Global Health Donors Viewed as Regulators of Monopolistic Service Providers: Lessons from Regulatory Literature

Han Ye

Abstract

Controlling healthcare costs while promoting maximum health impact in the recipient countries is one the biggest challenges for global health donors. This paper views global health donors as the regulators of monopolistic service providers, and explores potential optimal fund payment systems under asymmetric information. It provides a summary and assessment of the prevailing optimal price regulation designs for monopolistic serviceproviders. A set of non-Bayesian approaches that are relevant and applicable for the global health donors are discussed. It also reviews incentive contracting experience between the public and private sectors.

Keywords: healthcare costs, global health donors, monopolistic service providers, incentive contracting.

Center for Global Development
www.cgdev.org

Working Paper 424
February 2016

Global Health Donors Viewed as Regulators of Monopolistic Service Providers: Lessons from Regulatory Literature

Han Ye
Boston University
yehan@bu.edu

This paper was commissioned by Center for Global Development (CGD) as a background paper for CGD Working Group on Next Generation Financing Models in Global Health - see Silverman et al. [2015] for the final report. I would like to thank Sebastian Bauhoff, Randall P. Ellis, Paul Gertler, Rachel Silverman, and Ingo Vogelsang for critical feedback and valuable suggestions. I am especially grateful to Mead Over for helpful comments and suggestions. I am grateful to all participants at the April 2015 technical workshop, and the labor reading group of Boston University. All errors and omissions are mine. Comments are welcome.

Han Ye. 2016. "Global Health Donors Viewed as Regulators of Monopolistic Service Providers: Lessons from Regulatory Literature." CGD Working Paper 424. Washington, DC: Center for Global Development.
http://www.cgdev.org/publication/global-health-donors-viewed-regulators-monopolistic-service-providers-lessons-regulatory

Center for Global Development
2055 L Street NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

Foreword

This paper – and the companion paper by Liam Wren-Lewis (Working Paper 425) – were commissioned as background papers for a CGD working group entitled: "Next Generation Financing Models in Global Health". Funded by a grant from the Bill & Melinda Gates Foundation, the working group's purpose has been to advise the Global Fund for AIDS, TB & Malaria (GFATM) on how they can improve the value for money from their grants for health service delivery in recipient countries. This working group is thus delving more deeply into a key "domain" of opportunities for improved GFATM efficiency, as discussed in from CGD's 2013 study on the GFATM, "More Health for the Money".

Both of these papers can be viewed as contributions to the literature on "Cash-on-Delivery" approaches to the financing of foreign assistance pioneered by the Center for Global Development. The writings on the "COD" model have analyzed and advocated shifting from cost-reimbursement or input financing to payments for verified outputs, a shift that would be impossible if outputs could not be measured or verified. These papers, and indeed all of the work of the working group that commissioned them, share with the COD literature the assumption that a substantial portion of global health aid supports the delivery of services which *can* be measured and verified. This work on incentive design asks the next question: Among all the ways to condition payments on performance, which contract designs would generate the most powerful incentives towards improved efficiency.

These two papers both start from the observation that global health donors such as the Global Fund for TB, AIDS & Malaria and the US PEPFAR program can be viewed as *purchasing agencies* with the mandate to purchase health care services on behalf of poor beneficiaries in recipient countries. Both papers adopt as a stylized fact that the contractors from whom donors purchase these services have monopoly power in the delivery of the services to their constituents and monopsony power in their bargaining relationship with the donor. Under these assumptions, the donor's objective of assuring the delivery of quality services to the beneficiaries at the lowest sustainable cost can be compared to the problem of the regulator of a natural monopoly, a problem that has received decades of attention in the economics literature within the sub-fields of industrial organization, mechanism design and the theory of regulation. From this economic perspective, the global health donor is a "principal" while the contractor within the recipient country, whether a government agency, a non-governmental organization or a UN agency, is the donor's "agent". The relationship between the two is characterized by asymmetric information, with the agent knowing more than the principal about the current cost of service delivery of any specified quality and about opportunities for cost reduction.

Han-Ye's paper survey's the incentive mechanisms that have been proposed in the theory of regulation, asking how each of these could be adapted for use by a global health donor, or principal, in order to improve the efficiency of its contracting procedures vis-à-vis its "agents". She weighs the pros and cons of each reviewed mechanism and highlights the tradeoff between the power of its incentives and the cost of its information requirements,

providing options from which global health donors might select contractual features for experimentation.

Liam Wren-Lewis's paper starts from the same assumptions, but goes deep rather than wide. As currently practiced by the GFATM and other health donors, contracting is essentially cost-reimbursement or input financing. A common result from contracting theory reflected, for example, in the "efficiency wage" literature in labor economics, is that a cost-reimbursement contract would be the principal's best choice when the agent's performance is impossible or costly to monitor. Wren-Lewis' paper suggests that a health donor could offer a contract that would give the agent/contractor the option to choose, at the end of the contract period, whether to receive a conventional reimbursement of its previous period's costs or, alternatively, to receive a payment proportional to the number of verified units of output it had produced. The Wren-Lewis paper shows that this two-choice "menu contract" is the simplest version of a more elaborate multi-choice menu contract in the mechanism design literature and, despite its simplicity, can achieve up to 80% of the more complex contract's efficiency gains. By building up his efficiency-enhancing contract designs from familiar elements like the cost-reimbursement contract, Wren-Lewis's paper offers the global health donor an incremental path towards more powerful contracting.

As suggested by their origins within the policy context of rich-country regulation of natural monopolies, the contract designs discussed in these two papers could also be used to structure the contractual relationship between a national government and the subnational entities such as states, provinces, or NGOs that will deliver health or other services.

The Center for Global Development publishes these two background papers with the hope that they will spark the growth of a literature on efficiency enhancing mechanism design in the non-profit sector in general and in global health in particular.

Mead Over
Senior Fellow, Center for Global Development

Table of Contents

List of Tables and figures

1. Introduction

The Global Fund for AIDS, TB and Malaria has endeavored to provide funding to local intuitions to reduce the morbidity and mortality from these diseases around the world. The financial support the Global Fund (GF) provides is mediated by a bilateral relationship between the GF and the Principal Recipient (PR) [1]. The grant agreement between the GF and the PR is established upon the evaluation of a financing proposal from the PR. The grant allocation is largely based on expenses. To achieve the maximum health impacts of the GF's investments, a more adaptive and flexible financing model needs to be considered [2]. The similarity between the GF-PR relationship and the government-monopoly relationship suggests regulatory literature as a natural knowledge pool to draw from. This paper summarizes the prevailing incentive designs of price regulation for monopolistic service providers. Those incentive designs not only link price with costs but also with other targets, such as output, quality and performance.

In this paper, we view global health donors as the regulator of monopolies. This is because the objective of the GF, to maximizing its health impacts, is similar to the objective of a benevolent regulator. Both have welfare maximizing objectives related to efficiency and redistribution concerns. Both have information disadvantages and can-not verify the production function and effort level of the monopolies (PRs). Both can regulate the price (reimbursement) [3] level or the revenues of the firms. Both face many administrative challenges when implementing the optimal contracts. Also, both are subject to strategic behaviors of the regulated firms (PRs) on a multidimensional strategy space, such as output, quality and product allocation. In addition, we view PRs as regulated monopolies. Both have considerable market power. Also, the profit maximizing problem of the regulated monopoly firm is not unlike the surplus maximizing problem of the PR.

[1] The PR is usually the national government. Sometimes, it is a corporate entity created in response to the availability of GF financing for the sole purpose of submitting proposals for that financing and for subcontracting or otherwise ensuring the delivery of the health services and products agreed on with the GF.

[2] See Glassman et al. [2013] for a report on More Health for the Money.

[3] Throughout this paper, I use prices, payments and reimbursements interchangeably. When we consider the payments to the monopolies, we do not distinguish whether the payments are made by the regulator to the monopolies via reimbursements, or by the consumers to the monopolies via prices. The regulated payments/prices influence the monopoly's budget constrains, therefore its profit-maximizing decisions. In this paper, I only focus on the behavioral responses of monopolies to incentives. Clearly, the payment schedules will change consumer behaviors, and therefore the firms' behaviors. However, this paper focuses on the first order problem of the PRs' response to different payment schedules.

The current practice of grant payment from GF to PR is to passively reimburse all incurred expenditures reported by PR. Because of asymmetric information between GF and PR, PR will take advantage of this payment method and will exert zero cost-saving effort. Without observing the production technology and cost-saving effort, expenditure based payment methods encourages wasteful behavior, and rewards inflated expenses. Similar results have also been observed under the Rate of Return Regulation (RORR), which was used in the telecommunication industry, and the cost-reimbursement contracts in government contracting. As monitoring and inspection are costly and imperfect, and each PR has its own unique organization practice and faces different health care environment, it is especially important for GF to put incentives into its financing models to induce PRs to act voluntarily in ways desired by the GF.

Incentive regulation has been broadly studied in the regulatory literature since 1970s, and implemented to regulate public utilities (Baron and Myerson [1982], Laffont and Tirole [1986], Laffont [1993], Vogelsang [2002], Joskow [2005], Sappington and Weisman [2010]). Viewing PRs as the regulated utilities [4] helps us to capture the lessons of setting per unit price from this literature and apply them when designing the GF-PR contract.

On the other hand, government contracting has been widely used by government agencies to procure a broad range of services, such as public transportation, water distribution and ancillary service in public education system. This public-private partnership is very similar to the GF-RP relationship. Viewing PRs as monopolistic service providers helps us to understand the advantages and disadvantages of different contract types. Moreover, the altruistic characteristic of monopolistic providers of certain services, such as public education, employment counseling and healthcare, sheds light on some unique challenges faced by the global health donors. We will also discuss relevant experiences on incentive design in this aspect.

This paper provides a summary assessment of the prevailing optimal price regulations for public utilities. It also reviews incentive contracting experiences between the public and private sectors. In the following section, I will briefly introduce the asymmetric information problems and the theoretical rationales for the various regulation instruments designed to

[4] Global health donors can be viewed as regulators of the monopolistic service providers due to the above mentioned similarities. However, most of the discussions in this paper are based on the institutional background of developed countries. When applying them to the innovative grant-making design for the GF, local institutional contexts have to be recognized.

circumvent the challenges. Section 2 reviews the most prevailing price regulations in public utilities, and discusses the lessons from each type of regulation for the GF-PR relationship. Trade-offs and unintended consequences are discussed. Section 3 discusses experiences from government contracting with for-profit monopolies, and provides empirical evidences on potential strategic behaviors of the monopolies. Section 4 focuses on the altruistic characteristic of nonprofit monopolies, and extends the results to the GF-PR relationship. Section 5 concludes.

2. GF-PR viewed as regulator-monopoly

The fundamental goal of the GF is to induce the PR to take actions desired by the GF that would not occur without monitoring and oversight. To be more specific, the primary regulatory objectives of both the GF and the regulators of monopolistic service providers include the following: fair price, allocative and operative efficiency, universal service provision and efficient investment (see Table 1). To achieve those goals, the regulators maximize social welfare by imposing incentives and constraints that influence the regulated firms' behaviors. Similarly, the GF can achieve "more health for the money" by designing payment systems with targeted incentives.

The welfare function of the GF can be simplified as

$$W = B - F, \tag{2.1}$$

where B is the benefit level of fund beneficiaries (consumers), and P is the fund payment transferred to the PR. The GF maximizes the benefit of the service provision B net of the fund payment F [5].

The GF maximizes its own welfare function, subject to the utility maximization decisions made by the PR. Assume the utility function of the PR is

$$U[\Pi, B] = \Pi + \alpha B = F - C + \alpha B, \tag{2.2}$$

[5]Here, we focus on GF's objective for each given project. The opportunity cost of fund payment transferred to a given PR is not taken into account in this setting. However, it is still meaningful because cost reduction of a given PR will leave GF with more funds available for other projects (PRs) , hence maximizing the health impact with less money.

where α represents the altruistic preference of the PR. C is the total costs, and profit is bounded by zero. For most of the paper, we assume $\alpha = 0$, then PR's utility is

$$U[\Pi, B] = F - C. \tag{2.3}$$

Table 1: GF-PR viewed as regulator-monopoly

	Regulator	The Global Fund
	Max W=CS+PS*	Max W=B-F
Objectives	Maximizing social welfare	Maximizing health impact
	1) Efficiency concerns	1) Efficiency concerns
	2) Redistribution concerns	2) Fairness concerns
	3) Budgetary concerns	3) Save funding for other countries in need
Policy instrument	Have discretion over price	Have discretion over fund payment
	(price regulation)	(fund financing model)
	Monopolist service provider	The Principal Recipient
	Max Profit=pq-C	Max Surplus=F-C=fq-C
	s.t.: price regulation	s.t.: fund financing model
Objectives	Maximizing profits	Maximizing surplus
Market Power	Government-grant monopoly	Contract-granted monopoly
Actions	Multidimensional strategy space	Multidimensional strategy space
	(effort, quality, performance)	(effort, quality, performance)
Information problem	Asymmetric and uncertainty	Asymmetric and uncertainty

Source: Author's construction. * CS: consumer surplus. PS: producer surplus.

The exclusive partnership between GF and PR in a given country creates barrier to entry, and gives the PR monopoly power in the local health market. This monopoly power gives the PR the ability to determine the units of service provision by altering the accessibility of the service. Even if PR internalizes all the costs, as a monopoly, it will provide too few service, and most likely only to a small group of consumers. This brings out the question: how could the GF design a optimal payment system that induces the PR to produce the maximum health impact efficiently?

