with individuals that differ in their willingness to pay for the public service.

[3] Currie and Delbosc (2013) assert that "*(. . .) in 2012 the officially recorded fare avasion rate in Melbourne was 9.3%, yet a survey that same year found that 21% of Melbourne's population (and 34% of public transport users) admitted to fare evading at least once in the past year.*"

[4] Smith (2004) provided similar estimations for electricity theft.

[5] Barabino et al. (2020) review the recent literature that deals with fare evasion in transport systems.

the following piece of evidence. Figure 1 illustrates the correlation between the price of public transport in an unbalanced panel of 77 cities between 1979 and 2018 (adjusted to acknowledge differences in their respective living costs) and estimated values of the mcpf for the respective countries where these cities are located.[6]

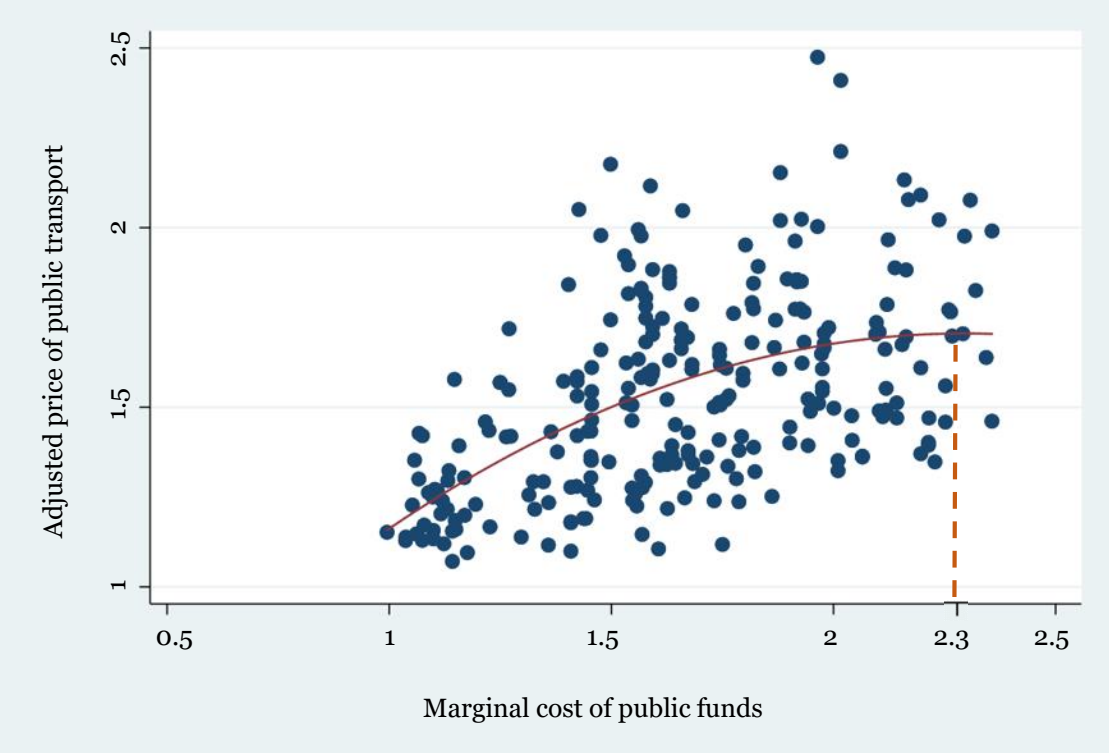


Figure 1: Correlation between marginal cost of public funds and prices of transport systems across the world, 1979-2018. Sources: Union des Banques Suisses (UBS) data and own estimations based on Barrios et al. (2013).

The best fit curve (in red) depicts a hump-shape between the mcpf and adjusted public transport prices, with the peak at 2.29.[7] One could argue that the form of this curve cannot be taken as a proof that the conclusions of the traditional regulation theory

[6] In the Online Appendix, we provide details about the data and the methodology used to build this figure.

[7] To better grasp the quantitative implications of this exercise, we calculate the elasticity of the adjusted price with respect to the mcpf at some cities along the fitted curve. For example, Sao Paulo presents a mcpf of 1.58 (adjacent to the mean of the mcpf, across all years and cities) and an elasticity of 0.47. On the other hand, when the mcpf equals the mean plus 1 standard deviation (close to 1.97), Kiev has an elasticity equal to 0.25. Finally, at the maximum value of the mcpf, 2.52, which is above the peak of the fitted line, Stockholm presents a negative elasticity equal to -0.22.

do not hold in our data, because as the mcpf increases prices may be converging to the monopoly price. This claim is questionable, for the following two reasons. First, the peak is reached for estimated values of the mcpf that are plausible, and clearly far from infinity. Second, according to EMTA (2020), prices of public transport in cities with a mcpf close to the peak of the curve (such as Copenhagen, Helsinki, Oslo, and Stockholm) are heavily subsidized, and so are undoubtedly lower than the corresponding monopoly price.

Naturally, regulators also have other instruments to address fare evasion, like enforcement (Killias et al. 2009), technological or institutional innovations to facilitate payments (Currie and Reynolds 2016), appeals to intrinsic motivations (Ayal et al. 2021), and quality of service improvements (Guarda et al. 2016), among others. Many papers, like Barabino and Salis (2019), presented optimization models that incorporate one or only some of them. But, to the very best of our knowledge, there is no contribution to the literature that characterizes the optimal regulation of a natural monopoly under the threat of evasion, combining the use of prices and these abovementioned instruments in a setting with a distorted tax system.[8] The goal of this paper is precisely to fill this gap.

We consider a model featuring a unit mass of risk-neutral individuals that earn identical income and consume two goods. One of these goods is provided by a monopolist (hereinafter, we refer to it as 'the good'), while the other is the numeraire and is produced in a perfectly competitive market. Individuals choose whether to consume a unit of the good formally or informally (i.e., without paying its price), or not consuming it at all. Individuals are heterogeneous concerning the subjective cost incurred when they are detected evading, and they trade-off the monetary payment and the expected subjective cost. In our model formal buyers never coexist with individuals that spend all their income in the numeraire. However, this is not true regarding the other margin of decision: individuals with relatively low subjective cost evade, while the others consume formally. As expected, when the price increases the mass of evaders also grows.

The monopolist produces the good with a commonly known decreasing average cost technology, for any given level of quality. Crucially, the production cost depends upon total demand, which includes consumption by evaders. To deal with this issue, the firm exerts a costly effort that increases the likelihood of a high level of detection, deterring in this way informal consumption. The monopoly's earnings come from formal consumers and public transfers that depend upon the realized level of evasion.

[8]Focusing on urban transport systems, Basso and Silva (2014) characterized the efficiency between different congestion management policies (which include prices), and their relation to the mcpf. But they do not incorporate fare evasion to their analysis. Informal consumption is not even mentioned in Sherman (1989), one of the main books about the conventional theory of monopoly's regulation, or in more recent surveys that present the advances of the incentive theory of regulation, like Laffont (1994) and Armstrong and Sappington (2007).

Initially, a regulator decides whether the firm will provide the good. If so, it then chooses the price and the level of effort to maximize the expected social welfare, net of transfers to the firm. These transfers are costly because they are financed with distortionary taxation, and thus society bears a mcpf greater than one. In any case, the regulator must induce the monopoly to remain active, because there are very high losses if the latter withdraws from the market after initiating its activities.

We first characterize the optimal regulation when the quality of the good is exogenous. When choosing a higher price, the regulator trades off the extra revenues the firm receives (which are socially valuable because they save on transfers) against the rise in evasion caused by such a price increase. In our setup, there exists a hard bound on prices, above which there are no formal consumers (they either evade or they do not buy the good at all). As long as the price is lower than this bound, production costs are sunk, since they must be incurred whether consumers pay or evade. In particular, the regulator does not deal with the typical Ramsey-Boiteux concerns. Instead, it sets the price to choose optimally the expected marginal formal consumer. We show that the optimal price and effort are both increasing in the mcpf, up to the point where the former cannot rise any more. Further increases in the mcpf lead to higher enforcement, until a level where the firm should be shut down.

We then assume that the quality of the good is endogenous, and thus becomes another regulatory instrument. Increasing quality is obviously costly, but it enables to raise the maximum price that can be charged by the firm. When the price is below this cap, an increase in the mcpf leads to higher prices and effort levels, but to a lower quality. We present sufficient conditions ensuring that when the maximum price is reached, however, the optimal mechanism involves even lower prices, accompanied by a deterioration of the quality. The regulator chooses to save money on quality, even if this leads to lower prices, since the latter is at least mitigated by lower evasion. Therefore, under these circumstances, we recover a hump-shape between the mcpf and the optimal price, as depicted in Figure 1. To verify the robustness of these analytical results, we present numerical simulations of the model.

Finally, keeping the quality exogenous, we generalize the model in two directions. First, we relax the assumption of unit demand, and we derive the optimal pricing rule, establishing its relation to the Ramsey-Boiteux formula. Here, the classical distortion introduced by higher prices, through inefficiently low consumption, only applies to formal consumers. On the other hand, some agents become evaders and consume an inefficient amount of the good, which enlarges the firm's deficit and thus forces the regulator to increase the costly transfers to the firm. We find that the pricing rule formula

combines two terms: a term reflecting the Ramsey-Boiteux concern, and a new one, related to the financing of the extra costs caused by informal consumption. Notably, the mcpf does not appear explicitly in the second term. Second, we assume that effort is non observable. In this case, besides the fact that the optimal regulation incorporates the informational rent to let to the firm when evasion is low, the results are similar to those previously obtained. In particular, the pricing rule is not distorted by the presence of moral hazard, which is reminiscent to the "dichotomy property" exposed by Laffont and Tirole (1990). By comparing our finding to their, we explain why in the context of our model this result is unexpected.

1.1 Related Literature

Our paper connects with different strands of the literature. First, it is related to the theoretical work focusing on regulatory problems in developing countries. As Estache and Wren-Lewis (2009) claim, in those countries "*There is a clear concern that public institutions are unable to collect adequate revenue to allow direct subsidies when the ability of consumers to pay for services is limited.*" Our paper contributes to the study of this issue in at least two ways. We explicitly take into account that weak institutional settings allow consumers to choose whether to pay or not for the service. Then, by highlighting fare evasion as an important matter for regulators, we complement the analysis made by Laffont (2005) on the optimal way to solve the tension mentioned in the previous citation. Moreover, we bring this topic to the forefront of the regulatory discussion, since this tension is becoming relevant also in more advanced economies.

The two closest contributions to our paper, both from the methodology they use and the results obtained, are Silva and Kahn (1993) and Buehler et al. (2017). The former present a model with a firm that provides public transportation. Individuals decide to consume formally or informally, or not consuming it at all. In the last section of their paper, they consider heterogeneous individuals in terms of their preferred number of rides. To deter evaders, the firm can exert a costly monitoring effort. As in our model, the detection probability depends upon effort, but also on the individual and total number of travels. The firm chooses an individually-rational, incentive-compatible mechanism comprising the number of formal consumers and service quality, as well as the monitoring level and fare (not the fine, which is fixed by law). As in our paper, the optimal mechanism depends upon the gains and costs of increasing the mass of subscribers, and thus involves free riding for low users and selling formally the service to the others. Moreover, the firm

underprovides transportation with respect to the first best case (i.e., when monitoring evaders is costless) and shares equally the total cost among users. In particular, the authors claimed that the subscription fee is higher than under the first best.[9] Besides some differences in the formalization of the production costs, the IO approach of this article contrasts with our regulatory framework. As the authors recognized in the concluding remarks, their solution is not equivalent to a social planner's outcome. Moreover, they do not examine the possibility of an equilibrium with evaders and non-subscribers, situation that in our paper generates a cap on the maximum price to charge. Finally, as subsidies are not allowed, the firm has to break even and price according to its average total cost.

Buehler et al. (2017) develop a model where a profit-maximizing firm chooses the price paid by formal consumers, as well as the fine collected from detected evaders. Risk neutral individuals consume one unit of the good and differ in their willingness to pay. Individuals take the same three decisions than in the previous paper and have identical moral costs. The firm engages in price discrimination. Indeed, formal consumers pay more because, although some individuals evade when they face a price increase, the firm anticipates that some income will be recovered from their fines. When the authors consider a welfare maximizing firm, price discrimination vanishes, and thus marginal cost pricing emerges at the optimum. Although financial losses can occur, Buehler et al. (2017) did not formalize the funding of such deficits, neither did incorporate shutdown into their analysis. These two crucial features of our model explain why we find contrasting results.

We also build on the literature that analyzes individual fare evasion. Boyd et al. (1989) consider heterogeneous risk-neutral individuals that, based on the perceived detection probability, decide to evade payments. A firm (or the government) chooses the inspection level, that ultimately yields the actual rate of evasion detection. Taking prices and fines as given, they find the optimal level of inspection. In a similar vein, Kooreman (1993) considers risk-averse individuals that decide whether to evade the payment of a farecard. Individuals differ in their risk aversion. He derives a lower bound on the inspection rate that pushes individuals to evade. He then tests some of its comparative statics results. Concerning the formalization of evasion, our paper differs from these contributions by not incorporating fines into our model, and in turn assuming that individuals differ with respect to their subjective cost borne if caught.

More recent research investigates empirically the factors that affect fare evasion. Killias et al. (2009) analyze how increasing detection modifies the incentives to evade by studying a field experiment in Zurich, in which ticket inspections were reintroduced to

[9] Fraser (1996) showed that the derivation of this last result had a drawback, and presented necessary and sufficient conditions to ensure it.

the public transportation system. They find that evasion levels dropped, even in hours of the day when no new inspections took place, but that the fare evasion rate did not respond to increasing fines. Using a preferences survey applied to apprehended evaders of Transmilenio in Bogotá, Guzman et al. (2021) highlight the importance of incorporating observable and latent variables to understand fare dodging. In particular, they obtained that this misbehavior is negatively related to the satisfaction of the transport system, but that norm's nonconformity moderates this result. Although our approach is theoretical and has a normative perspective, we adopt many assumptions that can be justified by these empirical findings.

The layout of the remainder of the paper is as follows. Section 2 describes the model, and discusses some assumptions. Section 3 analyzes optimal regulation when quality is exogenous. Section 4 incorporates quality as a regulatory instrument. Section 5 extends the basic model, and considers non unit demand for the regulated good. Section 6 concludes. All proofs and the case with moral hazard are relegated to the Appendix. An Online Appendix contains further details about Figure 1.

2 The Model

We consider an economy with a unit mass of individuals. Two goods are produced. The good x is provided by a monopolist, while the numeraire m is produced in a perfectly competitive market.

2.1 Individuals

All individuals earn income y, and have unit demand for good x. Let $b(q)$ be the benefit that each individual derives from consuming a unit of good x, where q denotes the quality with which this good is provided. The strictly increasing and concave function $b(.)$ satisfies Inada conditions. We further assume that individuals have quasilinear utilities on good m.