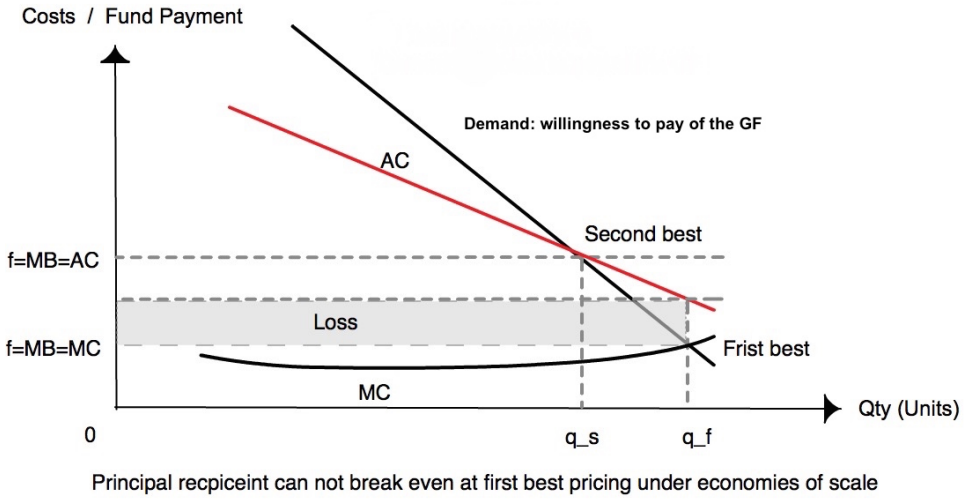
3. Problems of incomplete information

3.1. What is optimal price regulation?

The main challenges faced by the GF are the monopolistic nature of the PR and the asymmetric information between the GF and the PR. To copy with such challenges, the regulatory literature has studied optimal pricing mechanisms.

In regulatory literature, the definition of optimal price regulation has changed as the understandings of the monopolies increases. Under complete information, there are "first-best" pricing and "second-best" pricing. When the GF knows the exact cost function of the PR, the unconstrained "first-best" pricing implies that she will set the fund payments such that the marginal benefit to the GF equals the marginal cost to the PR. However, to establish the contract, GF has to ensure that the financial suitability of the PR. Due to PR's decreasing cost function, the "first-best" price cannot support PR to break-even. The "second-best" pricing is derived under an optimization problem with break-even constraints. It is also called Ramsey pricing [6].

Figure 2: An illustration of "first best" and "second best" pricing



Source: Author's construction.

[6] Ramsey outcomes are the price and quantity achieved under an optimization problem with the break-even constraints. The question was first raised by Ramsey [1927] in the context of optimal taxation.

Figure 2 demonstrates a simple example of a PR that could not break even at "first-best" pricing under economy of scale. If there exists economy of scale, and if the average cost curve is downward-sloping and above the marginal costs curve, then the participation constraint is binding. For a one-good situation, the efficient price equals the average cost. For a multiple-goods situation, Ramsey pricing suggests that per unit payment should be set up as marginal cost plus markup. The markup is inversely related to the demand elasticity. In the global health context, the Ramsey prices can be interpreted as an optimal allocation of funds to different healthcare services. The projects with inelastic health impact should be given less fund payment. The projects whose impact is more sensitive to funding should be given more funds. Such allocation provides the least distortion from the efficient level. However, it should be noted that sometimes Ramsey prices can contradict distributional objectives. For example, consumers who have inelastic demand for certain goods are mostly the ones who are most in need (Train [1991]). Similarly, projects that are less cost effective could benefit the patients who are most at risk.

However, Ramsey pricing requires detailed information on cost functions, demand functions and budget constraints. In practice, complete information is impossible[7]. Under incomplete information, optimal price regulation is defined as a set of policies that induce firms to act voluntarily in ways desired by the regulator. Such policies are influenced by Bayesian approaches and non-Bayesian approaches. Bayesian approaches derive the exact optimal contract by solving a well-defined principal-agent problem. The name "Bayesian" comes from the requirement of prior knowledge of probability distribution of the regulated firms' unobserved characteristics, such as cost types. It is also referred to as incentive regulation or mechanism design (Laffont [1993], Vogelsang [1999]).

Non-Bayesian approach differs from Bayesian approach in the following aspects: 1) it emphasizes practical applications of the regulatory design. Instead of solving the exact optimal contract, non-Bayesian approach provide a stepwise improvement towards the efficient outcomes. 2) it has a delayed convergence to optimal outcomes. This means the PR will be allowed to have profits for a longer period [8]. 3) it only require observed bookkeeping. The optimal pricing mechanisms using non-Bayesian approach are designed to achieve effi-

[7] In practice, either the GF nor regulators have access to the production technology of the PR and the regulated party. It is impossible to fully observe the interactions between the cost reduction and the managerial effort of the PR. In addition, the impact of changes in accessibility on demand induced by changes in fund payments is not perfectly understood by the donor and the PR.

[8] See Vogelsang [2002] for a discussion on Bayesian and non-Bayesian mechanisms.

ciency improvement under incomplete information, and to ensure that the efficiency gains are shared with the consumers over time.

Regulations based on non-Bayesian approach is more practical for the GF to adopt, especially in the initial stage of projects, when little information about cost distribution is known. In this paper, our focus will be on pricing mechanisms based on non-Bayesian approach. Price cap, RPI-X price cap, sliding scale, menu, and two-part tariff are examples of the prevailing incentive regulations in the public utility sectors. I will discuss more on each of them later, and discuss their application in the Global Fund context.

3.2. Adverse selection and moral hazard

Before we discuss the details of the prevailing incentive regulations, let us look at two well-recognized market failures:adverse selection and moral hazard problem. They arise from the conflicts of interests between the regulator and monopoly, and asymmetric information.

The consequence of Adverse selection is that only inefficient PRs exist in the market due to asymmetric information. This is because when fund payment is set such that the PR is financially sustainable, PR pretends to have a higher cost profile than it actually does, believing that higher costs or reported expenses of institutions will be granted more funding. Without the knowledge of the PR's actual cost opportunities, the GF can not distinguish high cost PRs and low cost PRs. If the GF passively reimburses all reported expenses, PR has incentive to report artificially high expenses.

In the regulatory literature, Rate of Return Regulation (RORR) is designed to address the adverse selection problem[9]. Under RORR, firms are allowed to change prices only if the realized rate of return on capital changes. By auditing the costs, or at least the realized expenditure of the production process, the regulator obtains information on the firm's production function. Furthermore, by restraining the firm to earn no more than the realized rate of return on the capital investment, the regulator leaves the firms with no room for rent extraction. RORR belongs to the set of mechanisms that influence firm behaviors by setting constraints[10]. The current status quo "cost reimbursement" fund payment system of the GF is similar to RORR. However, when there is no profit left for firms, the lack of incentives brings up another problem – the moral hazard problem.

[9]The monopolies extract surplus from the consumers by posting artificially high prices. Low cost firms can not be distinguished from high cost firms.

[10]The constraints can be set according to sales, outputs and costs.

Moral hazard problem originates from the fact that managerial effort is difficult to observe and monitor. Here, we assume that effort lowers production costs, improves efficiency, but generates greater cost for the manager. The per unit cost function is $c(q,e)$, where e is effort level, q is unit of output. Assume $c'(e) < 0$, $c"(e) > 0$, and the cost of exert effort is β. Assume economy of scale, that average cost decreases as output increases, $c'(q) < 0$, $c"(q) > 0$. The average profit function of the PR is $\pi = f-(c(q,e)+\beta)$, where f is per unit fund payment. Theoretically, if fund payments are adjusted instantly according to the realized costs, there is no incentive for cost minimization. PR produces inefficiently and the cost inflation is borne by consumers and donors. Price cap regulation has been applied in the public utility regulation to mitigate the moral hazard problem. Price is bounded by a price cap that is adjusted based on exogenous cost changes and performance benchmarks. Individual firm's regulated price is not directly related to the firm's costs. Under price cap regulation, firms have incentives to reduce cost, because the margin between the price cap and cost will be entirely captured by the firm. Similarly, if the payment to a PR is independent of the PR's realized costs, PR is incentivized to lower total spending. price cap regulation is one of the non-Bayesian regulation we will discuss later in this paper.

3.3. Uncertainty

Another problem of incomplete information comes from uncertainty. Moreover, adverse selection and moral hazard issues often interact with uncertainties. Even if the information is symmetric ex-ante between GF and PR, the existence of uncertainty makes it impossible to have the bilateral contract written down completely. Uncertainties, such as unanticipated environmental changes, exogenous fluctuation of input prices, changes in the regulatory environment, and unanticipated fluctuation of demand, make it difficult to contract everything. Uncertainties call for contract design. Apart from regulating per unit payment, what are the other aspects the GF should consider when design a GF-PR contract under uncertainties? Fortunately, with the developments in contract theory, and improvements in measurement technology, non-contractible regulatory targets become contractible. We will draw experience from lessons produced in government contracting literature, and discuss relevant lessons.

4. Put incentives into disbursement: experience from public utilities

The GF aims to design a disbursement method that links payment to desired targets, such as cost reduction. The price level regulation in the regulatory literature provides valuable lessons. Price level regulation incentivizes firms to comply to regulation objectives by regulating permissible price structures. Specifically, price per unit of output is regulated. In the GF context, that is the disbursement per unit of output [11]. Different permissible price structures represent various incentives[12]

We proceed with the review from the Vogelsang and Finsinger mechanism to the most widespread RPI-X regulation, and follow with benchmarking regulation, revenue/profit sharing regulation, menus and two-part tariff. The adaptability of these price regulations depends on the availability of data, the cost profile of the principle recipient, and the trade-offs between different regulatory objectives.

4.1. Vogelsang and Finsinger (V-F) mechanism

The Vogelsang and Finsinger (V-F) mechanism (Vogelsang and Finsinger [1979]) has been suggested as an exemplary regulatory mechanism for the Global Fund that is both attractive and simple (Glassman et al. [2013]). Compared to the complex Ramsey pricing, V-F requires much less information. The core of the V-F mechanism is the constraint R_t. At each regulatory period, regulator grants the regulated firm freedom to choose any prices within the constraint R_t, which is defined as

$$R_t = [p_t | q_{t-1} p_t - C(q_{t-1}) \leq 0], \tag{4.1}$$

where p is the price, q is the quantity and $C(q)$ is the total costs of the firm.

The constraint R_t requires that for each period prices are allowed, as long as when multiplied by the last period's output they do not exceed the previous period's total costs [13]. The

[11] The output level could be measured in terms of service provided, patients admitted, products sold or cases solved, depending on the information used.

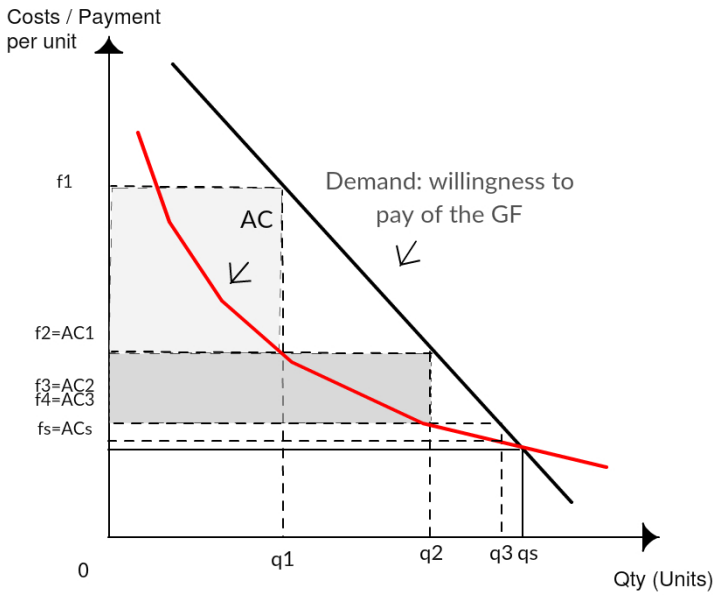
[12] In the regulatory theory literature, cost reduction is one of the most important objective of the regulator. cost saving incentive is called the "power" of a contract in contract theory.

[13] The result is based on the following assumptions: stationary cost and demand function, no intertemporal cost and demand effects, firms know their cost function. The regulator can observe the realized output, price

firm maximizes profits each period under the constraint of the permissible set of prices. The optimal price level is defined by the Laspeyres price, that is, a weighted average price index with the weights being last period's quantity.

Because the zero-profit criteria is defined based on last year's cost, the V-F mechanism allows the firm to profit from the cost reduction in the current period. This provides incentives for cost minimization. It also makes sure the firm transfers the surplus from cost reduction to consumers in the next period through this dynamic component. Consumer surplus increases each period. Theoretically, the V-F mechanism pushes the regulated prices and overtime to iteratively converge to the second-best Ramsey outcomes.