Individuals can consume good x in two ways: formal and informal. Formal consumption is done at price p,[10] while informal consumption *does not*. Informal consumption leads to an expected loss γz, where γ is the probability of being detected

[10] We assume that income y is sufficiently high, so that all individuals can afford good x.

evading the payment of p, and z is the money measure of the subjective cost faced in this circumstance, which may be caused by a reputation loss, waste of time or social stigma. From now on, we call z the "reputation cost", which is heterogeneous across individuals. Formally, z is distributed according to a strictly positive, continuous, and bounded density function f on $[0, \overline{z}]$, with cumulative distribution F. An individual might also decide not to consume good x and spend his entire income on m. In this case, we say that such an individual "leaves the market" for good x.[11]

Utilities of formal and informal consumers are denoted by U^F and U^I, respectively. Let U^O be the utility of an individual that leaves the market. These utilities are given by

$$U^F \equiv b(q) + y - p, \quad U^I \equiv b(q) + y - \gamma z, \quad U^O \equiv y. \tag{1}$$

We denote the individuals' indirect utility function by $\mathcal{U}_\gamma(p, q, z)$. We assume that, when they are indifferent between consuming or not good x, individuals choose to consume it. Similarly, when they are indifferent between consuming good x formally or informally, individuals choose to purchase it.

Comparing expressions (1), it is straightforward to show that if $p \leq b(q)$, no individual leaves the market for good x. Otherwise, no individual consumes the good formally. In other words, formal buyers never coexist with individuals that spend their whole income in the other good m. Moreover, let $\widehat{z}_\gamma(p) \equiv \frac{p}{\gamma}$ be the value of the reputation cost that makes an individual being indifferent between consuming good x formally and informally. As expected, individuals with relatively high reputation costs, $z \geq \widehat{z}_\gamma(p)$, consume good x formally.[12]

Figure 2 summarizes the previous discussion. Panel (a) depicts, in the (p, q) plane, the regions where different ways of consuming good x coincide. Informal consumption is feasible on both sides of the red boundary $b(q)$. But, for a given quality, demanding good x formally dominates leaving the market whenever $p < b(q)$. The opposite holds to the left of $b(q)$. At any point in the region F-I (like, for example, in A), panel (b) illustrates evasion decisions over the reputation cost line.

[11] We will hereinafter characterize individuals based on how they decide to purchase (or not) the good x.

[12] We further assume that $\overline{z} > b(q)/\gamma_\ell$, which implies that there are always individuals that consume good x formally.

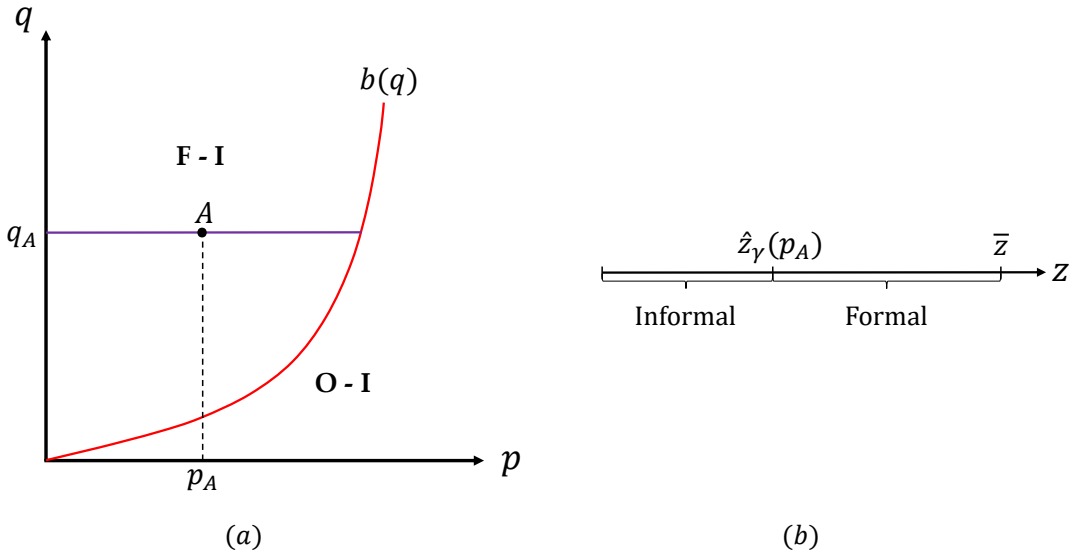


Figure 2: Individuals' decisions

2.2 The Firm

The monopolist produces good x using a decreasing average cost technology, for any given quality q. Specifically, the production cost is given by

$$C(X, q) = K + c_1\left(X^F + X^I\right) + c_2 q, \tag{2}$$

where K is a fixed cost, c_1 is the (constant) marginal production cost, c_2 is the (constant) marginal cost of quality, and X^F, X^I denote formal and informal aggregate consumption of good x, respectively. Therefore, the production cost depends upon total demand $X = X^F + X^I$, which includes consumption by evaders. To simplify notation, we only write $C(X)$ until quality becomes another regulatory instrument in Section 5. We assume that fixed and marginal costs are common knowledge.

For institutional reasons, the firm is unable to levy fines directly. However, it can deter informal consumption through a detection technology. The probability of catching evaders γ can adopt two values, γ_ℓ and γ_h, with $0 < \gamma_\ell < \gamma_h < 1$. By exerting a costly effort e, the firm affects the probability that $\gamma = \gamma_h$, $\rho(e)$.[13] The strictly increasing and

[13]For example, effort e can represent the number of inspectors the monopolist hires to catch evaders. The probability γ illustrates the fact that the effective rate of detection also depends upon other factors, out of the firm's complete control.

concave function $\rho : [0, \infty) \to [0, 1)$ is such that $\rho(0) = 0$.

From (1), we can see that the realization of γ has an impact on the individuals' decisions regarding consumption of x, and thus on the number of formal and informal consumers. We then write X^F_γ, X^I_γ and X_γ.

The firm's earnings come from formal consumers and transfers from the regulator, which can depend upon the realized level of evasion. As it is equivalent, and for the sake of simplicity, we make transfers contingent on the current probability of detection γ. For a given realization of γ, the firm's ex-post utility is given by

$$V_\gamma(p, e) \equiv pX^F_\gamma - C(X^F_\gamma + X^I_\gamma) + T_\gamma - \theta e, \tag{3}$$

where T_γ stands for transfers received from the regulator, and θe is the effort cost. We assume that the firm can leave the market at any time, and we normalize its outside option to zero.

2.3 The Regulator

First, the regulator decides if good x will be provided, by letting the firm operate or shutting it down. If the market is active, the regulator chooses a price p, and contingent, non-negative transfers T_γ.[14] As in Laffont and Tirole (1986, 1993), the regulator maximizes the expected value of the social welfare

$$W_\gamma(p, e) \equiv CS_\gamma(p) + V_\gamma(p, e) - (1 + \lambda)T_\gamma, \tag{4}$$

where

$$CS_\gamma(p) \equiv \int_{[0,\bar{z}]} \mathcal{U}_\gamma(p, q, z)dF(z),$$

stands for net consumer surplus. In order to raise T_γ, the government taxes other (non-modelled) sectors of the economy in a distortionary way. Let $\lambda > 0$ represent the deadweight loss of taxation, and $1 + \lambda$, the marginal cost of public funds. Solving for T in (3), and replacing in (4), we rewrite $W_\gamma(p, e)$ as

$$W_\gamma(p, e) = S_\gamma + \lambda pX^F_\gamma - (1 + \lambda)[C(X_\gamma) + \theta e] - \lambda V_\gamma(p, e), \tag{5}$$

[14] Since the regulator always chooses a price below the determined by an unregulated monopoly, in this context fixing the price is formally equivalent to price-cap regulation.

where S_γ is the gross consumer surplus. The regulator values positively income from sales to formal consumers (because they decrease the need to make transfers), but dislikes leaving rents to the firm (because they are socially costly).

We assume that society bears very high costs if the firm withdraws after being allowed to operate.[15] Therefore, in order to induce it to stay active under any circumstance, the regulator must satisfy the firm's ex-post voluntary participation constraints $V_\gamma(p, e) \geq 0$.

2.4 Timing

At $t = 0$, Nature draws z from the distribution F. Then, at $t = 1$, the regulator decides whether the firm will provide or not good x. If so, at $t = 2$, the regulator sets the price p and contingent transfers T_γ. At $t = 3$, after observing the regulator's decision, the firm chooses the level of effort e. At $t = 4$, individuals learn about the realization of γ and decide either to consume or not good x, and if they consume, whether to do it formally or informally.[16] Finally, production takes places as to meet total consumption, and all payoffs are realized. Figure (3) summarizes the timing.

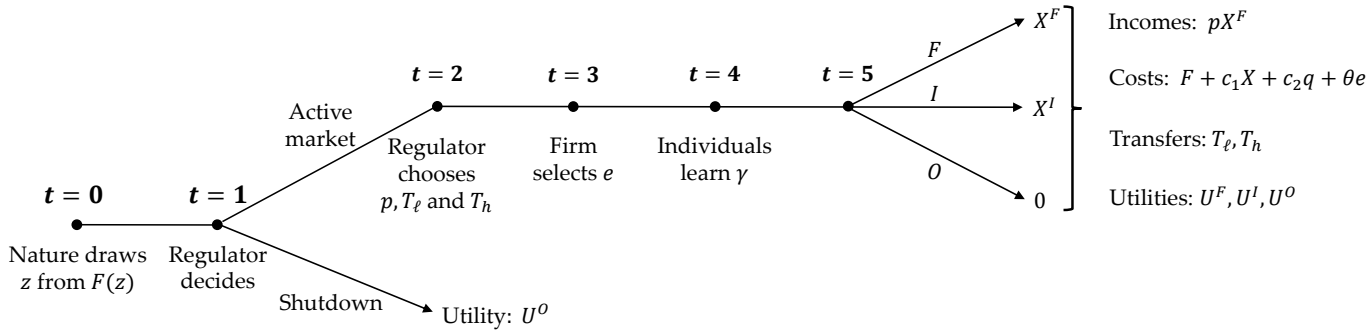


Figure 3: Timing of the model

[15] If good x represents public transport, many long term decisions depend upon its normal operation. Consider for example urban design or the localization of firms. In case of a collapse of this public service, the regulator may face social discontent, bankruptcies, etc.

[16] The fact that we assume that individuals observe the detection probability is consistent with Bucciol et al. (2013), who found that passengers' beliefs on ticket inspection frequency were very close to the actual figures.

2.5 Discussion

Some features of the model deserve some comments. First, we assume that individuals have unit demand for good x. For some public services, this normalization seems reasonable, since individual demand is almost constant, conditional on purchase. To illustrate this assertion, Table 1 shows the average weekday number of rides in New York subway, for the last 6 years.[17]

Table 1: Subway ridership in New York

Year	2013	2014	2015	2016	2017	2018
Population	19,624,447	19,651,049	19,654,666	19,633,428	19,589,572	19,530,351
Average weekday rides	5,465,034	5,597,551	5,650,610	5,655,755	5,580,845	5,437,587
Average weekday rides, per capita	0.278	0.285	0.287	0.288	0.285	0.278

Source: : Metropolitan Transportation Agency, http://web.mta.info/nyct/facts/ridership/; United States Census Bureau.

Similar figures can be obtained for different transport modes, in other cities. But there are other public services where evasion is also an issue and their demand is not constant, like electricity. Therefore, we relax this assumption, and consider that individuals have an elastic demand for good x in Section 6.

Second, we assume that the monopoly cannot impose fines to individuals that are caught evading. Indeed, in some countries, firms that provide public services are either not allowed to fine evaders directly or they have to rely upon a costly and uncertain judicial procedure to do so.[18] Also, fines do not seem to have a big impact over the evasion decision, as found by Killias et al. (2009). More importantly, they are seldom an important source of income, either for providers of public services or for the government, and even less if judicial costs are considered. Therefore, as our focus is the relation between the firm's pricing and the extent of evasion, not incorporating fines does not seem to be too restrictive.

Individuals face an idiosyncratic cost z if they are caught evading. As we have

[17] To obtain the per capita ridership, we divide the total number of rides by the yearly estimated population of the State of New York.

[18] For example, in Chile, to fine evaders, inspectors from private providers need to be accompanied by the police. And even if caught, evaders might not be penalized after all. Calvo (2015) shows that, in all *comunas* of Santiago, the ratio of sanctions to inspections is less than 9 percent. But this incapacity to fine fare evaders is not a distinctive feature of developing countries, as the study of Bijleveld (2007) in the Netherlands revealed.

already mentioned, this cost represents the social stigma and reputation loss due to being detected,[19] and the waste of time and disutility from subsequent prosecution. All these act as deterrents for evasion. It has been documented that moral costs or social sanctions do matter for individuals that evade the payment for public services (see, among others, Sterner and Sheng 2013, Dai et al. 2019 and Ayal et al. 2021). Moreover, with the goal to affect the reputation of a particular individual, enforcement authorities have sometimes exposed evaders in the social networks.[20]

3 Regulation with exogenous quality

In this section, we characterize the optimal regulatory scheme, when the quality of good x is exogenous. We assume that effort e is observable and contractible, and thus that the regulator can impose its level to the firm.[21]

We first analyze the case of an active market, and then we incorporate the option of shutting down the firm. Conditional on delivering good x, the regulator chooses the price, the effort and the firm's rents to maximize the expected social welfare, while inducing the firm's ex-post voluntary participation. Formally, the problem faced by the regulator is

$$\max_{p,e,V_\ell,V_h} \mathbb{E}W \equiv \rho(e)W_h(p,e) + (1-\rho(e))W_\ell(p,e) \tag{6}$$

$$\text{s.t. } V_h(p,e) \geq 0, \quad (VP_h)$$

$$V_\ell(p,e) \geq 0 \quad (VP_\ell).$$

To simplify notation, we let $W_i(p,e) = W_{\gamma_i}(p,e)$ and $V_i(p,e) = V_{\gamma_i}(p,e)$, for $i \in \{h,\ell\}$.

Both the net consumer surplus and the firm's utilities depend directly on the price and the effort chosen by the regulator, and also indirectly, through the individuals' evasion

[19] In some cases, individuals even feel harassed by inspectors. See https://www.theage.com.au/national/victoria/fare-evasion-on-melbourne-public-transport-at-lowest-recorded-level-20150102-12gty3.html

[20] In Argentina, the electricity company *Edenor* published in January 2019 on *Twitter* that one of their inspections found an illegal electricity installation in a franchise of Maru Botana, a well known baker and TV presenter. After a couple of months, Maru Botana decides to settle out of court and pay the corresponding fine. See https://www.clarin.com/sociedad/edenor-escracho-maru-botana-colgarse-luz-locales_0_thDI9lXJ_.html.

[21] In Appendix B, we relax this assumption and we consider the case of non-observable effort. Now the regulator has to provide incentives to induce the firm to exert effort, which becomes more expensive because of informational rents.

decisions.[22] Let's denote by p^E, e^E, V_h^E and V_ℓ^E the solutions to (6). The following lemma outlines some of their properties.

Lemma 1 *Optimal regulation requires* $V_\ell^E = V_h^E = 0$ *and* $p^E \leq b(q)$.

Since transfers, and therefore rents to the firm, are costly, the regulator designs a scheme which grants no rents to the former regardless of the realization of γ. Furthermore, optimal regulation never involves $p > b(q)$. Such a situation would be dominated by setting $p = b(q)$, which leaves total consumer surplus constant, but reduces the firm's financial needs by bringing individuals out of the market into formal consumption. Note that quality, even if it does not affect evasion decisions, plays an important role in the optimal regulatory scheme. The fact that nobody purchases the good when $p > b(q)$ gives rise to an endogenous price cap.

As $p^E \leq b(q)$, all individuals consume good x. These results allow us to rewrite (6) as

$$\max_{p,e} \mathbb{E}W = \underbrace{y + b(q)}_{\text{(a)}} - \underbrace{\mathbb{E}\int_0^{\widehat{z}_\gamma(p)} \gamma z dF(z)}_{\text{(b)}} + \underbrace{\lambda p \mathbb{E}\left[1 - F(\widehat{z}_\gamma(p))\right]}_{\text{(c)}} - \underbrace{(1+\lambda)\big(C(1) + \theta e\big)}_{\text{(d)}} \qquad (7)$$

$$\text{s.t.} \quad p \leq b(q),$$

where expectations are taken with respect to γ, whose distribution is determined by e. Block (a) captures the fact that all individuals receive a benefit $y + b(q)$ from consumption of the numeraire and good x, regardless of their behavior. Hence, all trades-off that the regulator faces are captured by (b)-(d). Block (b) corresponds to the expected aggregate welfare loss from reputational costs borne by evaders. Block (c) represents the expected social value of revenues from formal sales, stemming from a reduction of costly transfers. Finally, (d) corresponds to the social value of production and effort costs. Since everyone consumes good x, this part of the problem is independent of p.

In order to characterize the optimal regulation, we need to understand how reputational costs and revenues (blocks (b) and (c) in (7)) change with e. Effort pushes the latter upwards, since it increases the likelihood of the detection probability γ_h, reducing the number of evaders, and thus increasing formal revenues. The impact of e on block (b) is more nuanced. By differentiating the aggregate reputational cost with respect to effort

[22] We do not restrict transfers to be positive. Since this is the relevant case to consider, we impose (when necessary) sufficient conditions to ensure that this holds in the optimal regulatory scheme.

we obtain $\rho'(e)\Omega(p)$, where $\Omega(p)$ is given by

$$\Omega(p) \equiv \gamma_\ell \int_0^{\widehat{z}_{\gamma_\ell}(p)} z dF(z) - \gamma_h \int_0^{\widehat{z}_{\gamma_h}(p)} z dF(z) = \underbrace{(\gamma_\ell - \gamma_h) \int_0^{\widehat{z}_{\gamma_h}(p)} z dF(z)}_{\text{Inframarginal effect}} + \underbrace{\gamma_\ell \int_{\widehat{z}_{\gamma_h}(p)}^{\widehat{z}_{\gamma_\ell}(p)} z dF(z)}_{\text{Marginal effect}},$$

and represents the impact of a higher likelihood of γ_h on the expected reputational costs borne by evaders. On the one hand, there is an inframarginal effect: consumers that always evade experience the reputational cost with higher probability. On the other hand, there is a marginal effect: for individuals that only evade when $\gamma = \gamma_\ell$, a higher likelihood of γ_h discourages them from being informal, and thus they avoid suffering the reputational cost. We make assumptions on the primitives of the model so that effort decreases reputational costs. Moreover this effect, on its own, is not big enough to justify exerting effort.

Assumption 1 *The density function f is log-concave, with $\bar{z}\frac{f'(\bar{z})}{f(\bar{z})} > -1$.*

Assumption 2 *The function $\rho(e)$ satisfies $\rho'(0) < \theta/\Omega(b(q))$.*

These assumptions enable us to prove the following lemma.

Lemma 2 *The reputational impact of effort on the expected social welfare satisfies the following inequalities: $0 < \rho'(e)\Omega(p) < \theta$.*

This lemma is consistent with our intention of modelling effort as an instrument that helps to finance the firm's deficit. *Ceteris paribus*, increasing effort has a positive reputational effect on social welfare, in the sense that the above mentioned marginal effect always dominates.[23] However, this reputational benefit is not sufficient, on its own, to induce the firm to exert a positive amount of effort, because of its marginal cost θ. Therefore, the financial role of effort, through evasion deterrence, is the driving force of its utilization.

The following proposition characterizes the optimal interior price-effort scheme.

Proposition 1 *When $p^E < b(q)$, the price p^E and the effort e^E are characterized by*

$$\mathbb{E}\left[f(\widehat{z}_\gamma(p^E))\widehat{z}'_\gamma(p^E)(1+\lambda)p^E\right] = \lambda \mathbb{E}\left[\int_{\widehat{z}_\gamma(p^E)}^{\bar{z}} dF(z)\right] \tag{8}$$

[23] In other words, even if the regulator is utilitarian and puts equal social weights on formal and informal consumers, he prefers to face less evaders.

$$\rho'(e^E)\left(\Omega(p^E)+\lambda p^E\int_{\widehat{z}_{\gamma_h}(p^E)}^{\widehat{z}_{\gamma_\ell}(p^E)}dF(z)\right)\leq(1+\lambda)\theta, \textit{with equality if } e^E>0. \tag{9}$$

Equation (8) highlights the marginal benefits and costs of an increase in p. The right-hand side corresponds to the marginal benefit, which is the extra revenue coming from inframarginal consumers. These resources are valued at λ, since increasing revenues allow for a reduction in transfers. The left hand side corresponds to the marginal cost. There is a mass $f(\widehat{z}_\gamma(p^E))\widehat{z}'_\gamma(p^E)$ of new evaders, each one causing a direct financial loss of λp^E. Moreover, these individuals bear a reputational cost of $\gamma\widehat{z}(p^E)$, which is equal to p^E for the *marginal evader* (who is indifferent between facing the reputational cost and paying for formal consumption).

Examining (8), we can assert that, to attenuate the use of socially costly transfers to finance the firm, the optimal price p^E is always strictly positive. But this policy has a cost: it exacerbates evasion. Indeed, unlike standard monopoly regulation, individuals do not stop buying when they face a price increase: they just stop paying. Since at any price total consumption $X=1$, the marginal cost of production c_1 plays no role; all production costs are fixed. Therefore, neither average cost pricing nor Ramsey-Boiteux-Laffont-Tirole regulation are optimal.[24] Instead, the logic of the regulator consists in fixing the price to choose optimally the expected marginal formal buyer.

To determine the optimal level of effort, in (9) the regulator balances the marginal increase in the likelihood of additional revenues accruing from less evasion, net of the change in reputation costs faced by evaders, and its social marginal cost.

The following proposition presents the main comparative statics results of an interior optimal regulatory scheme.

Proposition 2 *Assume that $p^E<b(q)$. When $e^E>0$, the price p^E and effort e^E decrease with θ, and increase with γ_h and λ. Moreover, the expected optimal transfers $\mathbb{E}T^E_\gamma$ increase with θ, and decrease with γ_h and λ.*

At the optimum, when $e^E>0$, price and effort are *complements*. Indeed, as p^E goes up, more individuals consume informally, and thus the marginal benefit of increasing effort to deter evasion increases. The comparative statics with respect to θ and γ_h result from this complementarity. When the marginal cost of effort θ increases, e^E should optimally

[24] If individuals could not evade, it is straightforward to show that, at the optimum, the regulator makes no transfer to the firm and charges average cost pricing.

decrease, pushing the expected number of evaders up. To attenuate this rise in evasion, it is optimal to decrease p^E as well. On the other hand, when γ_h increases, the marginal benefit of effort goes up, pushing e^E upwards. As a consequence, the expected number of evaders decreases, price becomes a less costly way to finance the firm, and thus it should be raised. This contradicts the assertion, often made in administrative and political circles, that an improvement in the detection technology and the resulting decline in evasion could foster a fare reduction in public transport.[25]

Finally, as λ goes up, the social cost of transfers increases, making more attractive the use of higher prices to finance the firm. The same happens with effort, which deters evasion. The aforementioned complementarity between price and effort reinforces these effects, pushing both regulatory instruments further upwards.

The proposition also shows that, as increasing λ affects the marginal value of prices and effort as regulatory tools, their changes must be compensated through transfers, whose expected value decreases but whose social cost grows.

Unlike the optimal price, the regulator does not always set a strictly positive level of effort, as the next corollary states.

Corollary 1 *There exists a threshold $\lambda_1 > 0$ such that, if $\lambda > \lambda_1, e^E > 0$. Otherwise, $e^E = 0$.*

For sufficiently low values of λ, the raising social cost of transfers is not so important. Therefore, the marginal benefit of effort is lower than its social cost, and thus $e^E = 0$. As the marginal cost of public funds increases, additional revenues are more valuable, and thus effort becomes positive.

Next, we characterize non-interior optimal regulatory schemes and we present their main comparative statics properties.

Proposition 3 *If $p^E = b(q)$, the optimal level of effort e^E is given by the first-order condition (9), and, when it is strictly positive, it increases with λ. On the other hand, the expected optimal transfers $\mathbb{E}T_\gamma^E$ decrease with λ.*

As λ rises, the optimal response to mitigate the growing social cost of transfers is to increase effort e^E, since the price p^E cannot escalate further. Although this pushes the cost of the firm up, increment that must be financed with additional transfers (for any

[25] *"Reducing the number of people who travel without a ticket is not only in our interest as the operator, but also in the interest of our fare-paying customers. Few of us want to pay more for our tickets because some people avoid paying (...)"* See https://www.northernrailway.co.uk/legal/penalty-fares.

realization of γ), it also limits the extent of evasion, implying an increase in the firm's revenues. This in turn decreases the above mentioned need for extra transfers. Despite this fact, the expected social cost of the optimal transfers raises with λ.

This proposition implicitly assumes that there are values of the deadweight loss λ such that the market of good x is active and the optimal price $p^E = b(q)$. But such existence is not guaranteed a priori: it could be the case that the regulator *consistently* (i.e., for any set of primitives of the model) prefers to shut down the firm before the price reaches the cap.

In fact, this is not true. In the Appendix we show that we can find parameter conditions such that, for high values of λ, i.e., $\lambda \geq \widehat{\lambda}_q$, the regulator lets the firm to operate and sets its price at the boundary. But why would the regulator charge a price that cancels the net consumer surplus of good x of formal buyers? Because their expenditures finance a large fraction of the costs and enables the firm to deliver the good to all individuals, in particular to evaders, whose strictly positive surplus has social value.

Finally, we incorporate into our analysis the option that good x may not be provided. The next corollary naturally follows from the previous proposition.

Corollary 2 *There exists a threshold $\overline{\lambda}_q$ such that, when $\lambda > \overline{\lambda}_q$, the firm is optimally shut down.*

Based on one possible case characterized in the Appendix, we illustrate these last results. For a given quality, Figure 4 depicts one possible path of the optimal price p^E, as a function of the deadweight loss of taxation. First, it increases if $\lambda < \widehat{\lambda}_q$. Then it remains constant when $\lambda \in [\widehat{\lambda}_q, \overline{\lambda}_q]$. Finally, for higher values of λ, the regulator shuts down the firm.[26]

[26]The other possible case of figure is an increasing optimal price, but that never reaches the boundary $b(q)$ because the firm is shutdown before that happens.

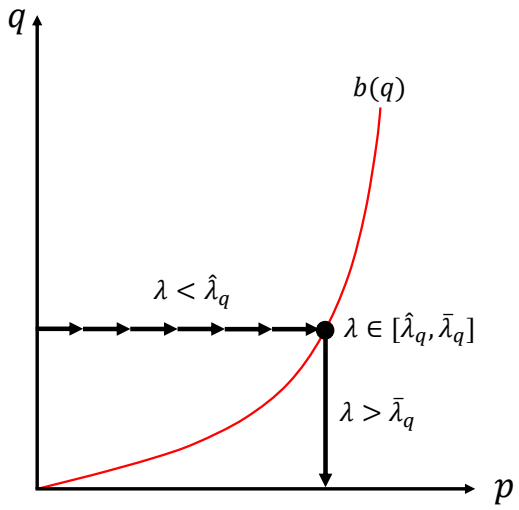


Figure 4: A feasible path of the optimal price

4 Regulation with endogenous quality

Although the assumption concerning the exogenous and constant quality of good x seems quite realistic to analyze some public services, in others it is natural to consider such dimension as endogenous. For example, in public transportation, buses comfort levels and waiting times at stops can be adjusted. To address this issue, in this section, we consider that q is chosen by the regulator.[27]

When operating the market for good x is optimal, the regulator solves

$$\max_{p,e,q,V_\ell,V_h} \mathbb{E}W \equiv \rho(e)W_h(p,e,q) + (1-\rho(e))W_\ell(p,e,q) \tag{10}$$

$$\text{s.t. } V_h(p,e,q) \geq 0, \quad (VP_h)$$

$$V_\ell(p,e,q) \geq 0 \quad (VP_\ell).$$

This problem is formally equivalent to (6), except for the fact that the social welfare and the firm's ex-post utility also depend upon q.[28]

Let's denote by p^E, e^E, q^E, V_h^E and V_ℓ^E the solutions to (10). Again, it is straightforward

[27] In many contexts, the regulator does not select directly the quality of the service. Instead, he sets a minimum standard $\underline{q}$, and the provider complies by choosing $q \geq \underline{q}$. In our setting, this yields the same outcome. As any optimal scheme would satisfy $p \leq b(q)$, and quality does not affect the trade-off between formal and informal consumption, the firm would always pick $q = \underline{q}$.

[28] Now, we make the dependence of W and V on q explicit.

to verify that $V_l^E = V_h^E = 0$ and $p^E \leq b(q^E)$. The following proposition characterizes the optimal regulatory scheme in this new environment.

Proposition 4 *When $p^E < b(q^E)$, the price p^E and the effort e^E are characterized by the system of first-order conditions (8) and (9). The optimal quality q^E is given by*

$$b'(q^E) = (1 + \lambda)c_2. \tag{11}$$

The equation that determines the optimal level of quality is independent of e^E and p^E. This is a consequence of the quasi-linearity of the utility functions, the cost separability between quantity and quality, and the possibility of informal consumption.[29]

As a corollary to these observations, the qualitative properties of the optimal price p^E and e^E are identical to those presented in the previous section, and q^E is always strictly positive. As an increase in c_2 or λ raises its social marginal cost, the individuals' marginal benefit of it must raise as well. Since $b(\cdot)$ is concave, q^E decreases.

However, the price and the quality are not completely disconnected, since the latter limits the price the regulator can charge. The main difference with respect to the previous section is that here the value of this cap is endogenously determined.