Figure 3: A one-good example of V-F mechanism



Source: Author's construction based on Figure 5.1 of Train [1991]

Figure 3 exhibits the V-F mechanism with a simple one-good example. Figure 3 uses the same setting as Train [1991]. The average cost curve is downward-sloping due to economy of scale. The pseudo demand curve represents the willingness to pay of GF. It is also downward-sloping due to diminishing marginal benefit. Suppose prior to adjustments, PR has been given a per unit payment of f_1, and provides q_1. The realized average cost is AC_1 according

and last period's total cost (Vogelsang and Finsinger [1979]).

to accounting records. Guided by the V-F mechanism, in period 2, GF will adjust the per unit payment such that $f_2 q_2 \leq AC_1 q_1$. In a one-good situation, the average cost determines the per-unit fund payment. f_2 is set to AC_1. In response to the adjustment, PR will exert the necessary effort to minimize costs, and supply q_2 amount of service. The residual "net revenue" PR receives from the excess fund payment, the grey shaded area in Figure 3, creates incentives for reducing costs. By observing q_2, GF learns the average costs corresponding to q_2, and sets the ceiling for fund payment in period 3 to be AC_2. This process continues. The area of rent extraction in Figure 3 decreases over time. Because of economy of scale, PR supplies more services, until f_s and q_s, where PR breaks-even.

There are two components of the V-F mechanism that are worth noting. One is the length of the price adjustment period. The period can be one year, one month, or 3 to 5 years. The length depends on a joint consideration of data feasibility and the cost of price adjustment. Secondly, the mechanism achieves allocative efficiency in a multi-products situation. A set of price combinations is permissible under the V-F mechanism. Without knowing the detailed breakdown of the costs of each single product, the price combination can still reach efficient Ramsey prices under the V-F mechanism. It fully utilizes the firm's private information without dictating the price level of products.

From the perspective of the GF, the analogous constraint is

$$R_t = [f_t | f_t q_{t-1} - C(q_{t-1}) \leq 0], \tag{4.2}$$

where f_t stands for per unit fund payment the GF gives to the PR. Here, we assume no lump-sum reimbursement. The length of the adjustment period t depends on the funding cycle of the GF.

Box 1: A Simple Example[a]

Suppose there are two different types of services, A and B. Each was delivered at 500 units last year. The total cost of the combined outputs was $600,000. Under V-F, the donor agrees to reimburse the recipient $600 per unit of output this year, regardless of the allocation between A and B. The donor only needs to know 1) last year's total cost, and 2) last year's level of each output. If it is less costly to reduce the marginal cost of service A, then the recipient will voluntarily increase the output level of A to profit from the economy of scale, and invest more effort and capital to reduce the cost of A, until the ratio of marginal revenue of A and B equals the ratio of marginal costs of A and B. This process continues until PR reaches its zero profit constraint (Figure 4 illustrates the process).

[a]Based on the example provided on page 60 in the appendix 2 of More Health For the Money.

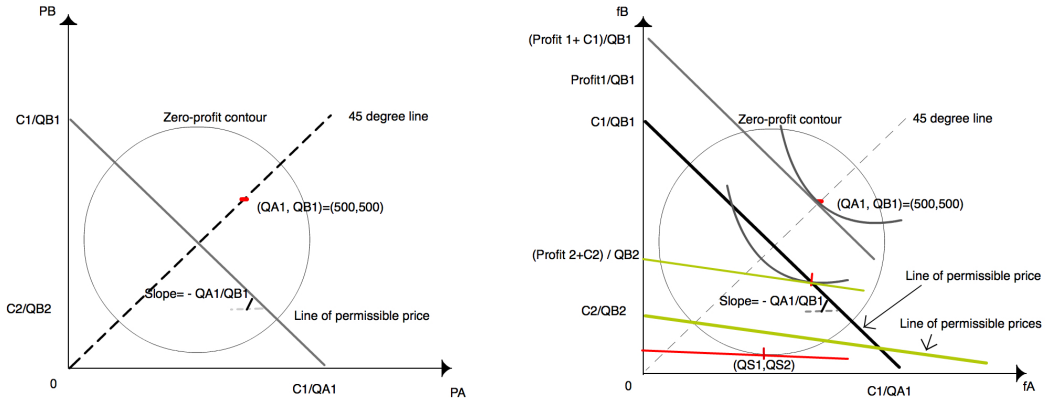
Figure 4 illustrates the example described above in Box 1. The set-up of the graphs is along the line of Figure 5.4 in Train [1991]. At period 1, Q_A^1 and Q_B^1 are 500 units each. Fund payments in period 2 are regulated along the constraint $f_A^2 Q_A^1 + f_B^2 Q_B^1 \leq C^1 = \$600,000$. The solid black line on the left-hand side of figure 2 is the line of permissible prices for period 2. [14]

Now, remember that we assume it is less costly to reduce the marginal cost of service A. In period 2, PR chooses to supply more Q_A^2. Each period, the line of permissible prices moves below the tangency line by $\frac{\Pi_t}{Q_B^t}$, which is the "profit" PR earned in period t divided by the output of service B at period t. The line of permissible prices is lowered each period as long as there are positive "profits". This process continues until PR is operating at zero profit. (Q_A^s, Q_B^s) is one possible output combination at zero profit. Each period, the VF mechanism transfers the efficient gain from cost reduction to the GF by lowering the required payments to the PR. This allows the GF to invest in more countries and areas in need.

In practice, there are two big challenges that impede the application of the V-F mechanism. One is the assumption of stationary demand and cost function. In reality, the regulatory environment changes constantly. It is unrealistic to assume constant cost function, and invariant demand. The other is the assumption that a firm maximizes profit every period,

[14]By rearranging $f_A^2 Q_A^1 + f_B^2 Q_B^1 \leq C^1$, we have $f_B^2 \leq \frac{C^1}{Q_B^1} - \frac{Q_A^1}{Q_B^1} f_A^2$. Train [1991] proves that the tangent line to the iso-benefit curve (social utility curve) at (Q_A^1, Q_B^1) has the same slope as the line of permissible prices.

Figure 4: A two-goods example of V-F mechanism

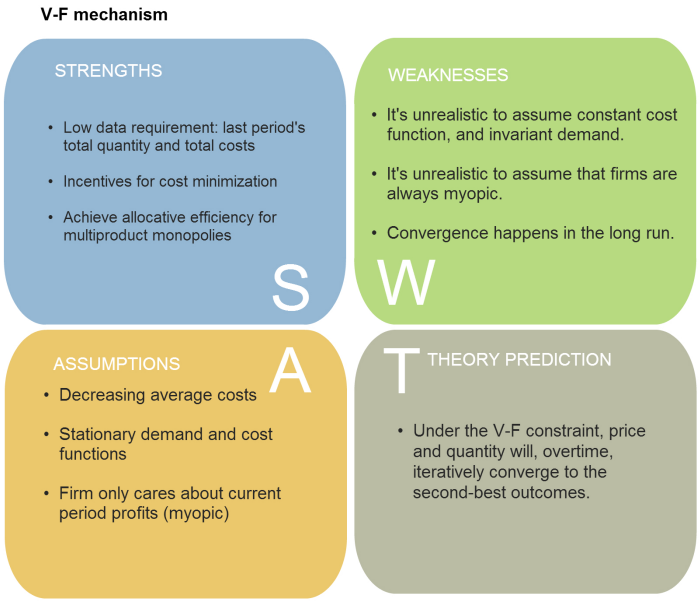


Source: Author's construction based on Figure 5.4 of Train [1991]

and loses that profit in the next period. Anticipating the scheduled payment adjustments, PR may engage in gaming. One potential strategic behavior is to incur wasteful costs. PR may choose not to improve efficiency to avoid facing more stringent constraints. [15].

The PR may also choose to incur unnecessarily high costs in one period to obtain higher permissible expenses in the subsequent period. Figure 5 illustrates a simple example of such wasteful behaviors. Assume the average cost is constant at \$600; at period 1 the fund payment prior regulation is set at \$800, Π_1=\$200*600=\$120,000. The recipient anticipates a fund payment adjustment in period 2. It also understands that the adjustment is based on the costs incurred in period 1. If PR does not manipulate, payment in period 2 is set to AC, $\Pi_2 = 0$. If PR chooses to inflate its costs in period 1 by W amount (\$100), the first period profits are halved to \$60,000. Now, at period 2, PR still has a profit margin of \$70,000. Assume at period 2, PR does not incur wasteful costs, and AC_2=AC=\$600. Whether it is profitable for PR to waste depends on the PR's discount factor, interest rate and the shape of the demand and cost functions. If we assume the interest rate is 10%, the net present value of wasteful behavior is \$12,300, higher than \$12,000. PR will choose to waste. However, if PR has a very high discount rate, such wasteful behavior is not the best response.

[15]Appendix 2 of *More Health For the Money* (Glassman et al. [2013]) suggests three reasons why VF is less vulnerable to this problem. It is important to notice that the recipients may share many of the donor's objectives, and they are not purely profit maximizing agents. Nevertheless, the lack of investment and innovation is a real concern. Moreover, the trade-off between high-powered incentives and investments has been a heated debate recently in the fields of telecommunication and other public utilities (Vogelsang [2013]).



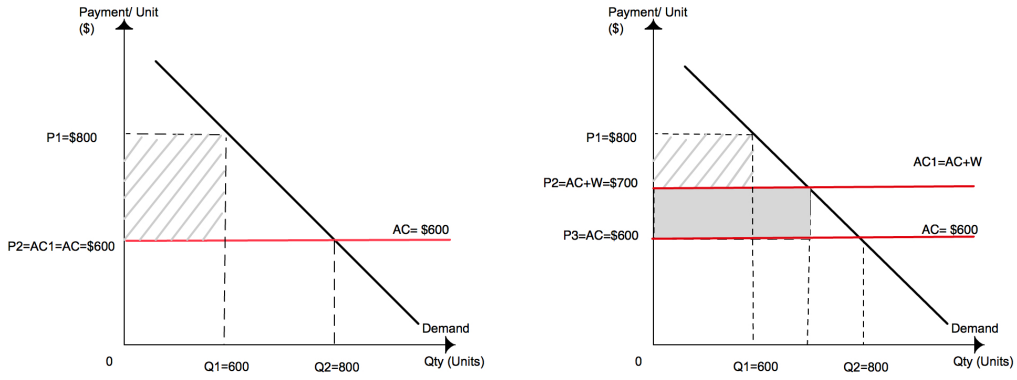
Source: Author's construction.

Another potential strategic behavior of the recipient is to misreport the actual cost. If PR misreports, profit in period 1 is still $120,000, but PR is allowed to have an addition $700,000 in profits in period 2. Similar to the adverse selection problem, in the belief that higher cost recipients will receive higher grants the next period, the recipients will claim to have a higher cost profile than they actually do. The PR extracts positive rents from the Global Fund by receiving higher than necessary financial supports. Random auditing with severe penalties can potentially deter such behaviors.

One possible remedy is a longer regulatory period. A longer period provides a more stable and viable cost reduction environment for the recipients. It enhances the incentives for efficiency. However, the longer the regulatory lag, the more likely the environment changes. For example, without adjusting for technology improvement, the recipient/firm will be granted too much money. It could also be that the recipient suffers financial distress due to input prices inflation.

To strike a balance between incentive and flexibility, some modifications need to be made. In practice, the UK Office of Gas and Electricity Markets (OFGEM) implemented a regulatory

Figure 5: Wasteful behavior in period 1- a one-good example



Source: Author's construction based on Figure 5.8 in Train [1991]

framework: RPI-X regulation. Here, RPI stands for retail price index. It adjusts for the change of input prices. X here is the efficiency factor. It adjusts the prices for technology changes. RPI-X is sometimes referred to as price cap regulation. It belongs to a class of incentive-based regulations. I will discuss more on the RPI-X regulation in the next section.

4.2. Price-cap regulation: the RPI-X formula

Price-cap regulation, sometimes known as RPI-X regulation, has been the most prevelant incentive mechanism in the public utility sector across the world. RPI-X is a dynamic long-term fixed price contract with a price adjustment formula. Prices are regulated at or below a specified fixed price cap. Note that the price cap is usually a weighted average of multiple products or services. The regulated firm has the discretion to adjust the prices of individual products.

For each price control period, the fixed price cap is set up based on the RPI-X formula. It is defined as:

$$p_t = p_{t-1}(1 + RPI - X), \tag{4.3}$$

where p_t is the per unit price cap the regulated firm faced at period t under RPI-X.

Now, consider the application for the GF. Assume last period's fund payment is estimated according to last period's average cost c_{t-1}, then the per unit fund payment to PR at period t

is:

$$f_t = f_{t-1}(1 + RPI_t - X_t) = c_{t-1}(1 + RPI_t - X_t). \tag{3.4}$$

And the per unit retained profits of PR at period t is:

$$\pi_t = f_t - c_t = f_{t-1}(1 + RPI_t - X_t) - ct = c_{t-1}(RPI_t - X_t) - (c_t - c_{t-1}). \tag{4.5}$$

The ceiling price level resets according to three components:

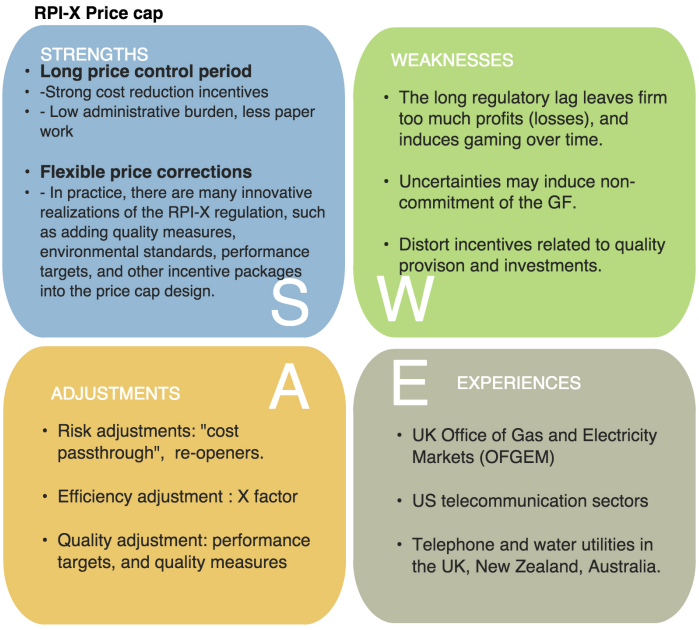
1) **RPI**: RPI is an adjustment for input price changes. Both retail price index (RPI) and consumer price index (CPI) are good measures of inflation. The main purpose of this adjustment is to estimate the economy-wide changes in input costs. It protects the regulated firms from economy-wide factor price changes. Similarly, adjustments on payments to PR should be made to accommodate the changes of input prices.