When the marginal cost of public funds rises, p^E increases while q^E decreases. Hence, we can expect that there exist parameters such that, for a relatively high value of λ, i.e., $\lambda = \tilde{\lambda}$, these optimal regulatory instruments reach the boundary; i.e., $p^E = b(q^E)$ in an active market. In the Appendix, we show that this is indeed correct. Next, we undertake comparative statics at the boundary. We find that, under plausible conditions, increases in λ imply *lowering* the optimal price, a result which is somewhat unexpected.

As we have already mentioned, in an interior solution, increasing λ has only one direct effect on q^E. At the boundary, however, quality has an additional benefit for the regulator: increasing it allows him to raise the price, while still keeping formal consumers in the market for good x. An increase in λ, which makes higher prices a more attractive instrument to finance the firm, also generates an indirect effect, through a countervailing incentive to push the quality upwards. Therefore, at the boundary, the total effect of an increase in λ on q^E is ambiguous. In the next proposition, we establish a sufficient condition for the direct effect to dominate, and thus for q^E to be decreasing in λ.

[29] Since all individuals consume good x, either as formal buyers or as evaders, the marginal benefit of quality affects identically all of them, regardless of the price.

Proposition 5 *If* $p^E = b(q^E)$ *and*

$$c_2 > b'(q^E)\left[1 + R(e^E)\left(1 + \frac{1+\tilde{\lambda}}{\tilde{\lambda}}\bar{z}f(\bar{z})\right)\right],$$

the price p^E *and the quality* q^E *decrease when* λ *increases.*

The left hand side represents the direct effect of increasing λ on the provision of q, which is given by its marginal cost c_2. On the right hand side we have the indirect effect: $b'(q)$ is the marginal increase in p that becomes feasible after an increase in q. The magnitude of this indirect effect is amplified by $R(e)$, the inverse of the curvature of the function $\rho(\cdot)$. The reason for this is the following: the lower the curvature, the larger the increase in e, as a reaction to a change in λ. This in turn makes price increases even more attractive for the regulator, due to the complementarity between p and e.[30]

If the condition displayed in Proposition 5 holds, further increases in λ cannot be met by price increases at the boundary. To mitigate the resulting fiscal expansion caused by the use of transfers, the regulator brings down the quality (which decreases costs) and the price (otherwise there would be no formal consumers), while keeping the market active. Therefore, there is a non-monotonic relationship between the optimal price and the marginal cost of public funds. The presence of evaders gives rise to the possibility that price and quality do not comove, and therefore public services of different qualities may be equally priced.

Finally, we incorporate into our analysis the option that good x may not be provided. The next corollary naturally follows from the previous proposition.

Corollary 3 *There exists a threshold* $\overline{\overline{\lambda}}$ *such that, when* $\lambda > \overline{\overline{\lambda}}$*, the firm is optimally shut down.*

Despite the opposite interactions between regulatory instruments reported in the previous proposition, the expected social cost of transfers increase with λ. This implies that for sufficiently high values of the marginal cost of public funds, the regulator shuts down the firm, and thus good x is not provided.

Figure 5 illustrates the previous discussion, based again on a possible case characterized in the Appendix. In Panel (a) we depict one feasible path of the pair (p^E, q^E) as a function of the deadweight loss of taxation: decreasing to the right if $\lambda < \widetilde{\lambda}$, then

[30] The term multiplying $R(e)$ is an upper bound of the complementarity between price and effort, as discussed in Section 4.

decreasing to the left when $\lambda \in [\widetilde{\lambda}, \overline{\overline{\lambda}}]$. Finally, for higher values of λ, the regulator shuts down the firm. To visualize it better, Panel (b) shows how the optimal price p^E evolves when λ increases.

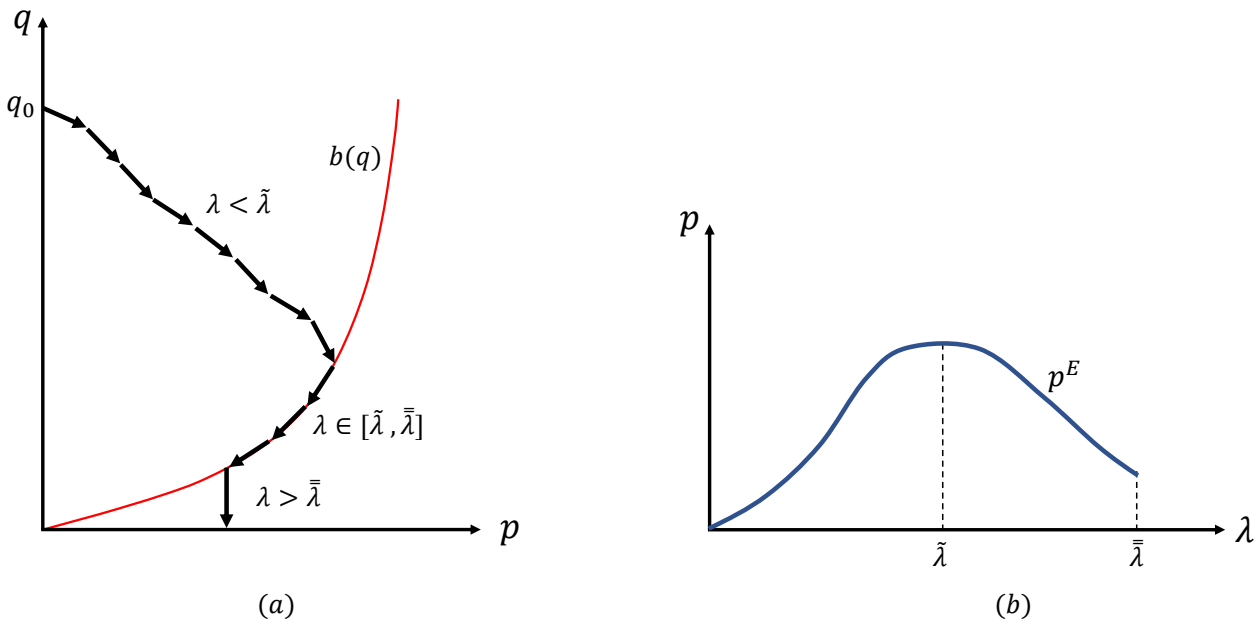


Figure 5: Feasible path of the optimal price-quality pair

To verify the robustness of the analytical results of this section, Figure 6 complements them with numerical simulations. For the baseline parameterization we assume that $\rho(e) = \frac{10e}{1+10e}$ and $b(q) = \sqrt{q}$. We normalize $\bar{z}$ to 1. We also set $\gamma_\ell = 0.5$, $\gamma_h = 0.55$, $\theta = 0.001$, $c_1 = 0.04$, and $c_2 = 1.5$.[31]

The figures on the left depict the boundary $p = b(q)$ as a dotted yellow curve and, in blue, the path of the optimal price-quality pair (p^E, q^E) as a function of the deadweight loss of taxation λ. At the center, we show the probability distribution $f(z)$, and the thresholds $\widehat{z}_h$ and $\widehat{z}_\ell$ at $\tilde{\lambda}$, where the price-quality pair reaches the boundary. On the right, we illustrate how the optimal price p^E varies with λ, and indicate $\tilde{\lambda}$.

First, we observe that the optimal path predicted by the model, and depicted in Figure 5(a), emerges under different probability distributions. The optimal pair (p^E, q^E) reaches the boundary at values of λ that are neither too small nor too large, and are indeed close to some of the estimated values shown in the Introduction. As λ increases, p^E and q^E decrease along the boundary, and finally the firm is shut down.

[31] K is chosen so that the boundary is reached and transfers are positive.

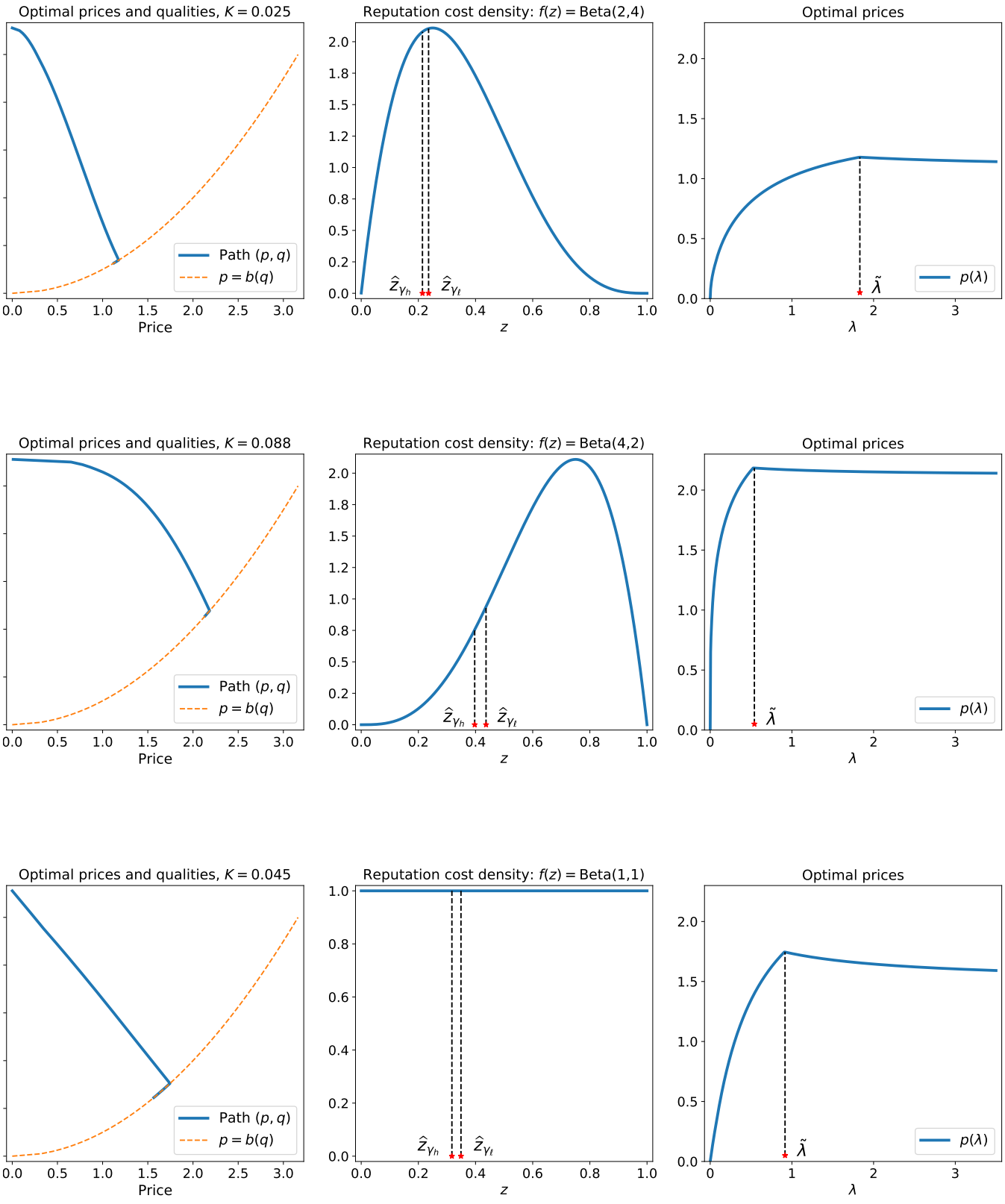


Figure 6: Numerical simulations

Second, these simulations help us visualize the channel through which λ impacts p^E and q^E. For a marginal change $d\lambda > 0$, the intensity of the price adjustment depends on the value $f(\widehat{z})$, which determines the mass of marginal evaders. Moreover, for any price $p^E < b(q)$, the thresholds $\widehat{z}_h(p^E)$ and $\widehat{z}_\ell(p^E)$ are located to the left of the corresponding thresholds at the boundary.

When $f(z) = \text{Beta}(2,4)$, the mass of marginal evaders is big, leading to a slow increase in the optimal price as λ goes up. Moreover, the critical level $\tilde{\lambda}$ is relatively high, since the regulator moves very reluctantly toward higher prices. In comparison, when $f(z) = \text{Beta}(4,2)$, there is a small mass of marginal evaders when $p < b(q)$, leading to sharp price increases as λ raises. This implies that $p^E = b(q^E)$ for a relatively low level $\tilde{\lambda}$.

It is also important to mention that, in both cases, after the boundary $p = b(q)$ has been reached, prices decrease very slowly with λ. Indeed, the 'small' section where the path (p^E, q^E) coincides with the boundary in the first column corresponds to a large set of λs, as can be appreciated in the third column.

Under a uniform distribution, the last simulation is somewhat between the two previous cases when $p^E < b(q^E)$. However, it is interesting to note that the price decrease after reaching the boundary is sharper and bigger in magnitude than under non-uniform distributions.

5 Non-unit demand

Our assumption of unit demand is based on the fact that many public services (like transport) are consumed in almost constant quantities, whose value can be normalized to one without loss of generality. This generates clear-cut results, with interesting intuitions, in particular those related to the pricing rule. However, other public services that also face the threat of evasion have non-unit individual demands, like electricity. In this section, we relax this assumption, and, when quality is exogenous, we derive a pricing rule that can be easily contrasted with the classic Ramsey-Boiteux-Laffont-Tirole formula.

Formal buyers choose the quantity x to purchase, deriving utility $U^F = v(x) + y - px$, where the strictly concave function $v(x)$ represents the individual benefit from consuming x units, and satisfies $v(0) = 0$ and $v'(0) = \infty$.[32] Individual demand $x^*(p)$ is implicitly given by $v'(x^*(p)) = p$.

For the sake of simplicity, we assume that all evaders get $\widehat{x} > 0$, and thus obtain a

[32]Since quality will remain fixed in this section, we drop any explicit mention of q in the benefit function.

benefit $v(\widehat{x})$.[33] As before, their utility is given by $U^I = v(\widehat{x}) + y - \gamma z$.

Let $\widehat{z}_\gamma(p) \equiv \frac{v(\widehat{x})-v(x^*(p))+px^*(p)}{\gamma}$ be the reputation cost of an individual that is indifferent between consuming good x formally and informally. As the benefit function $v(.)$ is strictly concave, $v(x^*) > px^*$, and thus no individual leaves the market.

As both participation constraints bind at the optimum, the regulator solves

$$\max_{p,e} \mathbb{E}W^E = y + \mathbb{E}\Big[F(\widehat{z}_\gamma(p))v(\widehat{x}) + (1 - F(\widehat{z}_\gamma(p)))v(x^*(p))\Big] - \mathbb{E}\int_0^{\widehat{z}_\gamma(p)} \gamma z dF(z)$$
$$+\mathbb{E}\Big[\lambda(1-F(\widehat{z}_\gamma(p)))px^*(p)\Big] - (1+\lambda)\Big(K+\mathbb{E}[F(\widehat{z}_\gamma(p))\widehat{x}+(1-F(\widehat{z}_\gamma(p)))x^*(p)]c_1+\theta e\Big) \tag{12}$$

This problem is conceptually identical to (7), except that formal buyers and evaders do not consume the same amount of good x. Moreover, a change in p now affects both the way individuals consume good x (formal or informal) and the amount purchased in the former case.