2) **X factor** : Price level has to be decreased by X percent to reflect productivity improvement. The enhanced productivity may come from cost reduction, technology innovation or economy of scale. The main purpose of the X factor is to share the efficiency gains with the consumers through lower prices in the next period. The entire average cost curve shifts down. It is designed to protect the consumers and the donors. With improvements in service delivery techniques and development of working knowledge in global health delivery, the efficiency improvement of PR allows the GF to finance fewer resources for the same amount of health impact. It is essential that the X factor is independent of the PR's individual costs to avoid suboptimal strategic behaviors. This is also the key difference between the RPI-X regulation and the V-F mechanism.

3) **The price control period** is usually 3 to 5 years for the public utility sector. This is intended to provide as much certainty as possible to the regulated firm. And, the initial price level p_0 at the start of the regulatory period is established using the rate of return method. It is worth noting that if the price review is conducted as frequently as the price review of RORR, the boundary between these two regulations blurs.

4.2.1. Properties of price-cap regulation

There are several interesting properties of the RPI-X regulation. First, the determination of the X factor is crucial to the viability of the price cap regulation. The productivity adjustment X is most often estimated based on projected long-run productivity growth trends. It



Source: Author's construction.

is usually estimated using statistical and programing techniques, such as benchmarking [16]. Secondly, RPI-X provides a feasible solution to the problems of the V-F mechanism. Vogelsang [2002] points out two attractive characteristics of price caps: 1) stable and viable cost reduction incentives, and 2) freedom and incentives for price rebalancing. The built-in adjustment of RPI-X formula provides a flexible contingency to avoid extreme profits or losses due to future uncertainties, while the fixed relatively long price control period gives the firms strong cost reduction incentives. The dynamic price review period also pushes prices towards the efficient Ramsey prices over time. Lastly, the price-cap regulation balances the risk bearing for the consumers and the regulated firms.

There are also some drawbacks of the price-cap regulation. First of all, the RPI-X regulation might distort incentives related to quality provision and investment. To maximize the profit margin, the regulated firm exerts less effort in quality improvement. When quality is difficult to monitor, and not related to the firms' revenue, the regulated firms are incentivized

[16] Benchmarking techniques measure the relative efficiency of individual firm's operating costs and service quality compared to their peers or a hypothetical efficient firm. See Online Appendix for more on yardstick and benchmarking regulation.

to compromise quality to achieve cost reductions. Adjustments to the incentive mechanism need to be made, such as linking price level with specified quality measures[17]. The cost reduction incentive encourages firms to adopt cost-reducing innovations. However, the price cap prevents the firm from obtaining all the profits from the investment. The firm will invest suboptimally in necessary innovative yet risky projects. In the long term, the lack of such investment slows the productive improvements. The trade-off between investment and high power incentive should be considered according to the specific circumstance[18].

In addition, there exist some exogenous firm-specific costs that are incurred due to events beyond the control of the regulated firm, such as tax changes. But, those costs are not adjusted in RPI-X. One variation of the price-cap formula addresses this issue. It takes the form of RPI-X+Y, where the Y factor is called "cost passthrough". Adding the Y factor alleviates the firm's uncertainty and encourages necessary investments. Last but not at least, the regulatory risk could be a potential threat to the price-cap regulation. The stable regulatory environment RPI-X offers relies on credible commitment of the regulators. We will discuss more on regulatory commitment in later sections.

4.2.2. Experiences with price-cap regulation

In the United States, more than 40 states have implemented price-cap regulation for the access and end-user prices in the telecommunication sector. Price-cap regulation is also the prevailing incentive mechanism adopted in the regulated segments of the privatized electricity, natural gas, telephone and water utilities in the UK, New Zealand, Australia and parts of Latin American.

In practice, the mechanism to create price caps is a combination of elements of different incentive-based regulations. RPI-X regulation is the most-widely adopted form of price-cap regulation. There are many innovative realizations of the price-cap regulations, such as adding quality measures, environmental standards, performance targets, technical efficiency goals and other incentive packages into the price cap design. For example, the GF can adjust up the price cap when the PR achieves critical treatment measurements, such as the 90% of HIV-positive key population on antiretroviral therapy. The RIIO (Revenue=Incentives+ Innovation+Outputs) model implemented by the UK Office of Gas and Electricity Markets

[17] See section 3.3.2.1 Quality shading.
[18] See section 3.3.1.1 Trade-offs between incentives and investments.

(OFGEM) in 2010 is one of the examples (OFGEM [2010]). The price control period is 8 years under this new regulatory framework. It emphasizes the benefit to future consumers. It includes an innovation stimulus package in the price control. The longer price control period of RIIO provides a more stable regulatory environment for investment. It encourages innovative investments that benefit current and future consumers. The RIIO also includes targeted incentives to ensure long-term value for money. Both financial and reputational incentives have been used.

4.3. Sliding scale regulation

Sliding scale regulation is sometimes called profit/revenue sharing regulation[19]. Sliding scale regulation was first used in England in the middle of the 19th century in the electricity supply industry (Schmalensee [1979]). Under sliding scale regulation, the regulated firm rebates to consumers a fraction of its profit (loss). Applied to the GF context, a payment system reimbursing a fraction of incurred current period costs is an example of sliding scale. The cost sharing is governed ex-ante by a a sliding scale s, where $0 < s < 1$.

$$f_t = a + sc_t \tag{4.6}$$

$$\pi_t = f_t - c_t = a + (1-s)c_t, \tag{4.7}$$

where a is a constant amount. a could be the average cost of per unit service provision in all developing countries involved in fighting HIV/AIDS, tuberculosis and malaria. a could also be the threshold when the GF starts to reimburse PR's excessive costs. a should not be linked with each PR's individual costs. When a is zero, the sharing schedule simply reimburses the PR a fraction of realized costs sc_t. As long as a is less than $(1-s)c_t$, the PR bears a positive share of the costs. The link between payment and realized costs is presented by s.

It is a flexible combination of RORR and pure price-cap regulation. Pure price-cap regulation passes all realized benefits to the consumers. Under RORR, all incurred costs are passed on to the consumers and the the firm has no incentive for cost reduction. Table 2 exhibits the outcomes under RORR, pure price-cap, sliding scale and the RPI-X regulation, respectively. (Vogelsang [1999], Mayer and Vickers [1996]).

[19] I will use cost sharing, revenue sharing, profit sharing and sliding scale interchangeable. The core of sliding scale is risk sharing.

Table 2: Comparison of different price regulations

	RORR	Sliding scale	RPI-X	Pure price cap
f_t	$r \simeq c_{t-1}$	$a + sc_t$	$\simeq c_{t-1}(1 + RPI - X)$	c_t
π_t	$r - c_t \simeq -\Delta c$	$a + (s-1)c_t$	$c_{t-1}(RPI - X) - \Delta c$	0
Cost reduction incentive	0	+	+	+
Risk bearing	0	(1-s)%	+	100%
Investment incentives	+	s%		-

Source: Author's construction. * $a < r, 0 < s < 1$ a is positively related with c_{t-1}.

In an ex-ante view, sliding scale regulation provides a risk sharing system between GF and PR. It reduces the revenue uncertainty of the PR, and facilitates investments. In an ex-post view, it prevents the PR from having abnormal excess profits or unexpected loss. The cost of healthcare service provision varies by patient-case. When PR bears zero risk, the cost reduction incentive is zero. When PR bears all the risk, such as in the pure fixed price situation, PR will have the tendency to select less costly patients, to lower the quality of service and avoid risky investments. The optimal sharing rule s can be derived by solving the maximization problem described in equation 2.1.

Box 2: Examples of sliding scale regulation

In practice, the sharing parameter *s* depends on the amount of deviation from the targeted threshold. One example is the sliding scale regulation adopted by the electricity company in Oregon from 1992 to 1996. Firm can keep all the revenue up to a predetermined threshold. After the threshold, the revenue is shared between the customers and the firm. Another example is the mandatory reinsurance under the Affordable Care Act Exchanges in the U.S. for the years 2014-2016. The reinsurance is set to be triggered by a threshold. It partially reimburses an insurer's costs when the average cost of the enrolled patients exceed the threshold (Geruso and McGuire [2014]).

Generally, the price level (payment level) is adjusted by a share of the difference between realized performance measure and target performance. Typically, a "deadband" around the targeted threshold is allowed before activating any rewards or penalties. A "deadband" defines a reasonable range of performance.

The most important advantage of sliding scale is its built-in fairness and self-rebalancing properties (Vogelsang [2002]). Compared to RPI-X regulation, sliding scale is simpler to apply.

However, it does not have a dynamic component. If the profit sharing adjustment happens annually, it is important to estimate the cost effectiveness of the incurred monitoring and administrating.

4.4. Menu

A menu offers the regulated firm a choice among cost contingent incentive plans. By choosing its preferred incentive scheme, the firm reveals its welfare-enhancing preference. However, the information disadvantage of the regulator and uncertainties of demand and costs put a heavy burden on the menu design. Issues such as differential incentives for operating costs and capital costs and misreporting of expenditure forecasts put obstacles in the application of the menu mechanism. In the United States, the menu approach was tired for the regional Bell operating companies in the 1990s (Vogelsang [2002]). However, it was abandoned after a few years on account of regulatory commitment problems.

Table 3: A simplified example of sliding scale matrix

Plant/firm/fund recipient	A	B	C	D	E
Estimated expenditure	100	110	120	130	140
Sliding Scale (efficiency incentive)	40%	35%	30%	25%	20%
	Rewards and penalties				
Performance target					
Allowed expenditure	105	107.5	110	112.25	115
Actual expenditure					
70	16.50	14.73	12.60	9.76	6.60
80	12.50	11.23	9.60	7.26	4.60
90	8.50	7.73	6.60	4.76	2.60
100	**4.50**	4.23	3.60	2.26	0.60
110	0.50	**0.73**	0.60	-0.24	-1.40
120	-3.50	-2.78	**-2.40**	-2.74	-3.40
130	-7.50	-6.28	-5.40	**-5.24**	-5.40
140	-11.50	-9.78	-8.40	-7.74	**-7.40**

Source: OFGEM [2004d] page 87

The sliding scale menu implemented by OFGEM in 2004 is one of the most direct applications of the menu of cost-contingent contracts. It allows the firms to self-select the optimal

sharing parameter from the provided menu (OFGEM [2004d], OFGEM [2010], Joskow [2005]). It is based on the difference between the allowed capital expenditure target chosen by the firm from the menu and the firm's actual capital expenditures during the five-year price control period (OFGEM [2004d]). Table 3 is a simplified and modified example based on the OFGEM menu. Based on their estimated capital expenditure, firms A, B, C, D, E choose a preferred sliding scale incentive from the matrix. Firm A expects to spend $100, and will choose a sliding scale of 40%. This is because under this contingent contract, firm A expects to receive the highest reward of 4.5% of the total expenditure. Observing the choices of the contingent contract, the regulator can infer the firm's type. The Global Fund can modify the target to be a range of performance indicators, such as patient retention rate.

4.5. Two-part tariff

V-F mechanism, RPI-X price cap and sliding scales provide different methods to regulate price per output. However, if fund payment only depends on unit of service provided, even when per unit cost is minimized, there is a tendency to provide too much service [20]. For example, there will be short but more frequent cares, excessive but low quality hospital admissions, and distributing bed nets regardless of their intended use [21]. Therefore, the payment scheme the GF sets should not only be determined by units of output but also depends on types of patient treated. When we consider the efficient allocation of outputs among different consumers, we need a way to regulate both fixed payment to each patient treated and per unit of output payment.Two-part tariff scheme in regulatory literature addresses allocation efficiency.

Under the two-part tariff scheme, the payment schedule is made of two parts: an one-time, fixed fee per consumer and a flow charge per-unit of service. In the context of the GF, an example of two-part tariff payment system for individual patient i, disease treatment q (bed nets, hospital visits, etc.) at period t is:

$$f_{t,i}(q) = l_{t,i} + s * c_{t,i}(q), \tag{4.8}$$

where $l_{t,i}$ is the lump-sum payments from GF to PR for each patient treated. This example

[20] Ellis and Miller [2007] point out the problem of too many admissions in hospitals under a pure capitation payment system.

[21] Situations just as "insecticide-treated bed net used as soccer net in Wassini Island, Kenya" are described in the picture on page 45 of the report of *More Money for the Health* (Glassman et al. [2013]).

is comparable to equation (3.6), where $0 < s < 1$, and $c_{t,i}$ is a function of quantity q. Note, the difference between sliding-scale and two-part tariff is that the fixed payment $l_{t,i}$ is per-patient i in two-part tariffs (equation (4.8)), whereas in sliding-scale a is a fixed payment per service provision q (Equation (4.6)). If the GF disburses PR with respect to per unit of service provision, regardless of different patient types, two-part tariff payment system is the same as sliding-scale. In the particular case of Equation 3.8, the PR has the incentive to admit more patients. This payment schedule will expose the GF to the risk of excessive low quality admission of patients. [22] However, it reduces the tendency of over-supplying service.

Let us take a step back to the industrial organization literature. One of the advantages of the two-part tariff scheme is the convergence to a welfare-improved pricing in the long run. When the per unit price equals marginal cost, and the total amount of fixed fee per consumer equals the consumer surplus, the two-part tariff induces the firms to act in accordance with social optimal outcomes [23]. The outcomes are more efficient than Ramsey prices. In practice, the unit price is often regulated following the RPI-X regulation formula. This provides incentives for efficient utilization of existing capital. The lump-sum fixed fee equals an approximation of change of consumer surplus.

Two-part tariff is especially attractive for electricity transmission and distribution, and for telecommunication networks, where long run capital investment is necessary, and the demand for access and demand for usage are distinguishable. An example of a two-part tariff under the V-F constraint is:

$$[p_t, l_t | (q_{t-1}p_t + l_t N) \leq (q_{t-1}p_{t-1} + l_{t-1}N)(1 + RPI - X)], \tag{4.9}$$

where N is the number of consumers[24]. The revenue of the current period is restricted by last period's revenue adjusted by inflation and efficiency improvement. Here, the fixed fee reflects the steady long-run capital costs of infrastructure investment, while the incremental

[22] In health economics literature, risk-adjusted payments by diagnostic related grouping (DRG) have been widely used. It is a hybrid payment system combining provider (hospital), patient, and service characteristics. See Ellis et al. [1998] for a review.

[23] Two-part tariff is an application of the surplus subsidy schemes, such as the Incremental Surplus Subsidy (ISS) scheme suggested by Sappington and Sibley [1988]. They equate the current period of ISS_t to the sum of the change of consumer surplus from last period and the firm's last period profit. Under ISS, the firm's profit maximization problem is the same as maximizing the change of total surplus. Sappington and Sibley [1988] show that this mechanism converges to welfare-optimal prices.

[24] This example is from Vogelsang [2001]. Here, I demonstrate with the Laspeyres price index with a constant number of consumers. Vogelsang [2001] showed that an index with idealized weights derived by solving an explicit profit maximization problem furthers welfare improvements.

charge reflects the relatively volatile short-run operational costs. Under two-part tariff, the lump-sum fixed payment provides a stable income to meet the needs for long-run capital expansion.

When facing heterogeneous demand, a menu of two-part tariffs is offered to consumers to obtain the most consumer surplus. This is most attractive feature of the two-part tariff scheme. For the consumer group with higher willingness-to-pay[25], the combination of a higher fixed charge and a flat incremental price is more suitable. For example, the "family and friends" plan promoted by telecommunication companies offers a bundle of high subscription fee and low per-minute charge. Recipients with different demographic groups of patients will self-select into the tariff structure that is more suitable.

In the GF context, the disparities in access to essential service suggest the existence of different "demand" curves for different population living with HIV. The key population groups, defined as men who have sex with men, sex workers, people who inject drugs and transgender people, have larger barrier to access to treatment than other groups (For example: the anti-homosexual legislation in Nigeria and Uganda). Extending the treatment to key populations has the most value for money to the GF. A menu of two-part tariffs creates most surplus when facing heterogeneous willingness-to-pay. For population group with high surplus and low expected expenditure, the combination of high headcount payment and low payment per service provision should be applied. For instance, treating HIV key population has higher marginal benefits to the GF (for example: the prevention benefits); a two-part fund payment system for the key population would involve a relatively higher per patient admitted payment, and a relatively lower payment to each service provided. See Box 3 for a one-good example.

[25] One technique to measure willingness to pay is using the contingent valuation method (CVM). It is preferred by economists because of its rigorous theoretical basis. See Piper and Martin [2001] for an application for rural water supplies.

Box 3: A Simple Examples of Two-part Tariff

Case A: Within a PR, two distinguishable population categories

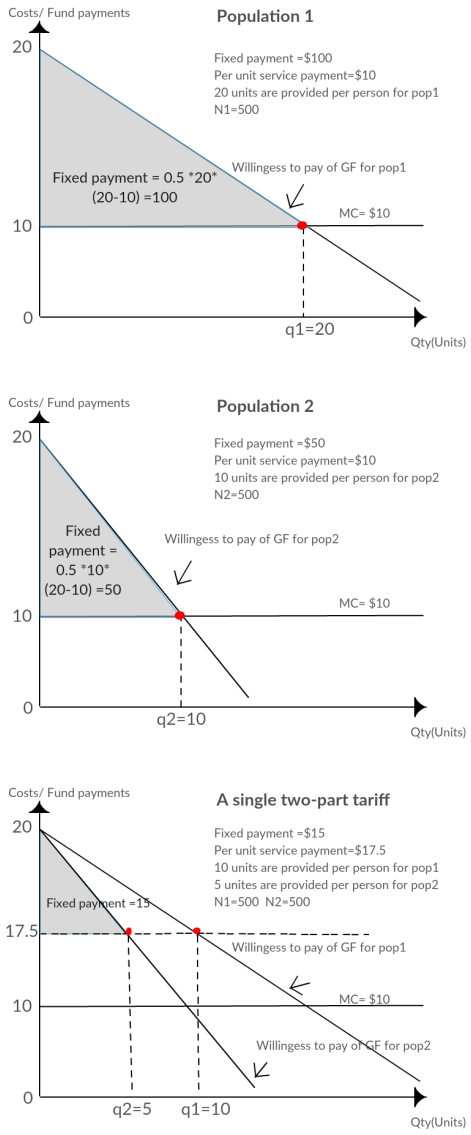
Assume GF knows the cost function of a given PR, and GF has estimated benefits of treating each population categories. Suppose the total cost is $50,000 + 10q$, marginal cost is \$10. The willingness-to-pay of population 1 (pop 1) is $f_1 = 20 - \frac{1}{2}q_1$, and the willingness-to-pay of population 2 (pop 2) is $f_2 = 20 - q_2$. Pop 1 has more value for money than pop 2. The estimated number of pop 1 is 500, and 500 for pop 2. The two-part tariff is expressed as $f_i(q) = l_i + t_i * q$, where $i = 1, 2.$, l_i is the fixed per person tariff, and t_i is the per unit payment. Now, suppose GF can distinguish the two population, and monitor PR perfectly, the social optimal two-part tariff schedule is (\$100,\$10) and (\$50,\$10) for population 1 and 2, respectively. That is a fixed payment of \$100 per patient admitted in pop1, \$50 per patient admitted in pop 2, and a \$10 payment per service provision. The total payment is \$225,000, total cost is \$155,000, the "consumer surplus" is \$75,000. See Figure 6 for an illustration.

Case B: within a PR, two non-distinguishable population categories

When GF cannot set two different payment schedules for the two population, the social optimal is set to maximize $500 * l + 500 * l + 500 * t * (40 - 2t) + 500 * t * (20 - t) - (50,000 + 10(500 * (40 - 2t) + 500 * (20 - t))$, where $l = \frac{(20-t)^2}{2}$. Solving the maximization problem, we can derive that the optimal two-part tariff schedule is (\$15,\$17.5). The total payment is \$146,250, total cost is \$125,000, and the "consumer surplus" is \$15,000.

It is worth noting that even though price discrimination is prohibited by regulators due to fairness considerations, non-discriminatory pricing creates dead weight loss. The dead weight loss comes from the consumer group with lower demand (or here, patients with higher expected risks) being excluded. Consider a one-product case with homogeneous demand – the fixed payment disbursed to each network (village/healthcare station) depends on the number of potential patient units that have access to the network. For geographic locations with less dense markets, a high fixed capitation payment is needed to for the lumpy long-run costs of essential infrastructures. When the willingness-to-pay of those consumers fails to match the long-run capital costs (or when the expected costs of those patients overwhelms

Figure 6: Examples of two-part tariff



the grants available), the profit-maximizing firm will discard the loss-maker markets. In such a situation, cross-subsidized price structures are enforced by welfare-maximizing regulators with motives, such as universal service. It takes the forms of subsidizing access by usage, cross-subsidization between geographic markets, and fund loss-marking services by profits from other services [26]. One simple example is the provision of rural bus routes at the expense

[26]Heterogeneous demand elasticities of different consumer groups and geographic differences in costs to sup-

of urban ones.

4.6. Practical issues

All the above suggested payment schemes require a certain level of information to grant the PR a fair surplus while rewarding cost reduction and other performance improvements. In practice, the GF faces many trade-offs when designing an optimal payment schedule. In this section, we describe three important trade-offs: cost-saving incentive and investment, long term and short term objectives, and delegation and flexibility.

4.6.1. Trade-offs

4.6.1..1 Trade-offs between cost-saving incentive and investment

As the existing infrastructure approaches the end of its life-cycle, and as new technologies emerge in the telecommunication industry, the investment issue become apparent. It has received extraordinary attention from policy makers. The trade-off between incentives and investments is one of the most obvious challenges faced by the regulators (Ai and Sappington [2002], Vogelsang [2010a]). Researchers and policy makers have spent a great deal of time studying the relationship between investment and the choice of tight, soft or intermediate regulatory regime [27] (Cambini and Jiang [2009], Vogelsang [2010b], Inderst and Peitz [2012]).

As long as the expected present value of cost savings exceeds the expected costs, a high incentive payment system encourages cost-reducing behaviors. However, the risk of future losses and the possibility of "truncated" profits by future payment adjustments lower the expected present value of investments. Prohibited from capturing all the profits from cost-saving, incentives cause sub-optimal investments. Ordinary investments [28] usually have a standard measure of expected costs and return, and suffer less under high power incentives. However, innovation is the game changer. Innovative investments are often very complex,

ply are the sources of cross-subsidization. As we discussed above, a complex dynamic price-cap regulation mechanism will push the price structure to the efficient Ramsey prices when cost reduction is the regulatory goal. However, regulators should be aware of harmful strategic behaviors. Regulated firms lower prices in some markets to deter competition. This predatory pricing in one market is funded by cross-subsidization from other services. Lack of competition narrows consumer's options, hinders innovation and leads to quality deterioration.

[27] Cambini and Rondi [2010] investigate the relationship between investment and regulatory regimes (incentive vs. rate-of-return regulation) for a sample of EU energy utilities from 1997 to 2007. In contrast to other studies, they show that investment rate is higher under incentive regulation than under rate of return regulation.

[28] The classification of "ordinary" and "innovative" investment is adopted from Vogelsang [2010b]. "Ordinary" investments are the ones with known costs and demands, such as legacy infrastructure maintenance and reconstruction. "Innovative" investments are the ones with highly uncertain costs and utilization.

unknown to the regulator, and very risky. From the GF-PR relation perspective, such investments can be interpreted as actions that cause high variation in costs, such as admitting sick patients who may incur non-reimbursable costs and adopting innovative interventions which may be ineffective.

There are two main aspects of incentive regulation that would deter innovation under certain conditions. One is truncated up-side return (Gans and King [2004], Vogelsang [2010b]). This is because the tight price constraint shifts the investment risks away from consumers to the regulated firm ex-ante. When the innovative project is very risky, and the expected present value of cost savings is not large enough, firms will not take up innovative investments [29]. Additionally, the regulator and the regulated firm have different valuations of the investment projects due to asymmetric information. Without fully understanding of industrial technology, regulation will distort efficiency. Regulators should give the firm responsibility for complex infrastructure investment decisions and innovations. The other issue is the lack of regulatory commitment. How can one prevent the regulator "crawling back" the excess profits of an investment by setting a low allowed price level after the capital expenditure is sunk? How can we ensure that the PR will not suffer from volatility in financial support due to commitment issues? We will discuss more on the ratchet effects and regulatory commitment issues below.

In practice, regulatory holiday has been adopted by policy makers around the world to promote investment incentives [30] and reduce regulatory uncertainty. For example, the contested German telecommunications law provides a regulatory holiday for innovative investment. Evidence has shown a positive relationship between innovative investments and regulatory holidays enforced by the judicial system (Gans et al. [2003]).

4.6.1..2 Trade-offs between multiple primary regulatory objectives

The regulators should internalize the trade-offs between different objectives before setting the incentives. Cost reduction incentives, output delivery incentives, and target performance

[29] It is important to notice the substantial heterogeneity in costs and benefits of different innovations. Darius Lakdawalla and Reif [2015] point out that "some innovations, like using aspirin to treat acute myocardial infraction, cost almost nothing while others", like Vertebroplasty, a expensive procedure that injects cement into the spine to stabilize vertebrae, has been shown to be ineffective.Some fall in-between the two extremes. For example, the inventions, like the antiretrovirals for HIV/AIDS, are costly but highly effective.

[30] When competition between incumbent and entrants is considered, one should note that regulation of the access prices in telecommunication actually promotes competition, including innovation competitions between the market participants.

incentives interact with each other. While some of the primary objectives may be complementary, the delivery of other objectives may require the regulated firm to make trade-offs. This may induce unintended consequences. When setting incentives, regulators should take account of the relationships between multiple regulatory objectives.