Let $\eta^I \equiv \frac{f(\widehat{z}_\gamma(p))\frac{\partial \widehat{z}_\gamma(p)}{\partial p}p}{F(\widehat{z}_\gamma(p))} > 0$ and $\eta^F \equiv -\varepsilon^F - \frac{F(\widehat{z}_\gamma(p))\eta^D_I}{1-F(\widehat{z}_\gamma(p))} < 0$ be the price-elasticity of the aggregate informal and formal consumption, respectively, where ε^F is the price-elasticity of the individual formal demand. The following proposition characterizes the optimal price-effort scheme.

Proposition 6 *The price p^E and the effort e^E are characterized by*

$$\frac{p^E - c_1}{p^E} = \underbrace{-\frac{\lambda}{1+\lambda}\frac{\mathbb{E}[1-F(\widehat{z}_\gamma(p^E))]}{\mathbb{E}[(1-F(\widehat{z}_\gamma(p^E)))\eta^F]}}_{(a)} + \underbrace{\frac{\mathbb{E}[\partial X^I_\gamma/\partial p]}{\mathbb{E}[\partial X^F_\gamma/\partial p]}\frac{c_1}{p^E}}_{(b)} \tag{13}$$

$$\rho'(e^E)\Bigg[\Big(v(x^*(p^E)) - v(\widehat{x}) - (1+\lambda)(x^*(p^E)-\widehat{x})c_1 + \lambda p^E x^*(p^E)\Big)[F(\widehat{z}_{\gamma_l}(p^E)) - F(\widehat{z}_{\gamma_h}(p^E))]$$
$$+\ \Omega(p^E)\Bigg] \leq (1+\lambda)\theta, \textit{with equality if } e^E > 0. \tag{14}$$

[33] For the sake of simplicity, this version of our model does not consider the possibility to evade by paying only a fraction of the due price. For example, this type of misbehavior may emerge in transport systems with fares increasing with the distance to travel. Evaders pay an amount that entitles them to ride up to a given place, but in fact they go further. Although incorporating this type of evasion could be more realistic, we believe that it will also complicate unnecessarily the analysis of the problem without modifying substantially the results.

Expression (13) defines the optimal pricing rule with non-unit demands.[34] On the left-hand side, we have the Lerner index, which evaluates the markup on *sold quantities*. The right-hand side is formed by the sum of two terms. The first one, (a), captures the Ramsey-Boiteux-Laffont-Tirole rationale for financing a natural monopoly with prices and costly transfers, where the marginal cost of public funds and the price-elasticity of demand play a crucial role. The mathematical expression of (a) differs with respect to the conventional formula, as it appears in Baumol and Bradford (1970), in two aspects. First, in the denominator, the price elasticity corresponds to that of the formal demand, which incorporates not only the effect of a price change on the individual demand of a formal buyer, but also on their number. Clearly, the presence of evaders pushes the value of the price elasticity upwards, which *ceteris paribus* implies a lower mark-up. Second, the regulator takes into account the random nature of enforcement, and thus considers the expected value of the formal demand.

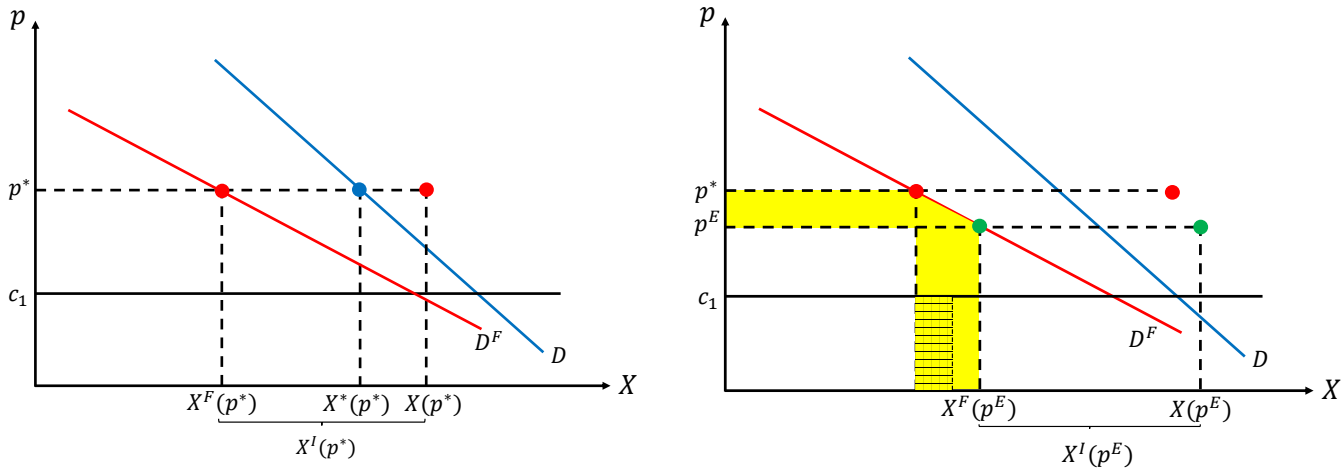
The second term (b) captures the fact that the regulator faces another effect induced by an increase in the price, which is absent in models without informal consumers: the variable cost of serving a higher number of evaders will increase. This second effect is measured with respect to the loss of income generated by the price increase, and thus is negative, reinforcing the above mentioned decrease in the optimal markup. Note that this second term does not depend directly upon the marginal cost of public funds. Thus, the optimal pricing rule separates the Ramsey-Boiteux rationale and the financial concern for the extra deficit caused by evaders.

The term (a) is positive, while (b) is negative. Therefore, contrary to the traditional natural-monopoly regulation, there is no clear-cut prediction concerning the comparison between the price and the marginal cost. For example, if the cost of serving more evaders were relatively high, the second expression may dominate, and thus the price p^E must be optimally set below c_1.[35] By explicitly focusing on the funding of the firm's deficit without fines, our results contrast those found by Buehler et al. (2017). In particular, the expression (13) yields marginal cost pricing provided very particular conditions hold, whereas these authors always obtain such a sharp result.

The following figures illustrate these considerations.

[34] Expression (14) has the same intuition than (9); it just incorporates the fact that formal and informal consumption are not equivalent. Therefore, in the remainder of this section, we do not discuss about the optimal level of effort.

[35] We are not the first to obtain such result. For example, in the context of a natural monopoly financed by a two-part tariff, Ng and Weisser (1974) showed that if the demand of the marginal individual (i.e., the one that is indifferent between consuming or not the good produced by the firm) is above the average demand of inframarginal consumers, the regulator optimally sets the variable price below the marginal cost.



(a) Ramsey-Boiteux pricing with evasion

(b) Optimal pricing with evasion

Figure 7: Optimal pricing with non-unit demands

In Panel (a), we draw the demand for good x when there is no evasion, D (in blue), the marginal cost c_1, and the optimal Ramsey-Boiteux price p^*. Then, we let evasion to emerge; but we keep the price constant. The red curve depicts the demand of formal buyers, D^F, and thus $X^F(p^*)$ is the total quantity purchased. Finally, we also show total informal consumption $X^I(p^*)$, and total consumption $X(p^*)$. In Panel (b), we let the regulator to react optimally to the presence of evaders, and we depict the optimal price and quantities (in green). When deciding whether and how much to modify the price, the regulator has to take into account i) the Ramsey-Boiteux-Laffont-Tirole concern about financing the firm's deficit with price and costly transfers, and ii) the financial consideration of the extra deficit caused by evaders' consumption. As we have already mentioned, $p^E \leq p^*$. So, when it decreases the price optimally, the regulator balances the social values of the net change in the firm's income and the increasing consumer surplus (in yellow) and the decrease in the cost of serving evaders (striped area).

6 Conclusions

Most contributions studying the provision of public services under the threat of evasion focus on deterrence, and do not incorporate the use of prices as a feasible instrument to deal with this pervasive problem. Even among those studies that consider

them, they analyze a profit-maximizing firm instead of a regulated monopoly, without formalizing how financial losses are covered. In this paper, we provide a normative model in which regulation of a natural monopoly is the focus, and in particular, the tension between assuring the firm's participation via socially costly transfers or by other mechanism, such as prices and enforcement. We highlight the main channels through which we believe fare evasion alters the received theory of natural monopoly regulation, stressing the relation between prices, quality, and the marginal cost of public funds.

Our results generate some important policy implications. First, when the service is consumed in fixed quantities, the existence of evasion should prevent the regulator to consider marginal costs as a pertinent variable in its optimization. Second, regulatory authorities should try to improve their understanding of the propensity to evade (i.e., the distribution of subjective costs) in the population. Finally, the technological characteristics of a transport system and the design of the pricing/transfer schemes cannot be decoupled. Indeed, the quality provided to consumers limits the price the firm can charge.

We conclude by emphasizing some limitations of our analysis and present some possible extensions. We have assumed that individuals have quasi-linear utilities. This implies that income effects play no role in terms of the decision between consuming the good formally or as an evader. Although this assumption seems plausible in general, one can argue that low-income individuals, for which the expenditure in public services represents a non-negligible fraction of their income, can face income effects when prices increase. Also, we have ignored that firms may have different efficiency levels to detect evaders. Incorporating this adverse selection feature into the model obviously modifies our results. But, more importantly, it could enable us to go one step ahead, and study the regulator's decision of granting the monopoly when it faces a pool of diverse potential entrants in a context of pervasive fare evasion. This analysis can help regulators to improve the design of bidding mechanisms for the provision of public transport when such misbehavior is an institutional concern.

References

Armstrong, M. and D. Sappington (2007). Recent developments in the theory of regulation. In M. Armstrong and R. Porter (Eds.), *Handbook of Industrial Organization Volume 3*, pp. 1557–1700. North Holland, Elsevier, London.

Ayal, S., J. Celse, and G. Hochman (2021). Crafting messages to fight dishonesty: A field investigation of the effects of social norms and watching eye cues on fare evasion. *Organizational Behavior and Human Decision Processes 166*, 9–19.

Barabino, B., C. Lai, and A. Olivo (2020). Fare evasion in public transport systems: a review of the literature. *Public Transport 12*, 27–88.

Barabino, B. and S. Salis (2019). Moving towards a more accurate level of inspection against fare evasion in proof-of-payment transit systems. *Networks and Spatial Economics 19*, 1319–1346.

Barrios, S., J. Pycroft, and B. Saveyn (2013). The marginal cost of public funds in the EU: The case of labour versus green taxes. In M. Marino, M. Tasso, and P. Tommasino (Eds.), *Fiscal Policy and Growth*, pp. 403–431. Banca d'Italia, Roma.

Basso, L. and H. Silva (2014). Efficiency and Substitutability of Transit Subsidies and Other Urban Transport Policies. *American Economic Journal: Economic Policy 6*(4), 1–33.

Baumol, W. and D. Bradford (1970). Optimal departures from marginal cost pricing. *American Economic Review 60*(3), 265–283.

Bijleveld, C. (2007). Fare dodging and the strong arm of the law An experimental evaluation of two different penalty schemes for fare evasion. *Journal of Experimental Criminology 3*(2), 183–199.

Boiteux, M. (1956). Sur la gestion des monopoles publics astreints à l'équilibre budgétaire. *Econometrica 24*(1), 22–40.

Bonfanti, G. and T. Wagenknecht (2010). Human factors reduce aggression and fare evasion. *Public Transportation International 59*(1), 28–32.

Boyd, C., C. Martini, J. Rickard, and A. Russell (1989). Fare evasion and non-compliance: A simple model. *Journal of Transport Economics and Policy 23*(2), 189–197.

Bucciol, A., F. Landini, and M. Piovesan (2013). Unethical behavior in the field: Demographic characteristics and beliefs of the cheater. *Journal of Economic Behavior & Organization 93*, 248–257.

Buehler, S., D. Halbheer, and M. Lechner (2017). Payment evasion. *The Journal of Industrial Economics 65*(4), 804–832.

Calvo, M. (2015). Análisis desagregado del comportamiento de usuarios de transporte público utilizando datos masivos. Memoria, Departamento de Ingeniería Civil, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile.

Currie, G. and A. Delbosc (2013). Understanding the psychology of fare evasion: Final report. Melbourne: Public Transport, Victoria.

Currie, G. and J. Reynolds (2016). Evaluating Pay-on-Entry Versus Proof-of-Payment Ticketing in Light Rail Transit. *Transportation Research Record* (2540), 39–45.

Dahlby, B. (2008). *The Marginal Cost of Public Funds: Theory and Applications*. The MIT Press, Cambridge, Massachusetts.

Dai, Z., F. Galeotti, and M. Villeval (2019). Fare-dodging in the lab and the moral cost of dishonesty. In A. Bucciol and N. Montinari (Eds.), *Dishonesty in Behavioral Economics*, pp. 245–265. Academic Press, Elsevier, London.

EMTA (2020). Barometer. European Metropolitan Transport Authorities.

Estache, A. and L. Wren-Lewis (2009). Toward a Theory of Regulation for Developing Countries: Following Jean-Jacques Laffont's Lead. *Journal of Economic Literature 47*(3), 729–770.

Fraser, C. (1996). Exclusion and moral hazard: A further analysis. *Journal of Public Economics 60*(2), 295–301.

Guarda, P., P. Galilea, L. Paget-Seekins, and J. de D. Ortúzar (2016). What is behind fare evasion in urban bus systems? An econometric approach. *Transportation Research Part A: Policy and Practice 84*, 55–71.

Guzman, L., J. Arellana, and J. P. Camargo (2021). A hybrid discrete choice model to understand the effect of public policy on fare evasion discouragement in Bogotá's Bus Rapid Transit. *Transportation Research Part A: Policy and Practice 151*, 140–153.

Killias, M., D. Scheidegger, and P. Nordenson (2009). The effects of increasing the certainty of punishment: A field experiment on public transportation. *European Journal of Criminology 6*(5), 387–400.

Kooreman, P. (1993). Fare evasion as a result of expected utility maximisation: some empirical support. *Journal of Transport Economics and Policy 27*(1), 69–74.

Laffont, J.-J. (1994). The New Economics of Regulation Ten Years After. *Econometrica 62*(3), 507–537.

Laffont, J.-J. (2005). *Regulation and Development*. Cambridge University Press, Cambridge.

Laffont, J.-J. and J. Tirole (1986). Using Cost Observation to Regulate Firms. *Journal of Political Economy 94*(3), 614–641.

Laffont, J.-J. and J. Tirole (1990). The Regulation of Multiproduct Firms Part I: Theory. *Journal of Public Economics 43*(1), 1–36.

Laffont, J.-J. and J. Tirole (1993). *A Theory of Incentives in Procurement and Regulation*. The MIT Press, Cambridge, Massachusetts.

Ng, Y.-K. and M. Weisser (1974). Optimal pricing with a budget constraint: the case of the two-part tariff. *The Review of Economic Studies 41*(3), 337–345.

Porath, K. and P. Galilea (2020). Temporal analysis of fare evasion in Transantiago: A socio-political view. *Research in Transportation Economics 83*.

Sherman, R. (1989). *The regulation of monopoly*. Cambridge University Press, Cambridge.

Silva, E. and C. Kahn (1993). Exclusion and moral hazard The case of identical demand. *Journal of Public Economics 52*(2), 217–235.

Smith, T. (2004). Electricity theft: a comparative analysis. *Energy Policy 32*(18), 2067–2076.

Sterner, A. and S. Sheng (2013). The effect of social stigma on fare evasion in Stockholm's public transport. *Journal of Transport Literature 7*(4), 50–74.

Troncoso, R. and L. de Grange (2017). Fare evasion in public transport: A time series approach. *Transportation Research Part A: Policy and Practice 100*, 311–318.

Appendix A

Proof of Lemma 1

i) Assume (wlog) that VP_h does not bind at the optimum. If so, we can always decrease V_h^E by a small $\epsilon > 0$ such that VP_h still holds. This change leads to an increase in expected welfare $\mathbb{E}W$, which is a contradiction. A similar argument can be used to prove that VP_ℓ also binds at the optimum.
ii) Assume that running the market for good x is efficient, i.e., $b(q) > C(1,q) = K + c_1 + c_2 q$. Let $\widetilde{z}_\gamma(b(q)) \equiv \frac{b(q)}{\gamma}$ be the value of the reputation cost that makes an individual indifferent between consuming good x informally or leaving the market. Moreover, assume that the regulator optimally sets $p^E > b(q)$ and e^E. This implies that $\widetilde{z}_\gamma(b(q)) < \widehat{z}_\gamma(p^E)$. Hence, individuals characterized by $z \leq \widetilde{z}_\gamma(b(q))$ decide to consume informally, whereas those with a higher reputation cost leave the market for good x. Under these circumstances, the expected welfare associated to the regulatory scheme (p^E, e^E) is given by

$$y + \mathbb{E}\left[\int_0^{\widetilde{z}_\gamma(b(q))} (b(q) - \gamma z)\, dF(z)\right] - (1+\lambda)\left(K + \mathbb{E}\left[c_1 \int_0^{\widetilde{z}_\gamma(b(q))} dF(z)\right] + c_2 q + \theta e^E\right). \tag{15}$$

Suppose now that the regulator offers a new scheme (p', e'), with $p' = b(q)$ and $e' = e^E$. This new scheme leads to an expected welfare given by

$$\begin{aligned} &\underbrace{y + \mathbb{E}\left[\int_0^{\widetilde{z}_\gamma(b(q))} b(q) dF(z)\right]}_{\text{(a)}} - \underbrace{\mathbb{E}\left[\int_0^{\widetilde{z}_\gamma(b(q))} \gamma z dF(z)\right]}_{\text{(b)}} + (1+\lambda)\underbrace{\mathbb{E}\left[\int_{\widetilde{z}_\gamma(b(q))}^{\overline{z}} b(q) dF(z)\right]}_{\text{(c)}} \\ &\qquad -(1+\lambda)\underbrace{\left(C(1,q) + \theta e^E\right)}_{\text{(d)}}, \end{aligned} \tag{16}$$

where we use $\widehat{z}_\gamma(p') = \widetilde{z}_\gamma(b(q))$ and the fact that individuals with high reputation costs consume formally when $p' = b(q)$.

The difference in welfare terms between both regulatory schemes is given by

$$(16) - (15) = (1+\lambda)(b(q) - c_1)\mathbb{E}\left[\int_{\widetilde{z}_\gamma(b(q))}^{\overline{z}} dF(z)\right].$$

As $b(q) > c_1$, this difference is strictly positive, implying that the proposed deviation is welfare improving. This contradicts the optimal nature of the initial regulatory scheme. If good x has to be provided, it is always optimal to set $p \leq b(q)$, and thus to induce all individuals to choose between consuming good x formally or informally. ∎

Proof of Lemma 2

The reputational impact of increasing effort marginally on the expected social welfare is equal to $\rho'(e)\Omega(p)$. To obtain the sign of $\Omega(p)$, we define the function

$$\Phi(\gamma) \equiv \gamma \int_0^{\frac{p}{\gamma}} z dF(z),$$

and we compute

$$\frac{d\Phi(\gamma)}{d\gamma} = \int_0^{\frac{p}{\gamma}} z dF(z) - \left(\frac{p}{\gamma}\right)^2 f\left(\frac{p}{\gamma}\right).$$

For the sake of simplicity, let $t = \frac{p}{\gamma}$. As f is bounded, $\lim_{t\to 0} \frac{d\Phi(\gamma)}{d\gamma} = 0$ and

$$\frac{d}{dt}\left(\frac{d\Phi(\gamma)}{d\gamma}\right) = -tf(t)\left[1 + t\frac{f'(t)}{f(t)}\right]. \tag{17}$$

Since f is log-concave, $\frac{f'(t)}{f(t)}$ decreases with t, reaching the minimal value $f'(\overline{z})/f(\overline{z})$ when $t = \overline{z}$.
If $f'(\overline{z})/f(\overline{z}) \geq 0$, $f'(z)/f(z) \geq 0$ for all z. Hence, (17) is negative. Otherwise, if $f'(\overline{z})/f(\overline{z}) < 0$, we need to consider the following two sub-cases:

- $\frac{f'(\overline{z})}{f(\overline{z})} \leq \frac{f'(t)}{f(t)} < 0$, which yields $\overline{z}\frac{f'(\overline{z})}{f(\overline{z})} \leq t\frac{f'(t)}{f(t)}$, and thus by Assumption 1, (17) is negative.
- $\frac{f'(\overline{z})}{f(\overline{z})} < 0 < \frac{f'(t)}{f(t)}$, which also implies that (17) is negative.

As (17) is negative, $\frac{d\Phi(\gamma)}{d\gamma} < 0$ (except when $t = 0$), and thus $\Phi(\gamma_\ell) > \Phi(\gamma_h)$. Therefore, $\Omega(p) > 0$, and so $\rho'(e)\Omega(p) > 0$.

In order to prove the second inequality of the lemma, we show first that $\Omega(p)$ increases with p.

Let $\Psi(\gamma_h) \equiv \frac{d\Omega(p)}{dp} = \frac{p}{\gamma_\ell} f(\frac{p}{\gamma_\ell}) - \frac{p}{\gamma_h} f(\frac{p}{\gamma_h})$. Clearly, $lim_{\gamma_h \to \gamma_\ell} \Psi(\gamma_h) = 0$ and

$$\frac{d\Psi(\gamma_h)}{d\gamma_h} = \frac{\widehat{z}_{\gamma_h}}{\gamma_h} f(\widehat{z}_{\gamma_h}) \left[1 + \widehat{z}_{\gamma_h} \frac{f'(\widehat{z}_{\gamma_h})}{f(\widehat{z}_{\gamma_h})} \right] > 0.$$

So $\Psi(\gamma_h)$ is always positive, implying that $\Omega(p)$ increases with p. The proof concludes using Assumption 2 and the fact that $\rho(e)$ is strictly concave. ∎

Proof of Proposition 1

The result follows from direct differentiation of the objective function (7) with respect to p and e. ∎

Proof of Proposition 2

(i). First, we study the comparative statics of p^E and e^E with respect to λ. Using a standard approach, we only need to prove that $\frac{\partial^2 \mathbb{E}W}{\partial p \partial \lambda}$, $\frac{\partial^2 \mathbb{E}W}{\partial e \partial \lambda}$, and $\frac{\partial^2 \mathbb{E}W}{\partial p \partial e}$ are all strictly positive. We can show that

$$\begin{aligned}
\frac{\partial \mathbb{E}W}{\partial p} &= -\frac{\partial}{\partial p}\left(\mathbb{E} \int_0^{\widehat{z}_\gamma(p)} \gamma z dF(z) \right) + \lambda \frac{\partial^2 \mathbb{E}W}{\partial p \partial \lambda} \\
\frac{\partial \mathbb{E}W}{\partial e} &= \left[-\frac{\partial}{\partial e}\left(\mathbb{E} \int_0^{\widehat{z}_\gamma(p)} \gamma z dF(z) \right) - \theta \right] + \lambda \frac{\partial^2 \mathbb{E}W}{\partial e \partial \lambda} \\
\frac{\partial^2 \mathbb{E}W}{\partial e \partial p} &= \rho'(e) \frac{\partial}{\partial p}\Big(W_h(p,e) - W_\ell(p,e) \Big).
\end{aligned}$$

Let's consider an interior solution, where $\frac{\partial \mathbb{E}W}{\partial p} = 0$. Therefore, as $\frac{\partial}{\partial p}\left(\mathbb{E} \int_0^{\widehat{z}_\gamma(p)} \gamma z dF(z) \right) > 0$, $\frac{\partial^2 \mathbb{E}W}{\partial p \partial \lambda} > 0$. Moreover, Lemma 2 guarantees that the bracketed term in the second equation is negative. Using an analogous argument allows us to conclude that $\frac{\partial^2 EW}{\partial e \partial \lambda} > 0$. Finally, to prove that the third derivative is positive, we compute

$$\begin{aligned}
\frac{\partial W(p,\gamma)}{\partial \gamma \partial p} &= \frac{\partial}{\partial \gamma}\left[\lambda \int_{\widehat{z}_\gamma(p)}^{\overline{z}} dF(z) - \frac{\partial \widehat{z}_\gamma(p)}{\partial p}(1+\lambda) p f(\widehat{z}_\gamma(p)) \right] \\
&= -\frac{\partial \widehat{z}_\gamma(p)}{\partial \gamma}\left[(1+2\lambda) f(\widehat{z}_\gamma(p)) + (1+\lambda)\widehat{z}_\gamma(p) f'(\widehat{z}_\gamma(p)) \right].
\end{aligned} \tag{18}$$

In the proof of Lemma 2 we have shown that, for all z, $z\frac{f'(z)}{f(z)} > -1$. Therefore,

$$(1+\lambda)\widehat{z}_\gamma(p) f'(\widehat{z}_\gamma(p)) > -(1+\lambda) f(\widehat{z}_\gamma(p)),$$

and thus (18) is strictly positive. Using the local supermodularity derived above and applying the Implicit Function Theorem, we obtain that p^E and e^E increase with λ.

Consider now the comparative statics with respect to θ. It is immediate that $\frac{\partial \mathbb{E}W}{\partial e \partial \theta} = -(1+\lambda)$ and $\frac{\partial \mathbb{E}W}{\partial p \partial \theta} = 0$. Given that $\frac{\partial^2 \mathbb{E}W}{\partial e \partial p} > 0$ around the optimal policy, we can conclude that $\mathbb{E}W$ is supermodular in $(-\theta, p, e)$ at the solution. It follows that p^E and e^E weakly decrease as functions of θ.

Finally, we analyze the comparative statics with respect to γ_h. We already established that $\frac{\partial^2 \mathbb{E}W}{\partial e \partial p} > 0$ holds. Moreover, the same assumptions and lemmas allow us to prove that $\frac{\partial^2 \mathbb{E}W}{\partial p \partial \gamma_h} = \rho(e)\frac{\partial^2 W(p,\gamma_h)}{\partial p \partial \gamma_h} > 0$ at an interior solution. Noting that $\frac{\partial^2 \mathbb{E}W}{\partial e \partial \gamma_h} = \rho'(e)\frac{\partial W(p,\gamma)}{\partial \gamma} > 0$, we conclude that p^E and e^E increase with γ_h.

(ii). The response of the expected optimal transfers

$$\mathbb{E}\left[T_\gamma^E\right] = C(1,q) + \theta e - \mathbb{E}[pX_\gamma^F] \tag{19}$$

to changes in θ, γ_h, and λ, are the following

$$\begin{aligned}
\frac{d\mathbb{E}\left[T_\gamma^E\right]}{d\theta} &= \frac{\partial \mathbb{E}\left[T_\gamma^E\right]}{\partial e}\frac{de^E}{d\theta} + \frac{\partial \mathbb{E}\left[T_\gamma^E\right]}{\partial p}\frac{dp^E}{d\theta} \\
\frac{d\mathbb{E}\left[T_\gamma^E\right]}{d\gamma_h} &= \frac{\partial \mathbb{E}\left[T_\gamma^E\right]}{\partial e}\frac{de^E}{d\gamma_h} + \frac{\partial \mathbb{E}\left[T_\gamma^E\right]}{\partial p}\frac{dp^E}{d\gamma_h} \\
\frac{d\mathbb{E}\left[T_\gamma^E\right]}{d\lambda} &= \frac{\partial \mathbb{E}\left[T_\gamma^E\right]}{\partial e}\frac{de^E}{d\lambda} + \frac{\partial \mathbb{E}\left[T_\gamma^E\right]}{\partial p}\frac{dp^E}{d\lambda}
\end{aligned} \tag{20}$$

Differentiating (19) with respect to p and e, and using the first-order conditions (8) and (9) yields

$$\begin{aligned}
\frac{\partial \mathbb{E}\left[T_\gamma^E\right]}{\partial p} &= -\mathbb{E}\left[X_\gamma^F + p^E\frac{\partial X_\gamma^F}{\partial p}\right] < 0 \\
\frac{\partial \mathbb{E}\left[T_\gamma^E\right]}{\partial e} &= \theta - \rho'(e^E)p^E\left[X_{\gamma_h}^F - X_{\gamma_\ell}^F\right] < 0.
\end{aligned} \tag{21}$$

As we know that $\frac{dp^E}{d\theta}, \frac{de^E}{d\theta} < 0$ and $\frac{dp^E}{d\gamma_h}, \frac{de^E}{d\gamma_h}, \frac{dp^E}{d\lambda}, \frac{de^E}{d\lambda} > 0$, the results mentioned in the proposition follow from combining all these signs. ∎

Proof of Corollary 1

Assumption 2 and a continuity argument implies that $e^E = 0$ when $\lambda \approx 0$. Moreover, the existence of λ_1 is direct from the previous proposition, since e^E is increasing in λ. Note that it is possible to obtain $\lambda_1 = \infty$, for example if θ is very high. ∎

Proof of Proposition 3

As

$$\frac{\partial^2 \mathbb{E}W}{\partial p^2} = -(1+\lambda)\mathbb{E}\left[\frac{f(\widehat{z}_\gamma(p))}{\gamma}\left(1+\widehat{z}_\gamma(p)\frac{f'(\widehat{z}_\gamma(p))}{f(\widehat{z}_\gamma(p))}\right)\right] - \lambda\mathbb{E}\left[\frac{f(\widehat{z}_\gamma(p))}{\gamma}\right] < 0,$$

the expected welfare is concave in the price. Hence, once the boundary $p = b(q)$ is reached, the best thing the regulator can do is to just keep the price constant, as λ increases. After this, the problem depends only upon the effort. The proof of the proposition is omitted because it is a direct application of previous results. ∎

Existence of parameters supporting optimal regulatory schemes at the boundary $p = b(q)$, when quality is exogenous

Proposition 7 *Assume that z is uniformly distributed and that $2\gamma_\ell > \gamma_h$. Under some parameter conditions, there exist $\widehat{\lambda}_q < \infty$ such that, when $\lambda \geq \widehat{\lambda}_q$, the optimal price is $p^E = b(q)$ and the optimal level of effort is given by the first-order condition (9).*

Proof. We follow a constructive approach to prove this proposition. First, we find sufficient conditions that ensure the existence of an interval for the quality q under which the results hold. Then, we find the lowest value of λ such that the optimal price is on the frontier. Finally, we identify parameters such that the market for good x is active and transfers are positive under these circumstances.