The first trade-off is the integration between efficiency concerns and redistribution concerns. The regulator's ranking of those concerns determines the regulatory design. Efficiency improvement at the expense of exclusion of a group of consumers has a negative impact from the equality perspective (Steinberg [1986], Crampes and Estache [1998]).

Another trade-off is the conflict between primary objective (short-term) and secondary (long-term) objectives. If the incentives only focus on the delivery of primary outputs, the firms are encouraged to achieve those performance targets at the lowest cost, potentially at the expense of measures that could help reduce the cost of delivering primary performance targets in the long term (OFGEM [2010]). Those measures are typically projects that improve efficiencies in a longer term with actions taken today. Or, the projects may require large upfront costs and have potential, but with uncertainty, to deliver benefits in the long term.

Lastly, regulators should be be aware of the inconsistency of firms' objectives across different institutions and hierarchies. The state-owned monopolies usually have incentives other than profit maximization, such as political orientation. For instance, they allocate service branches based on employment subsidy decisions other than profit maximization.

4.6.1..3 Trade-offs between delegation and flexibility

Monitoring and inspection are important dimensions of price regulation. However, the enforcement of monitoring in practice is very difficult. Moreover, if managerial effort is non-contractible, delegation offers little benefits, as the regulator spends time and resources monitoring the managerial personnel of local authorities (Akcigit et al. [2014]). The scale and scope of the monopoly indicate the need for delegation of supervision. In addition, delegating inspections to local authorities can increase the flexibility of the regulation and employ local level information. However, such discretion brings managerial conflicts. It gives room for regulation capture and corruption. The collusion between the regulated firm and local authorities can hamper the quality of accurate accounting and monitoring. Duflo et al. [2013] investigates the effect of higher inspection rates on pollution reduction. They conducted a two-year experiment in India, and found no significant impact on pollution reduction, even

though both the actual and perceived inspection rates were higher for the treatment group. Another form of corruption is collusion with local regulatory staff. Collusion between the contractor and government employees can also condemn the quality of accurate accounting and monitoring. Crampes and Estache [1998] suggest that high enough salaries should be given to the employees to minimize the incentive to accept the side payments.

Table 4: Comparison of trade-offs and unintended consequences

	Cost reduction incentives	Investment incentives	Uncertainty	Administrative burden	Strategic behaviors on other dimensions		
					Gaming over time	Multitasking issue	Quality Shading
Rent extraction by PR	+	+					+
Regulatory lag	+	+	+	-	+		
Cost sharing by PR	+		+				+
Finer payment structure			-	+		+	-
Flexibility			-	+			

Source: Author's construction.

Table 4 highlights some trade-offs GF should consider when design the fund payment contact. The key tradeoff is the tradeoff between rent extraction and cost saving. The higher allowed revenue residual PR can retain, the more cost saving incentives the payment system provides. Higher revenue residual also encourages investment that facilities cost savings, but might at the expense of service quality. How frequently should the GF adjust the payments? Longer price control period exposes both GF and PR under uncertainty, but facilitates long term investments. Cost sharing can be realized through revenue sharing, and other correction factors of the fixed payment level. The more costs are shared by PR, the stronger the cost saving incentive is. How fine should the price structure be? The finer the payment structure is, the less risk PR bears. However, it makes monitoring and accounting more challenge. And, it could cause multitasking issues. If the fund payment only has one target. For example, it only rewards enrollment quantity. In this case, PR allocates effort towards achieving the target number of patients at the expense of low numbers of hospital revisits, which are essential to TB treatment. Lastly, how flexible should the payment system be? Ex-ante spec-

ified correction terms and performance targets reduce uncertainties, yet the design of the terms requires great administrative efforts.

4.6.2. Unintended consequences and strategic behaviors

One of the key challenges of price regulation is to deter gaming of the incentive regulation mechanism. Because a firm operates in a multidimensional strategy space, it is nearly impossible to achieve the desired outcome by implementing a simple one-dimensional pricing strategy. It is worth noting that sometimes incentives may backfire. There are three main unintended consequences under incentive-based price regulation: quality shading, ratchet effect and market discrimination[31]. We will elaborate on each of them, and discuss the possible solutions to prevent such unintended behaviors.

4.6.2..1 Quality shading

Examples

Quality shading is the most obvious unintended consequence. In response to rewards for cost reduction, instead of improving efficiency, the PR achieves cost reduction by compromising quality of service. It exhibits unfavorable outcomes such as inadequate infrastructure, lack of managerial effort towards service improvement and poor product quality.

In the electricity distribution sector, reliability is one of the most important measures of service quality. The electricity outages of California in 2000 and 2001, and the northeast blackouts of 2003 are examples of quality shading as a consequence of inadequate incentives for maintaining quality of service. Ter-Martirosyan and Kwoka [2010] measure service outage as an indicator of poor service quality for the US electric power industry. They suggest that incentive regulation is indeed associated with significantly longer service outages. Other studies on the UK electricity distribution sector have suggested a positive association between quality shading and incentive regulation (Giannakis et al. [2005]). However, because quality shading also reduces demand, the degree of quality deterioration is restricted by the loss of total revenue. When the demand for a product or service is elastic or when there are certain levels of competition, the positive connection between quality shading and incentive pricing is not necessary. Vogelsang [2002] suggested that there are little empirical evidence for quality shading induced by incentive regulation within the telecommunication industry.

[31] Market selection and "cream skimming" will be discussed in 4.1.2.

In the Global Fund context, grant payment systems with strong cost reduction incentives will encourage recipients to reduce service quality. For example, if the grant payment unit is a fixed amount based on population served by the recipient, then such payment will induce too many admissions of low-risk patients, short but more frequent care, low service intensity, and in general low quality services.

Box 4: Example from Southern California Edison Company

"The Southern California Edison performance based regulation incorporates service quality performance incentives into the RPI-X price cap formula. The service quality incentives comprise service reliability, customer satisfaction, and employee health and safety. Service reliability comprises outage duration and outage frequency benchmarks. For example, the outage duration benchmark is 59 minutes average customer outage in 1997. The reward or penalty is $1 million per minute for two year averages with an upper limit of $18 million for outage and frequency. Similarly, the frequency benchmark is 10,900 annual interruptions with $1 million reward or penalty for each 183 interruptions. Customer satisfaction is measured by independent surveys and the satisfaction benchmark is 64%. The main reward and penalty is $2 million for each percentage below or above a 3% dead-band. The employee health and safety benchmark in brief is a ratio index of the number of accidents and illnesses. The benchmark value is 13 with a dead-band of 0.3 and $555,000 reward or penalty for 0.1 deviation increments beyond the band."

**This example is quoted directly from Jamasb and Pollitt [2001]. This quality shading action was in fact declared to be fraudulent.*

Solutions

Subsequent to a disastrous experience, quality shading has been brought to the attention of the regulators of the utility sector. As we mentioned before, both RPI-X and benchmarking regulation have variations forms that link quality to performance measurement and price resetting. In the electricity distribution sector, policies have been designed to include incentives for maintaining and enhancing service quality. Actions range from quality monitoring to setting quality benchmarks. In the U.S. electric power sector, more than 20 states set up explicit penalties for performance difference relative to the established quality standards

(Ter-Martirosyan and Kwoka [2010]).

Benchmarking of service quality is mostly done in terms of setting minimum standards. In electricity distribution, service quality incentives usually comprise three main dimensions: service reliability, customer satisfaction and employee health and safety. The minimum standards for service reliability are measured by frequency and duration of outages. One important aspect of quality benchmarking is to link the quality performance with financial rewards and penalties. For example, in Brazil, Colombia and Chile, the service quality standards are applied together with non-attainment penalties or rewards. The performance standards also vary under different exogenous events. For example, the OFGEM sets performance thresholds for normal and extreme weather conditions.

From the perspective of the Global Fund, an adaptive healthcare quality standard framework has to be established. For example, a fixed payment system with incremental rewards for each unit of service that achieved the targeted quality standard would mitigate the quality shading problem [32].

4.6.2..2 Gaming over time and ratchet effect

Gaming over time by the regulated firms

As we have discussed above, the construction of a price cap involves elements of rate of return regulation, benchmarking and profit sharing. Plus, the price resetting at each price review period to reflect the realized costs creates a ratchet. The "ratchet effect" refers to a situation where regulator uses private information revealed by firms' early actions to the firms' disadvantage later. The efficiency improvements achieved during the price control period are punished by tighter incentives in the next period. Anticipating this, firms may choose to refrain from efficiency improvement to avoid facing more stringent constraints. A distinct strategic behavior of the regulated firm is to manipulate the timing of cost reduction. Evidence has confirmed that firm's effort to reduce costs decreases as the review approaches (Sappington and Stiglitz [1987], Freixas et al. [1985], Baker et al. [2002])

The revenue the PR is allowed to collect from donors in the next price control period depends on the following: the actual audited expenditures during the price control period, the forecast expenditures during the price control, and the incentive rates applied to different

[32] There is a large literature on the prospective payment system in healthcare. It includes fee-for-service, capitation payments and other form of payments with high cost reduction incentives. (Ellis and McGuire [1986], Ellis and Miller [2007])

specific categories of expenditure. Following are some typical manipulations of expenditures over time:

- **Shifting costs within the price control period** The strategic shifting of costs allows the PR to make a large profit in the first half of the price control, and to pretend to be inefficient by over-spending when a price review approaches. The ratchet leads to differential incentives for cost reduction depending on how close the firm is to the next price review.

- **Multitasking problems** When the incentives are not symmetric among different tasks, multitasking problems emerge. Under RORR, there exist imbalanced incentives applicable to operating and capital costs. We observe that a firm is encouraged to capitalize part of the operating costs, in order to reach the target of operating costs reduction. Similarly, if the fund payment only rewards by number of patient enrolled, PR might allocate effort towards achieving the target number of patients enrolled at the expense of hospital revisits. A performance target with broader guidance would mitigate multitasking problems (Olken et al. [2014]).

- **Influencing the benchmark:** information on past costs is not required within the price cap period, but is likely to influence the resetting of X. PR may submit a high forecast at an early stage to influence the baseline benchmarks, then lower the forecast later when choosing the incentive scheme. For example, PR could submit inflated capital expenditure forecasts at the negotiation period to influence the level of available fund payments. PR earns high returns later during the contract period by underspending those forecasts. Regulators of monopolistic firms have implemented a "use it or leave it" mechanism to monitor and encourage firms to submit accurate expenditure forecasts[33].

What is the optimal length of a payment control period? The optimal length of a regulatory control period reflects a trade-off between a reduced incentive to engage in irreversible investments and to control costs resulting from more frequent payment adjustment, and benefits from reduced "ratchet effect" and faster convergence to efficiency. Another trade-off

[33] "use it or leave it" mechanism sets aside allowed revenue for a specific capital investment. If the revenue is not used for this project as intended, the allowance will be removed.

is between the challenges of accounting, monitoring and administering at each period, and the benefit of fewer uncertainties regarding the regulatory environment during a short regulatory lag. Currently, the funding cycle of the GF is one year, as a component of the budgeting process. Keeping the annual funding cycle, GF could improve cost reduction incentives by an informal/formal promise of continuing disbursement in the following year, based on PR's performance.

4.7. Summary

When designing the optimal pricing mechanism, regulators have to take account of the trade-offs and potential unintended consequences. Table 5 summarizes the advantages and limitations of some of the discussed price regulations. It also highlights the cost reduction incentives and data requirements of those approaches. Incentive-based optimal pricing needs to strike a balance between reflecting the main performance drivers and reducing incentives for engaging in unproductive methods or incentive-induced strategic behaviors. Low-powered incentives are desirable in situations such as noisy performance measurement, multitasking, collusion with monitors or regulatory capture and repeated interactions. In addition, regulators should promise to not make retrospective adjustments to revenue in the event that costs turn out to be different. For example, turnover in administration personnel harms the commitment to the policies enforced by the former regulator.

Besides the power of incentives, another dimension of payment system design that is worth noting is the choice of tight, intermediate and soft regulations. The softness and tightness of a payment system is related to the PR's participation constrains. It is still possible to have a high power incentive system under soft regulation. Under tight regulation, high power incentive payment systems are used, frontier benchmarking is adopted, and the sliding scale leaves no excess profits to the firm. Under soft/intermediate regulation, the permissible payment level is higher than the long-run average incremental costs, the fixed payment caps are targeted at average benchmarks with a longer control periods, and more cost sharing is allowed under the sliding scale regulation. Overall, soft regulation is more suitable at the early stage of incentive regulation, when the regulator has little experience based on cautious considerations Vogelsang [2010b]. As the regulator reduces its information disadvantage through systematic accounting systems and performance monitoring programs, tighter regulations can be implemented.

Table 5: Comparison of different price regulations

	Cost reimbursement	V-F mechanism	RPI-X cap	Two-part tariff	Sliding scale	Yardstick competition
Advantage	Simplicity	Allocative efficiency	Flexibility	Additional instrument	Bayesian approach	Relative performance
Cost reduction incentives	No	Strong	Strong	Depends on per unit payment design	Depends on sharing rule	Strong (depends on benchmark)
Limitation	No cost reduction	Unrealistic assumpions	Complexity	PR retains large profits	Depends on PR's ability to understand the sharing rule	Authority, sensitive to outliers
Data requirement	Auditing	Bookkeeping data (BD)	BD+ Estimating X factor, inflation rate.	BD+ Estimating consumer surplus	BD+ Prior knowledge of costs distribution	BD+ Industrial or district level data
Examples	Complex construction projects, early stage of defends contracts	UK electricity transmission and telecommunication sectors. Public transportation uses subsidy cap in Italy.		Brazil electricity market, physician payment system, water supply sector	Sliding scale matrix for capital expenditure allowance for OFGEM	Olken et al. [2014]

Source: Author's construction.