Let p^E_{lim} and e^E_{lim} be the limit values of p^E and e^E given by the first-order conditions (8) and (9) when $\lambda \to \infty$. From (8) and using the fact that z is distributed uniformly, we obtain

$$p^E_{lim} = \frac{\overline{z}}{2\left[\rho(e^E_{lim})\frac{1}{\gamma_h} + (1-\rho(e^E_{lim}))\frac{1}{\gamma_\ell}\right]}, \tag{22}$$

.

Thus, for any $\delta < 1$,

$$p^E_{lim} > \underline{p} \equiv \delta\frac{\overline{z}\gamma_\ell}{2}. \tag{23}$$

Let $\tilde{q}$ be implicitly defined by $b(\tilde{q}) = \underline{p}$. From now on, we only consider quality levels $q \leq \tilde{q}$.

Now, we find $\widehat{\lambda}_q$, the smallest value of λ such that $p^E = b(q)$. Since p^E is increasing in λ, the first-order condition (8) implies that this happens when

$$b(q)(1+2\widehat{\lambda}_q)\mathbb{E}\left[\frac{1}{\gamma}\right] = \widehat{\lambda}_q\overline{z}. \tag{24}$$

Rearranging (24) allows us to obtain

$$\frac{1+2\widehat{\lambda}_q}{1+\widehat{\lambda}_q} = \frac{\overline{z}}{\overline{z} - \mathbb{E}\left[\widehat{z}_\gamma(b(q))\right]}. \tag{25}$$

If operating the market for good x is optimal when $\lambda = \widehat{\lambda}_q$, the expected welfare under the optimal regulatory scheme is higher than under shutdown,

$$b(q) - \mathbb{E}\left[\int_0^{\widehat{z}_\gamma(b(q))} \gamma z dF(z)\right] + \widehat{\lambda}_q b(q) \int_{\widehat{z}_\gamma(b(q))}^{\overline{z}} dF(z) - (1+\widehat{\lambda}_q)(K + c_1 + c_2 q + \theta e^E) > 0,$$

which can be rewritten as,

$$\Sigma(q) \equiv b(q)\left(1 - \frac{\mathbb{E}\left[\widehat{z}_\gamma(b(q))\right]}{\overline{z}}\left(\frac{\frac{1}{2}+\widehat{\lambda}_q}{1+\widehat{\lambda}_q}\right)\right) > \left(K + c_1 + c_2 q + \theta e^E\right) = C(q). \tag{26}$$

Moreover, if under the above mentioned optimal regulatory scheme, transfers are positive,

$$C(q) > b(q) \int_{\widehat{z}_\gamma(b(q))}^{\overline{z}} dF(z) = b(q)\left(1 - \frac{\widehat{z}_\gamma(b(q))}{\overline{z}}\right) \equiv g(q), \quad \gamma \in \{\gamma_h, \gamma_\ell\}. \tag{27}$$

Combining (26) and (27), the results of the proposition would follow if $\Sigma(q) > C(q) > g(q)$.
A necessary condition is that $\Sigma(q) > g(q)$. The following lemma shows that this condition can be satisfied for a wide range of parameters.

Lemma 3 *If $2\gamma_\ell > \gamma_h$, $\Sigma(q) > g(q)$ for any quality $q < \tilde{q}$.*

Proof. Using (25), condition $\Sigma(q) > g(q)$ can be written, for the most restrictive case, as follows

$$1 - \frac{\mathbb{E}\left[\widehat{z}_\gamma(b(q))\right]}{2\overline{z}}\left(\frac{\overline{z}}{\overline{z} - \mathbb{E}\left[\widehat{z}_\gamma(b(q))\right]}\right) > 1 - \frac{\widehat{z}_{\gamma_h}(b(q))}{\overline{z}}. \tag{28}$$

Rearranging, we can rewrite (28) as

$$\mathbb{E}\left[\frac{1}{\gamma}\right]\left(1 + \frac{2\widehat{z}_{\gamma_\ell}(b(q))}{\overline{z}}\right) < \frac{2}{\gamma_h}, \tag{29}$$

which is implied by

$$\frac{\gamma_h}{\gamma_\ell}\left(1 + \frac{2}{\overline{z}}\frac{b(q)}{\gamma_\ell}\right) < 2. \tag{30}$$

If $q < \tilde{q}$, the left-hand side of (30) satisfies

$$\frac{\gamma_h}{\gamma_\ell}\left(1 + \frac{2}{\overline{z}}\frac{b(q)}{\gamma_\ell}\right) < \frac{\gamma_h}{\gamma_\ell}(1+\delta). \tag{31}$$

If $2\gamma_\ell > \gamma_h$, then inequality (30) holds for some $\delta > 0$. ∎

From (9) evaluated at $\widehat{\lambda}_q$, we know that, if $q = 0$, $e^E = 0$. The following lemma shows how e^E evolves with q.

Lemma 4 *When $\lambda = \widehat{\lambda}_q$, if $e^E > 0$ then it increases with q.*

Proof. We differentiate totally the system of first-order conditions (8) and (9), evaluated at $\lambda = \widehat{\lambda}_q$, when $e^E > 0$. We apply the Implicit Function Theorem and obtain

$$\frac{de^E}{dq} = \frac{-2b'(q)\theta\overline{z}\gamma_h\gamma_\ell}{(b(q))^2(1+2\widehat{\lambda}_q)\Delta(\gamma_h-\gamma_\ell)}\left(\mathbb{E}\left[\frac{1}{\gamma}\right] + \frac{2(1+\widehat{\lambda}_q)\overline{z}}{b(q)}\right) > 0, \tag{32}$$

where

$$\Delta = \frac{\overline{z}}{1+2\widehat{\lambda}_q}\left[\frac{-2\rho'(e^E)\theta}{b(q)(1+2\widehat{\lambda}_q)} + \rho''(e^E)\right] < 0$$

is the determinant of the Hessian matrix, evaluated at the optimum. ∎

As e^E increases with q, let's denote by $\tilde{e}$ the value that e^E adopts when $q = \tilde{q}$ and $p^E = b(\tilde{q})$.[36] Consider the following geometric argument. Since $\Sigma(0) = g(0) = 0$, Inada conditions for $b(q)$ imply that both curves increase sharply, in a neighborhood of $q = 0$. Now, let's define $\overline{C}(q) = K + c_1 + c_2 q + \theta\tilde{e}$ and $\underline{C}(q) = K + c_1 + c_2 q$. We can always find parameters K, c_1, c_2 and θ such that there exists an interval $[q_1, q_2]$ where $\Sigma(q) > \overline{C}(q)$ and $\underline{C}(q) > g(q)$. Hence, as

$$\underline{C}(q) \leq C(q) \leq \overline{C}(q),$$

we have shown the existence of parameters such that, when $q < \tilde{q}$, $\Sigma(q) > C(q) > g(q)$ holds. Figure 6 provides an illustration of this reasoning, for an arbitrary $C(q)$.

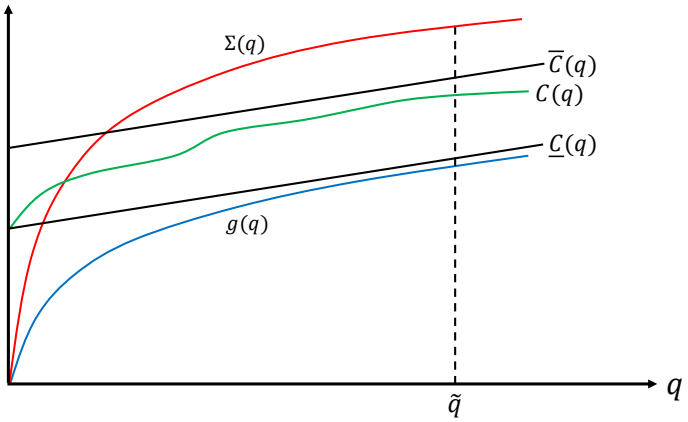


Figure 8: Possibility that $g(q) < C(q) < \Sigma(q)$

[36] We impose sufficient conditions to ensure that $\lambda_1 < \widehat{\lambda}_q$. For example, we can assume that the marginal cost of effort θ is sufficiently low.

Proof of Corollary 2

We present a proof that applies when the conditions that enable the optimal price to reach the boundary $b(q)$ with an active firm are satisfied.

Recall that, as the expected welfare $\mathbb{E}W$ is concave in the price, setting $p^E < b(q)$ is never optimal when $\lambda > \widehat{\lambda}_q$. Under these circumstances, again $p^E = b(q)$ and e^E is implicitly given by the first-order condition (9). Let $\Delta\mathbb{E}W \equiv \mathbb{E}W - y$ be the difference between the expected welfare when the market for good x is active and optimally regulated and when the regulator initially shuts down the firm. Applying an envelope argument, the expected welfare $\mathbb{E}W$ decreases with λ and diverges to $-\infty$. Hence, Bolzano's Theorem ensures that there exists a threshold $\overline{\lambda}_q > \widehat{\lambda}_q$ such that, when $\widehat{\lambda}_q < \lambda \leq \overline{\lambda}_q$, the market for good x should be active. Otherwise, the firm's shutdown is optimal.

A similar argument can be used to show that, if the conditions that enable the optimal price to reach the cap $b(q)$ are not satisfied, then there exists a threshold $\overline{\lambda}_q$ such that, when $\lambda > \overline{\lambda}_q$, the regulator optimally shuts down the firm. ∎

Proof of Proposition 4

Notice from (7) and (10) that q and (p, e) enter the objective function in an additive separable manner. As a consequence, the optimal price and effort are given by (8) and (9). The optimal quality follows from direct differentiation of the objective function with respect to q. The shape of the benefit function $b(\cdot)$ leads to the qualitative properties of q^E. ∎

Existence of parameters supporting non-interior optimal regulatory schemes when quality is endogenous

Proposition 8 *Assume that z is uniformly distributed and that $2\gamma_\ell > \gamma_h$. Under some parameter conditions, there exist $\widetilde{\lambda} < \infty$ such that, when $\lambda \geq \widetilde{\lambda}$, the optimal price-quality schedule satisfies $p^E = b(q^E)$ and the optimal level of effort is given by the first-order condition (9).*

Proof. Consider the relaxed problem

$$\max_{p,e,q,V_\ell,V_h} \mathbb{E}W. \tag{33}$$

Let $p(\lambda)$, $e(\lambda)$, $q(\lambda)$, $V_\ell(\lambda)$ and $V_h(\lambda)$ be the solutions to (33), as functions of λ. In particular, with a slight abuse of notation, let's denote by $q_0 \equiv \lim_{\lambda\to 0} q(\lambda)$ the optimal quality that is implicitly given by $b'(q_0) = c_2$. Also, $\lim_{\lambda\to\infty} q(\lambda) = 0$.

Now let $\Pi(\lambda)$ be the locus of points $(p(\lambda), q(\lambda))$, in the (p, q) plane. From

Propositions 2, 4 and the fact that q^E decreases with λ, we know that, for a given 4-tuple of parameters $\Gamma = (K, c_1, c_2, \theta)$, $\Pi(\lambda)$ is graphically characterized by a negative relation between p and q. Figure 9 depicts $\Pi(\lambda)$ and the boundary $b(q)$.

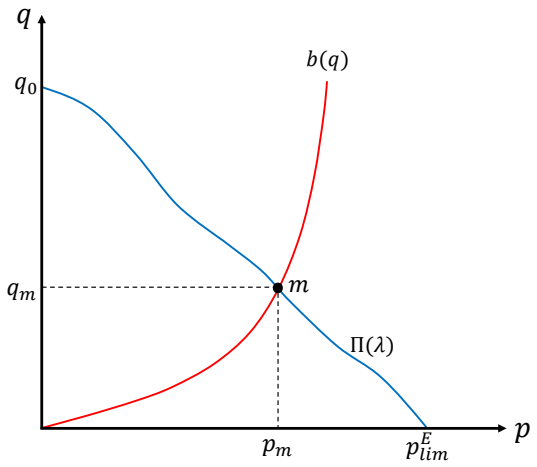


Figure 9: The locus $\Pi(\lambda)$ and the boundary $b(q)$

Due to the properties of both objects, we know that $b(q)$ crosses $\Pi(\lambda)$ only once. Let m be this geometric intersection, and let's denote by p_m and q_m its abscissa and its ordinate, respectively. From point m we infer the existence of a value λ_m, implicitly defined by $p(\lambda_m) = b(q(\lambda_m))$.

In the proof of Proposition 7, we have defined $\underline{p}$ as a lower bound for p^E_{lim}, and implicitly characterized $\tilde{q}$ as satisfying $b(\tilde{q}) = \underline{p}$. So, comparing $\underline{p}$ and p_m, two scenarii can potentially emerge: either $\underline{p} \geq p_m$ or $\underline{p} < p_m$.

1. If $\underline{p} \geq p_m$, $\tilde{q} \geq q_m$. By fixing the values of c_2 and θ in the 4-tuple Γ, $\Pi(\lambda)$ cannot be modified when the other two parameters K and c_1 change.[37] Applying the reasoning used in the proof of Proposition 7 to the value $q = q_m$, we know that we can always find values of K and c_1 such that there exists a value of λ, denoted by $\widetilde{\lambda}$, for which i) the market is active, ii) p_m and q_m solve problem (10), and iii) optimal transfers are positive.

2. If $\underline{p} < p_m$, then $\tilde{q} < q_m$. We can always increase c_2 in the 4-tuple Γ such that $\Pi(\lambda)$ shifts downwards enough, so that the previous scenario emerges again. ∎

[37] This is due to the fact that neither p^E nor e^E depend upon these two parameters.