5. Put incentives into contract: experience from government contracting

Contracting has been widely used by government agencies to procure a broad range of services, including public transportation, water distribution, sewage and garbage service, public defense, ancillary service in public education system and employment counseling. Government outsourcing public services offers consumers a choice of private-sector providers. The public-private partnership (PPP) [34] allows governments to take advantage of skilled specialization of the private sector. It also gives the governments a vehicle to avoid large and repetitive asset investments. Moreover, it entitles the governments the benefit from economies of scale regardless of the size of the government entity (Baker et al. [2002]). On average, above

[34] PPP refers to a variety of co-operative arrangements between the government and private sector. It is a method of involving the private sector in delivering public goods or services and/or securing the use of assets necessary to deliver public services. Partnerships also provide a vehicle for coordinating with non-governmental actor to undertake integrated, comprehensive efforts to meet community needs Poutvaara [2014]". Compare to other traditional procurement contracts, PPP contract has a longer contract period, and provides a bundled services/products to the government with private financing.

40 percent of the total costs of public goods are outsourced in the OECD countries (OECD [2013]). The public-private partnership is not different from the GF-PR relationship.

For standard services such as sewage and garbage service with fewer problems related to uncertainty and asymmetric information, outsourcing generally reduces costs without hurting service quality. However, when the service is more complex, a sophisticated contract is required to ensure optimal performance. Incentive contracting for public service delivery (government procurement) has been studied broadly in the literature (Grout and Stevens [2003], Vining and Boardman [2008], Auriol and Picard [2009], Poutvaara [2014]). Similar to optimal incentive regulation, the efficacy of incentive contracting is challenged by the incomplete and asymmetric information and uncertainties in the course of the contract.

The establishment of an initial contact usually comes in the form of negotiation, either formal or informal, over the contract clauses to reach a mutually agreed contract. Since unpredictable and incalculable events always occur, ex-post adaptation is inevitable. Bajari and Tadelis [2001] point out that in practice, little is known by either the government or the contractor at the onset of a procurement project. In addition, the existence of informal relational contracts makes complete written contracts impossible. Also, the effort exerted for cost reduction is not observed by the government, thus not contractable. Moreover, the cost of monitoring performance and getting contract adjustments enforced is very high. Overall, uncertainty induces moral hazard problem, intensifies lack of commitment, and complicates the contract contingencies. In the following section, we will briefly discuss the discuss two types of contracts, the fixed-price contact and cost-reimbursement contract, and introduce the experience of government contracting with for-profit and non-profit monopolies, respectively.

5.1. Fixed-price contracts and cost-reimbursement contracts

In practice, a vast majority of government contracts are variants based on simple fixed-price and cost-reimbursement contracts. Fixed-price contracts provide the contractor an adjustable price. It could be a price ceiling or a targeted revenue. Similar to a pure price cap, the fixed-price contract features a high level of contract design, strong cost-reduction incentives, yet very high renegotiation costs. For example, a fixed budget system for the PR is a type of fixed-price contract. It is designed based on the PR's characteristics, regardless of how many patients it treats or what services it provides. Such fixed budget system will motivate the

contractor to preferentially avoid high cost projects/customers/service, and under-provide quality. A fixed-price contract with finer classifications would mitigate the selection issue. However, a trade-off between the potentially improved quality and fairness and the challenges of monitoring and administering a fine system should be considered (Ellis and Miller [2007]). [35]

One variation of the pure fixed-price contract is a fixed-contract incentive contract. It provides flexibility for adjusting payments and establishing the fixed contract payment by a predetermined formula based on the deviation of negotiated total costs from the target costs.

Cost-reimbursement contracts provide the contractor with payment of allowable incurred costs, to the extent prescribed in the contract. The analogy of cost-reimbursement contract in optimal pricing is the rate-of-return regulation. Those contracts establish an estimation of the costs needed to deliver the service. One variation of the pure cost-reimbursement contract is the cost-plus-incentive-fee contract. According to the performance of the contractor, the reimbursement is adjusted upwards if the contractor delivers the service with less costs than the target costs, and downwards when total allowable costs exceed target costs. Compared to fixed-price contracts, cost-reimbursement contracts feature low levels of design and weaker incentives, and are more suitable when the project is complex. Bajari and Tadelis [2001] studied the choice between fixed-price and cost-reimbursement contracts under uncertainty. They suggest cost-reimbursement is preferred when the project is more complex and when the likelihood of future renegotiation is higher. Crocker and Reynolds [1993] provide evidence in defense procurement. They find that Air Force engine procurement contracts are mostly based on cost-reimbursement contracts at the initial stage when inevitable yet unanticipated adjustments are expected. At the later stages, when there are less uncertainties, fixed-price contracts are used.

5.2. Experience from outsourcing public education service, Job placement service and healthcare service

One area in which government outsourcing is particularly popular is in the delivery of public education. Historically, education outsourcing [36] is most common for the auxiliary services,

[35] Ellis and Miller [2007] provide a thorough discussion of different provider payment systems in the healthcare industry. The realizations of fixed-price contracts in provider payment design is are pure capitation system, a fixed budget/salary payment system and fee-for-service payment system. See 3.1.3 for a brief discussion.

[36] The experience with government outsource is not limited to contracting with for-profit providers. Especially in the health sector, nonprofit hospitals are very common partners to serve the healthcare needs of the public.

such as catering and school transportation. Recently, a broader scope of contracting was undertaken in the education sector: private provision of core education services, such as school voucher programs, charter schools, providing online public education.

Depending on each service, the contracting considerations are different. Auxiliary service (management, maintenance and support services) contracts usually take the form of fixed-price contracts. Operation services, such as infrastructure projects and school management, are more complex. Contract types such as build-own-transfer, build-own-operate, and performance are implemented[37]. For example, additional subsidies (payments) are granted to the school that demonstrates the highest test scores. Innovative contracting, such as a voucher system, introduces competition into school choice, and leads to more efficient and high quality eduction[38].

Federally sponsored employment and training programs are enforced through contracts with performance incentives. Such contracts reward/penalize the service provider according to the size of any incremental variation from the performance level used to ensure performance [39]. However, sometimes such incentives result in unintended consequences[40]. Marschke and Courty [2002], in a study of job training agencies, showed that public organizations can engage in gaming by timing their performance reports in order to benefit from awards. They show that performance incentives can come at a cost by having a negative effect on efficiency. Moreover, sometime strategic responses such as cream skimming are induced under performance incentives (Behaghel et al. [2014]).

Box 5.1: Cream-skimming in the school voucher system in Chile

Hsieh and Urquiola [2006] studied the universal school voucher system implemented in Chile since the 1980s. The voucher provided students the opportunity to attend private schools. However, they find that voucher increases the chance of sorting. The academic advantages of the private schools are associated with their ability to select the most able students and those with a greater ability to pay (cream-skimming).

[37] See Patrinos and Sosale [2007] for a review
[38] See Patrinos et al. [2009] for more on voucher systems. For unintended consequence see Box 5.1.
[39] See Barnow and Smith [2004] for a review on performance management of U.S. job training programs.
[40] See Box 5.2 for an example of strategic behaviors.

Box 5.2: Cream-skimming and parking in job counseling service provision

Behaghel et al. [2014] compared the effect of public and private provision of job counseling services. Exploring a large-scale randomized controlled experiment in France in 2007, they provide empirical evidence for two strategic behaviors of private job placement contractors under different incentive contracts: cream-skimming and parking. The public intensive program paid the social workers 657 euros per client, while the private programs on average received a minimum payment of 900 euros per client on enrollment, and a maximum of 3947 euros per client, conditional on the client being durably placed in a job. Behaghel et al. [2014] estimate that the public program increases exit to employment by 10.2 % after 6 months, whereas the private program impact is 4.5 %. Further, the public program increases the transitions to employment by 11% in the first 3 months, whereas the impact is yet insignificant in the private program.

They also find that when the conditional payment is relatively large (strong incentive), private programs maximize profits by selecting to enroll the job seekers with the best labor market prospects (cream-skimming); but if the the payment schedule is relatively flat (weak incentive), they may enroll any job seeker and just offer a bare minimum of services (parking).

Box 5.3: Creaming, skimping and dumping in healthcare service provision

Ellis [1998] suggested three strategic behaviors of healthcare providers under different incentive contracts. He points out that the design of reimbursement incentives will influence both the intensity of services and who is treated when patients differ in their severity of illness. He argues that cost-reimbursement contracts tend to induce overprovision of service to all types of patients. Fully prospective payment (fixed-price contract) entices healthcare providers to provide too much service to low severity (cost) patients (creaming) , and too little service to high severity (cost) patients (skimping). If the provider's participation constraint is binding under the prospective payment contract, the most severely ill patients will be avoided (dumping).

In practice, the perspective payment system is in the form of a risk-adjusted capitation payment system. Hennig-Schmidt et al. [2011] analyze the influence of incentives on physicians' supply of medical services by running a controlled laboratory experiment. They confirm the theory predictions of Ellis [1998]. They find that patients are over-served under fee-for-service and under-served under capitation.

5.3. Practical issues

The unintended consequences and trade-offs discussed in the section 4.6 also apply in the context of contract design. Here, I will focus more on issues based on limited commitment and informal contracts.

5.3.1. Dynamic contacts and costly renegotiation

Similar to the ratchet effect, the dynamic property also raises the possibility of renegotiation. Mutually-agreed renegotiation may improve social welfare, but it is usually very costly. It also impedes ongoing investments and weakens efficiency incentives.

If the global fund were to force a renegotiation, this action could be perceived as evidence of regulator non-commitment or of the "rachet effect". One way to prevent regulator non-commitment and non-mutual renegotiation is through judicious design. For example, compared to the European practice in utility networks regulation, the United States has a much longer history of implementing RORR. This is because the Supreme Court protects the firms against loss of fair return. The legal process prevents unnecessary adjustments. The rate of return resetting has to go through a series of Supreme Court cases. With the protection of the judicial system, regulated firms face a lower possibility of future price renegotiation. Another method to reduce renegotiation costs is to design an uncertainty mechanism. The uncertainty mechanism makes amendments to the contract contingent on prescribed explicit trigger events, removing the risks of both parties. It avoids undue complexity and provides transparency. Moreover, such a formal ex-post adaptation mechanism limits the opportunities for gaming. However, it is important that the trigger events cannot be influenced by the contractor.

When the renegotiation is forced by the contractor, it is phrased as a hold-up problem. A hold-up problem occurs when the contractor has more bargaining power ex-post. When the

unforeseeable event occurs, the contractor usually has private information about the costs of renegotiation. This happens mainly due to the following reasons: 1) the existing infrastructure investment is too large to back out; 2) the search for another appropriate contractor is time consuming and costly; 3) the contractor knows the externality of the public goods to the government.

5.3.2. Relational incentives and informal contracts

Sometimes the best measurement of performance is subjective and non-verifiable. No formal written contract can be enforced to ensure the desired performance. Informal contracts with implicit incentives can enforce dimensions that are non-contractable. A relational contract penalizes the contractor by removing future contracting opportunities. When social optimal outcomes cannot be achieved by contractual means, self-enforced informal contracts can improve the outcomes. Halac [2012] pointed out, when the contracting relationship is repeated and open-ended, informal relational contracts complement the written formal contracts. Those relational contracts are mainly sustained by the value of future relationships. For example, a procurement contact may include explicit penalties for late delivery. Meanwhile, the prospect of future business cooperation provides implicit incentives for performance, and deters moral hazard behaviors and commitment problems such as hold-ups.

The self-enforcing property of informal contracts is crucial. The content of the informal contracts is influenced by social norm, culture and trust (Banerjee et al. [2006], Levin [2003]). When the benefit of deviating from action in the implicit contract exceeds the present value of the future relationships, the informal relational contract is no longer self-enforcing (Gil and Marion [2012]). Halac [2012] studied the situation when the value of future relationships is not common knowledge, and discussed the source of inefficiency depending on the allocation of bargaining power. In particular, when the government has private information about her outside option, she will choose the power of incentive based on her bargaining power. If the contractor has stronger bargaining power, the government tends to overstate the value to capture a larger share of the surplus. If the government has stronger bargaining power, she may understate the value of the relationship.

5.4. The differences of nonprofit monopolies

Nonprofit companies play a prominent role in delivering public services. Areas where nonprofits are particularly active are the delivery of service for the civilian services, such as elder homes, nursing homes, and institutions for the disabled. Compared to for-profit monopolies, the nonprofits are different in the following attributes[41] :

- **Monitoring and accounting:** The nature of nonprofit companies' funding process and government lowers the risks of moral hazard. They are usually monitored by a combination of donors, clients and government officials (Young [2006]). However, the accountability demands of multiple monitors also raise challenges (Ebrahim [2010]). Additionally, the service provided by nonprofits usually has intangible outcomes. The measurement is difficult and subjective.

- **Objectives:** Nonprofits have similar maximizing objectives as the government, such as universal service (Steinberg and Weisbrod [2005]), redistribution of income (Frank and Salkever [1994], Norton and Staiger [1994]), maximizing employee income and improving service quality (Rose-Ackerman [1996]). The alignment of objectives deters opportunistic behaviors.

- **Incentives**: Employees at nonprofit entities are more responsive to non-financial incentives. Monetary or material incentives often do not work well. Sometimes, they even have negative impacts on intrinsic motivations, therefore performances (Gneezy et al. [2011]). These crowding-out effects are most acute in prosocial behaviors in the education and healthcare sectors (Frank and Salkever [1994], Dur and Tichem [2013]). For example, when a worker is more motivated by the perceived social image of doing good deeds, a financial reward visible to the public is detrimental to performance.

- **Altruistic preferences:** Employees at nonprofit organizations have strong altruism preference (Rose-Ackerman [1996]). However, altruistic workers have a tendency to serve the most disadvantaged, even at the expense of efficiency. Heckman et al. [1997] cited evidence from a job training program that the most disadvantaged are served more even though they get less out of the program.

[41] See Steinberg and Weisbrod [2005] and Gneezy et al. [2011] for detailed reviews on nonprofit entities.

6. Conclusion

There are strong analogies between the payment system design challenges of monopolies, such as network industries, and the grantmaking problems of the Global Fund. Existing practices that guide price regulation and contract payment design can be helpful for GF's innovative design of new grantmaking models. In practice, the mechanism with optimal incentives is a combination of elements of different incentive-based regulations. The most widely adopted RPI-X regulation contains elements of RORR, benchmarking, profit-sharing and other targeted incentives. Innovative realizations of the optimal incentive regulation consider many dimensions of performance other than cost-saving. They usually include quality measures, environmental standards, technical efficiency and other incentive packages in the pricing mechanism.

When designing a results-based price and contracting mechanism with the correct incentives, the Working Group should fully analyze the tradeoffs between cost-saving incentives and rent extraction, risk sharing, regulatory commitment and investments. Cost controls and efficiency improvements are more appropriate for standard tasks with easier to measure outputs, but not for complex tasks with high uncertainties. Unintended consequences, such as quality shading, gaming over markets and time, regulatory capture and hold-ups, should be included in the framework to provide an integrated healthcare service with more aligned incentives.

This paper aims to offer the Working Group and other practitioners a compact and unified framework to think about incentive designs in the global health sector. Development of a results-based agreement of grantmaking, specifically tailored to address the practical financing problems the Global Fund faces, must be left for future research. The ultimate goal is to strengthen the GF's ability to maximize the impact of the financial support it provides to the PR, and achieve the goal of accelerating the end of AIDS, TB and malaria as epidemics.

References

Chunrong Ai and David EM Sappington. The impact of state incentive regulation on the us telecommunications industry. *Journal of Regulatory Economics*, 22(2):133–160, 2002.

Ufuk Akcigit, Harun Alp, and Michael Peters. Lack of selection and limits to delegation: Firm dynamics in developing countries. 2014.

Emmanuelle Auriol and Pierre M Picard. Government outsourcing: Public contracting with private monopoly*. *The Economic Journal*, 119(540):1464–1493, 2009.

Patrick Bajari and Steven Tadelis. Incentives versus transaction costs: A theory of procurement contracts. *RAND Journal of Economics*, pages 387–407, 2001.

George Baker, Robert Gibbons, and Kevin J Murphy. Relational contracts and the theory of the firm. *Quarterly journal of economics*, pages 39–84, 2002.

Sudeshna Ghosh Banerjee, Jennifer M Oetzel, and Rupa Ranganathan. Private provision of infrastructure in emerging markets: do institutions matter? *Development Policy Review*, 24 (2):175–202, 2006.

Burt Barnow and Jeffery Smith. Performance management of us job training programs: Lessons from the job training partnership act. *Public Finance and Management*, 4(3):247–87, 2004.

David P Baron and Roger B Myerson. Regulating a monopolist with unknown costs. *Econometrica: Journal of the Econometric Society*, pages 911–930, 1982.

Luc Behaghel, Bruno Crépon, and Marc Gurgand. Private and public provision of counseling to job seekers: Evidence from a large controlled experiment. *American Economic Journal: Applied Economics*, 6(4):142–174, 2014.

Carlo Cambini and Yanyan Jiang. Broadband investment and regulation: A literature review. *Telecommunications Policy*, 33(10):559–574, 2009.

Carlo Cambini and Laura Rondi. Incentive regulation and investment: evidence from european energy utilities. *Journal of regulatory economics*, 38(1):1–26, 2010.

Claude Crampes and Antonio Estache. Regulatory trade-offs in the design of concession contracts. *Utilities Policy*, 7(1):1–13, 1998.

Keith J Crocker and Kenneth J Reynolds. The efficiency of incomplete contracts: an empirical analysis of air force engine procurement. *The RAND journal of economics*, pages 126–146, 1993.

Anup Malani Darius Lakdawalla and Julian Reif. The complex relationship between health care reform and innovation. page Chapter 12, 2015.

Esther Duflo, Michael Greenstone, Rohini Pande, and Nicholas Ryan. What does reputation buy? differentiation in a market for third-party auditors. *The American Economic Review*, 103(3):314–319, 2013.

Robert Dur and Jan Tichem. Altruism and relational incentives in the workplace. 2013.

Alnoor Ebrahim. The many faces of nonprofit accountability. *The Jossey-Bass Handbook of Nonprofit Leadership and Management,*, pages 101–123, 2010.

Randall P Ellis. Creaming, skimping and dumping: provider competition on the intensive and extensive margins. *Journal of health economics*, 17(5):537–555, 1998.

Randall P Ellis and Thomas G McGuire. Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of health economics*, 5(2):129–151, 1986.

Randall P Ellis and Michelle McKinnon Miller. Provider payment methods and incentives. 2007.

Randall P Ellis, Gregory C Pope, Lisa I Iezzoni, John Z Ayanian, David W Bates, Helen Burstin, and Arlene S Ash. Diagnosis-based risk adjustment for medicare capitation payments. 1998.

Richard G Frank and David S Salkever. Nonprofit organization in the health sector. *The Journal of Economic Perspectives*, pages 129–144, 1994.

Xavier Freixas, Roger Guesnerie, and Jean Tirole. Planning under incomplete information and the ratchet effect. *The review of economic studies*, 52(2):173–191, 1985.

Joshua Gans, Stephen King, et al. Access holidays for network infrastructure investment. *Agenda*, 10(2):163–178, 2003.

Joshua S Gans and Stephen P King. Access holidays and the timing of infrastructure investment. *Economic Record*, 80(248):89–100, 2004.

Michael Geruso and Thomas G. McGuire. Tradeoffs in the Design of Health Plan Payment Systems: Fit, Power and Balance. NBER Working Papers 20359, National Bureau of Economic Research, Inc, July 2014. URL http://ideas.repec.org/p/nbr/nberwo/20359.html.

Dimitrios Giannakis, Tooraj Jamasb, and Michael Pollitt. Benchmarking and incentive regulation of quality of service: an application to the uk electricity distribution networks. *Energy Policy*, 33(17):2256–2271, 2005.

Ricard Gil and Justin Marion. Self-enforcing agreements and relational contracting: evidence from california highway procurement. *Journal of Law, Economics, and Organization*, page ewr026, 2012.

Amanda Glassman, Victoria Fan, Mead Over, R Silverman, K McQueston, and D Duran. More health for the money. putting incentives to work for the global fund and its partners. 2013.

Uri Gneezy, Stephan Meier, and Pedro Rey-Biel. When and why incentives (don't) work to modify behavior. *The Journal of Economic Perspectives*, pages 191–209, 2011.

Paul A Grout and Margaret Stevens. Financing and managing public services: An assessment. 2003.

Marina Halac. Relational contracts and the value of relationships. *The American Economic Review*, 102(2):750–779, 2012.

James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654, 1997.

Heike Hennig-Schmidt, Reinhard Selten, and Daniel Wiesen. How payment systems affect physicians provision behaviouran experimental investigation. *Journal of Health Economics*, 30(4):637–646, 2011.

Chang-Tai Hsieh and Miguel Urquiola. The effects of generalized school choice on achievement and stratification: Evidence from chile's voucher program. *Journal of public Economics*, 90(8):1477–1503, 2006.

Roman Inderst and Martin Peitz. Network investment, access and competition. *Telecommunications Policy*, 36(5):407–418, 2012. URL http://ideas.repec.org/a/eee/telpol/v36y2012i5p407-418.html.

Tooraj Jamasb and Michael Pollitt. Benchmarking and regulation of electricity transmission and distribution utilities: Lessons from international experience. Technical report, Faculty of Economics, University of Cambridge, 2001.

Paul L Joskow. Incentive regulation in theory and practice: electricity distribution and transmission networks. *AEI-Brookings Joint Center Working Paper*, (05-18), 2005.

Jean-Jacques Laffont. *A theory of incentives in procurement and regulation*. MIT press, 1993.

Jean-Jacques Laffont and Jean Tirole. Using cost observation to regulate firms. *The Journal of Political Economy*, pages 614–641, 1986.

Jonathan Levin. Relational incentive contracts. *The American Economic Review*, 93(3):835–857, 2003.

Gerald Marschke and Pascal Courty. The challenge of measuring productivity: A case study of performance measurement in a job training program. Technical report, 2002.

Colin Mayer and John Vickers. Profit-sharing regulation: An economic appraisal. *Fiscal studies*, 17(1):1–18, 1996.

Edward C Norton and Douglas O Staiger. How hospital ownership affects access to care for the uninsured. *The Rand journal of economics*, pages 171–185, 1994.

OECD. Government at a glance 2013. pages 82–83, 2013.

OFGEM. Electricity transmission network reliability incentive scheme: Final proposals. Technical report, Office of Gas and Electricity Markets, 2004d.

OFGEM. Handbook for implementing the riio model. Technical report, 2010.

Benjamin A Olken, Junko Onishi, and Susan Wong. Should aid reward performance? evidence from a field experiment on health and education in indonesia. *American Economic Journal: Applied Economics*, 6(4):1–34, 2014.

Harry Anthony Patrinos and Shobhana Sosale. *Mobilizing the private sector for public education: A view from the trenches*. World Bank Publications, 2007.

Harry Anthony Patrinos, Felipe Barrera Osorio, and Juliana Guáqueta. *The role and impact of public-private partnerships in education*. World Bank Publications, 2009.

Steven Piper and Wade E Martin. Evaluating the accuracy of the benefit transfer method: a rural water supply application in the usa. *Journal of Environmental Management*, 63(3): 223–235, 2001.

Panu Poutvaara. Public-sector outsourcing. *IZA World of Labor*, 2014.

Frank P Ramsey. A contribution to the theory of taxation. *The Economic Journal*, pages 47–61, 1927.

Susan Rose-Ackerman. Altruism, nonprofits, and economic theory. *Journal of economic literature*, pages 701–728, 1996.

David EM Sappington and David S Sibley. Regulating without cost information: The incremental surplus subsidy scheme. *International Economic Review*, pages 297–306, 1988.

David EM Sappington and Joseph E Stiglitz. Privatization, information and incentives. *Journal of policy analysis and management*, 6(4):567–585, 1987.

David EM Sappington and Dennis L Weisman. Price cap regulation: what have we learned from 25 years of experience in the telecommunications industry? *Journal of Regulatory Economics*, 38(3):227–257, 2010.

Richard Schmalensee. Market structure, durability, and quality: a selective survey. *Economic Inquiry*, 17(2):177–196, 1979.

Rachel Silverman, Mead Over, and Sebastian Bauhoff. Aligning incentives, accelerating impact. next generation financing models for global health. 2015.

Richard Steinberg. The revealed objective functions of nonprofit firms. *The RAND Journal of Economics*, pages 508–526, 1986.

Richard Steinberg and Burton A Weisbrod. Nonprofits with distributional objectives: price discrimination and corner solutions. *Journal of Public Economics*, 89(11):2205–2230, 2005.

Anna Ter-Martirosyan and John Kwoka. Incentive regulation, service quality, and standards in us electricity distribution. *Journal of Regulatory Economics*, 38(3):258–273, 2010.

Kenneth Train. Optimal regulation: The economic theory of natural monopoly. Technical report, 1991.

Aidan R Vining and Anthony E Boardman. Public-private partnerships in canada: Theory and evidence. *Canadian public administration*, 51(1):9–44, 2008.

Ingo Vogelsang. Optimal price regulation for natural and legal monopolies. *Economia Mexicana. Nueva Epoca*, 8(1), 1999.

Ingo Vogelsang. Price regulation for independent transmission companies. *Journal of Regulatory Economics*, 20(2):141–165, 2001.

Ingo Vogelsang. Incentive regulation and competition in public utility markets: a 20-year perspective. *Journal of Regulatory Economics*, 22(1):5–27, 2002.

Ingo Vogelsang. Incentive regulation, investments and technological change. 2010a.

Ingo Vogelsang. The tension between incentive regulation and investments in network industries. *DICE REPORT*, 2010b.

Ingo Vogelsang. The endgame of telecommunications policy? a survey. Technical report, CESifo Working Paper, 2013.

Ingo Vogelsang and Jörg Finsinger. A regulatory adjustment process for optimal pricing by multiproduct monopoly firms. *The Bell journal of economics*, pages 157–171, 1979.

Dennis R Young. Complementary, supplementary, or adversarial? nonprofit-government relations. *Nonprofits & government: Collaboration & conflict*, 37, 2006.