Proof of Proposition 5

The optimal scheme solves

$$\max_{q,p,e} \mathbb{E}W \equiv y + b(q) + \mathbb{E}\left[\int_0^{\widehat{z}_\gamma(p)} -\gamma z dF(z) + \lambda p \int_{\widehat{z}_\gamma(p)}^{\overline{z}} dF(z)\right] - (1+\lambda)\left(K + c_1 + c_2 q + \theta e\right)$$

subject to

$$p \leq b(q).$$

The Lagrangian of this problem is

$$\mathcal{L} = y + b(q) + \mathbb{E}\left[\int_0^{\widehat{z}_\gamma(p)} -\gamma z dF(z) + \lambda p \int_{\widehat{z}_\gamma(p)}^{\overline{z}} dF(z)\right] - (1+\lambda)\left(K + c_1 + c_2 q + \theta e\right) - \Phi[p - b(q)]. \tag{34}$$

We only analyze the case when the constraint is binding, i.e. $\Phi > 0$. The solution to this program is characterized by the following system of first-order conditions, yielding to Φ^E, q^E, p^E and e^E,

$$\frac{\partial \mathcal{L}}{\partial \Phi} = b(q) - p = 0 \tag{35}$$

$$\frac{\partial \mathcal{L}}{\partial q} = (1+\Phi)b'(q) - (1+\lambda)c_2 = 0 \tag{36}$$

$$\frac{\partial \mathcal{L}}{\partial p} = \mathbb{E}\left[\lambda \int_{\widehat{z}_\gamma(p)}^{\overline{z}} dF(z) - (1+\lambda)\widehat{z}_\gamma(p) f(\widehat{z}_\gamma(p))\right] - \Phi = 0 \tag{37}$$

$$\frac{\partial \mathcal{L}}{\partial e} = \rho'(e)\left[\Omega(p) + \lambda p \int_{\widehat{z}_h(p)}^{\widehat{z}_\ell(p)} dF(z)\right] - (1+\lambda)\theta = 0 \tag{38}$$

To undertake comparative statics with respect to λ, we completely differentiate this system of first-order conditions. We assume that the second-order conditions for a maximum hold, so we can apply the Implicit Function Theorem. After some algebra,

we obtain

$$\frac{dq}{d\lambda} = \frac{\rho''(e)\Big[\Omega(p) + \lambda p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big]\Big]}{-|H|}\Bigg[c_2 + b'(q)\Bigg(-\mathbb{E}\big[1 - F(\widehat{z}_\gamma(p)) - \widehat{z}_\gamma(p)f(\widehat{z}_\gamma(p))\big] - \frac{\rho'(e)\Big[\lambda\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] + (1+\lambda)\big[\widehat{z}_\ell(p)f(\widehat{z}_\ell(p)) - \widehat{z}_h(p)f(\widehat{z}_h(p))\big]\Big]}{\rho''(e)\Big[\Omega(p) + \lambda p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big]\Big]} \cdot\Big[\theta - \rho'(e)p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big]\Big]\Bigg)\Bigg], \quad (39)$$

where $|H|$ is the determinant of the bordered Hessian matrix of (34). As the fulfilment of the second order conditions implies that $-|H| > 0$, the necessary and sufficient condition that ensures that (39) is negative is

$$c_2 > b'(q)\Bigg(\mathbb{E}\big[1 - F(\widehat{z}_\gamma(p)) - \widehat{z}_\gamma(p)f(\widehat{z}_\gamma(p))\big] + \left(-\frac{\rho'(e)}{\rho''(e)}\right)\frac{\lambda\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] + (1+\lambda)\big[\widehat{z}_\ell(p)f(\widehat{z}_\ell(p)) - \widehat{z}_h(p)f(\widehat{z}_h(p))\big]}{\Omega(p) + \lambda p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big]} \cdot\Big[\rho'(e)p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] - \theta\Big]\Bigg)$$

To find an upper bound for the expression in parenthesis, we proceed as follows.

- $\mathbb{E}\big[1 - F(\widehat{z}_\gamma(p)) - \widehat{z}_\gamma(p)f(\widehat{z}_\gamma(p))\big] < 1$.
- As $\widehat{z}_\ell(p) > \widehat{z}_h(p), F(\widehat{z}_h(p)) < F(\widehat{z}_\ell(p)) \leq 1$. Hence, $\lambda\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] < \lambda$.
  We can also show that the function $zf(z)$ increases with z (see the proof of Lemma 2). Hence, $(1+\lambda)\big[\widehat{z}_\ell(p)f(\widehat{z}_\ell(p)) - \widehat{z}_h(p)f(\widehat{z}_h(p))\big] < (1+\lambda)\widehat{z}_\ell(p)f(\widehat{z}_\ell(p)) < (1+\lambda)\overline{z}f(\overline{z})$.
  So, $\lambda\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] + (1+\lambda)\big[\widehat{z}_\ell(p)f(\widehat{z}_\ell(p)) - \widehat{z}_h(p)f(\widehat{z}_h(p))\big] < \lambda + (1+\lambda)\overline{z}f(\overline{z})$.
- Let's rewrite the first-order condition (38) as follows

$$\underbrace{\rho'(e)\Omega(p) - \theta}_{A} + \lambda\Big(\rho'(e)p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] - \theta\Big) = 0.$$

  By Lemma 2, $A < 0$. Hence, $\rho'(e)p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] - \theta > 0$.
  Moreover, $\rho'(e)p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] - \theta < \rho'(e)p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big]$.

- As we have shown in the proof of Lemma 2 that $\Omega(p) > 0$,

$$\Omega(p) + \lambda p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] > \lambda p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big].$$

With all these results, we have

$$\begin{aligned}
&\mathbb{E}\big[1 - F(\widehat{z}_\gamma(p)) - \widehat{z}_\gamma(p) f(\widehat{z}_\gamma(p))\big] + \\
&\quad\left(-\frac{\rho'(e)}{\rho''(e)}\right)\frac{\lambda\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] + (1+\lambda)\big[\widehat{z}_\ell(p) f(\widehat{z}_\ell(p)) - \widehat{z}_h(p) f(\widehat{z}_h(p))\big]}{\Omega(p) + \lambda p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big]} \\
&\qquad .\big[\rho'(e) p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] - \theta\big] \\
&< 1 + \left(-\frac{\rho'(e)}{\rho''(e)}\right)\frac{\lambda + (1+\lambda)\overline{z} f(\overline{z})}{\lambda p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big]}\rho'(e) p\big[F(\widehat{z}_\ell(p)) - F(\widehat{z}_h(p))\big] \\
&= 1 + \left(-\frac{(\rho'(e))^2}{\rho''(e)}\right)\left(1 + \frac{1+\lambda}{\lambda}\overline{z} f(\overline{z})\right).
\end{aligned}$$

Moreover, as $\frac{1+\lambda}{\lambda}$ decreases with λ, we know that

$$\frac{1+\lambda}{\lambda} < \frac{1+\tilde{\lambda}}{\tilde{\lambda}},$$

where $\tilde{\lambda}$ is the lowest value of λ such that $p = b(q)$.

Also, as the function $\rho(e)$ is increasing and concave in e,

$$-\frac{\rho'(e)}{\rho''(e)} < \left|\frac{\big(1 + (\rho'(e))^2\big)^{\frac{3}{2}}}{\rho''(e)}\right| \equiv R(e),$$

where $R(e)$ is the radius of the curvature of the function $\rho(e)$ at the point e.
So finally, if

$$c_2 > b'(q^E)\left[1 + R(e^E)\left(1 + \frac{1+\tilde{\lambda}}{\tilde{\lambda}}\overline{z} f(\overline{z})\right)\right],$$

then

$$\frac{dq^E}{d\lambda} < 0.$$

we conclude. ∎

Proof of Corollary 3

We present a proof that applies when the conditions that enable the optimal price-quality scheme to reach the boundary $b(q)$ are satisfied.

Let $\Delta\mathbb{E}W_f \equiv \mathbb{E}W_f^E - y$ be the difference between the expected welfare when the market for good x is active and optimally regulated, with $p^E = b(q^E)$ and when the regulator initially shuts down the firm. Applying an envelope argument, the expected welfare $\mathbb{E}W_f^E$ decreases with λ and diverges to $-\infty$. Hence, Bolzano's Theorem ensures that there exists $\overline{\overline{\lambda}} > \widetilde{\lambda}$ such that, when $\widetilde{\lambda} < \lambda < \overline{\overline{\lambda}}$, the market for good x should be active. Otherwise, the firm's shutdown is optimal.

A similar argument can be used to show that, if the conditions that enable the optimal price-quality scheme to reach the boundary $b(q)$ are not satisfied, then there exists a threshold $\overline{\overline{\lambda}}$ such that, when $\lambda > \overline{\overline{\lambda}}$, the regulator optimally shuts down the firm. ∎

Appendix B Regulation with moral hazard

We take the model of Section 4, but we assume that e is not observable. Therefore, the regulator must adjust V_ℓ and V_h to induce the firm to choose a particular level of effort. The problem to solve is

$$\max_{p,e,V_\ell,V_h} \ \mathbb{E}W \equiv \rho(e)W_h(p,e) + (1-\rho(e))W_\ell(p,e) \tag{40}$$

$$\begin{aligned} \text{s.t. } \ V_h(p,e) \ &\geq 0, \quad (VP_h) \\ V_\ell(p,e,) \ &\geq 0, \quad (VP_\ell) \\ e \in \underset{\tilde{e}}{\text{argmax}} \ \mathbb{E}\left[V_i(p,e)\right] \ &\quad (IC_{MH}). \end{aligned}$$

Problem (40) has the same structure than (7), with an additional incentive constraint (IC_{MH}). In the remainder of this section, we assume that IC_{MH} is completely characterized by its first-order condition. The following lemma characterizes the ex-post voluntary participation constraints.

Lemma 5 *If* $e^E > 0$, $V_h^E > V_\ell^E = 0$.

Proof. Consider that, at the optimum, $e^E > 0$. Assume first that constraints VP_ℓ and VP_h are slack. If so, we can always decrease V_ℓ^E and V_h^E by a small $\epsilon > 0$ such that both constraints still hold. Moreover, since the constraint (IC_{MH}) is completely characterized by its first-order condition

$$\rho'(e)\left(V_h - V_\ell\right) - \theta = 0, \tag{41}$$

then the firm's ex-post utilities in the new regulatory scheme also satisfies (IC_{MH}). These reductions in ex-post utilities lead to an increase in the expected welfare, which is a contradiction.

Now assume that VP_ℓ and VP_h bind. This implies that (41) becomes $-\theta < 0$ and so the profit maximizing level of effort is $e^E = 0$. This contradicts the initial assumption that $e^E > 0$.

Finally, assume that constraint VP_h binds, while VP_ℓ is slack at the optimum. Again, this implies that (41) becomes $-\rho'(e)V_\ell - \theta < 0$, and so again the profit maximizing level of effort is $e^E = 0$, which is a contradiction.

The unique alternative that does not lead to a contradiction is the one conjectured in the lemma. ∎

As is standard in moral hazard problems, optimal regulation implies letting some informational rents to the firm. Indeed, to induce the monopolist to exert any positive level of effort, it must have a stake in the low evasion scenario, when $\gamma = \gamma_h$.

As before, the only relevant regulatory schemes are those for which $p \leq b(q)$. Using the previous lemma, problem (40) becomes

$$\max_{p,e} \ \mathbb{E}W = \mathbb{E}W^E - (1+\lambda)\rho(e)\frac{\theta}{\rho'(e)}. \tag{42}$$

where $\mathbb{E}W^E$ is given by (7) and the second term corresponds to the expected rents left to the firm in order to induce it to exert effort. The next proposition characterizes the optimal price-effort scheme.

Proposition 9 *When $p^E < b(q^E)$, the price p^E and the effort e^E are characterized by*

$$(1+\lambda)\mathbb{E}\left[\widehat{z}_\gamma(p^E)f(\widehat{z}_\gamma(p^E))\right] = \lambda\mathbb{E}\left[\int_{\widehat{z}_\gamma(p^E)}^{\overline{z}} dF(z)\right] \tag{43}$$

$$\rho'(e^E)\left(\Omega(p^E) + \lambda p^E \int_{\widehat{z}_{\gamma_h}(p^E)}^{\widehat{z}_{\gamma_\ell}(p^E)} dF(z)\right)$$
$$\leq (1+\lambda)\theta\left(2 - \frac{\rho(e^E)}{\rho'(e^E)}\frac{\rho''(e^E))}{\rho'(e^E)}\right), \text{with equality if } e^E > 0. \tag{44}$$

Proof. Differentiating the function $\mathbb{E}W$ with respect to p and e, and rearranging, we obtain the first-order conditions (43) and (44). ∎

The only difference between this system of first-order conditions and (8)-(9) is the presence of the informational rent in (44). Moreover, conditional on the optimal level of effort e^E, the pricing rule given by (43) is identical to (8), when effort is observable. In other words, the pricing rule is not distorted by the presence of moral hazard.

This finding is reminiscent of the "dichotomy property" obtained by Laffont and Tirole (1990). At first glance, this similarity with their result could seem obvious because our total cost specification satisfies the separability condition that ensures that property.[38] But things are not that simple. In Laffont and Tirole's model, effort only has an impact on firm's costs.[39] In our model, effort pushes up the likelihood of a higher level of enforcement, which leads to more formal purchases. Therefore, effort also affects the expected formal demand of good x, and thus one may have expected the pricing rule to be adjusted according to the effort level e. The intuition for this generalization of the dichotomy property lies on the fact that, to balance the marginal benefits and costs of a price change, the regulator only needs to set optimally the expected marginal evader. And effort only affects the identity of this individual through the weights $\rho(e)$ and $1-\rho(e)$, which are independent of the price.

The following corollary shows how p^E and e^E compare their levels found in the previous sections.

Corollary 4 *The price and the level of effort are lower with moral hazard than those when effort is observable.*

Proof. Consider the following fictitious program,

$$\max_{p,e} \mathcal{L} \equiv \quad \alpha\, \mathbb{E}W^E + (1-\alpha)\, \mathbb{E}W^{E,MH}, \tag{45}$$

where $\mathbb{E}W^{E,MH}$ and $\mathbb{E}W^E$ are the optimal expected welfares, with and without moral hazard, respectively. Notice that when $\alpha = 0$ ($\alpha = 1$), $\mathcal{L}$ coincides with $\mathbb{E}W^{E,MH}$ ($\mathbb{E}W^E$). So we can conduct comparative statics with respect to α. Given our differentiability assumptions, we compute the following cross derivatives,

$$\begin{aligned}
\mathcal{L}_{\alpha p} &= 0 \\
\mathcal{L}_{\alpha e} &= (1+\lambda)\theta\left(1 - \frac{\rho(e)}{\rho'(e)}\frac{\rho''(e)}{\rho'(e)}\right) \geq 0 \\
\mathcal{L}_{pe} &= \alpha\frac{\partial^2 \mathbb{E}W^E}{\partial p \partial e} + (1-\alpha)\frac{\partial^2 \mathbb{E}W^{E,MH}}{\partial p \partial e} \geq 0
\end{aligned} \tag{46}$$

As the function $\mathcal{L}$ is supermodular with respect to (p, e, α), Topkis' Monotonicity Theorem implies that p and e increase with α, which proves the corollary. ∎

[38] See Proposition 3 in their paper.

[39] More precisely, effort decreases the ex-post observable cost.

Since moral hazard entails the extra cost of leaving rents to the firm when evasion is low, the regulator ends up choosing a lower level of effort than when the latter was observable. As this implies that changing the price has a smaller marginal benefit, the regulator settles on a lower price as well.