

Designing Contracts for the Global Fund: Lessons from the Theory of Incentives

Liam Wren-Lewis

Abstract

This paper uses contract theory to suggest simple contract designs that could be used by the Global Fund. Using a basic model of procurement, we lay out five alternative options and consider when each is likely to be most appropriate. The rest of the paper then discusses how one can build a real-world contract from these theoretical foundations, and how these contracts should be adapted to different contexts when the basic assumptions do not hold. Finally, we provide a synthesis of these various results with the aim of guiding policy makers as to when and how ‘results-based’ incentive contracts can be used in practice.

Keywords: contracts, Global Fund, contract theory, theory of incentives

Designing Contracts for the Global Fund: Lessons from the Theory of Incentives

Liam Wren-Lewis
Paris School of Economics

This paper was commissioned by CGD as a background paper for the Center for Global Development (CGD) Working Group on Next Generation Financing Models in Global Health - see Silverman, Over and Bauhoff (2015) for the final report. The author is grateful for comments and suggestions from all the participants at the April 2015 technical workshop, especially Mead Over and William Savedoff.

CGD is grateful for contributions from the Bill and Melinda Gates Foundation and the UK Department for International Development in support of this work.

Liam Wren-Lewis. 2016. "Designing Contracts for the Global Fund: Lessons from the Theory of Incentives." CGD Working Paper 425. Washington, DC: Center for Global Development.
<http://www.cgdev.org/publication/designing-contracts-global-fund-lessons-theory-incentives-working-paper-425>

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

202.416.4000
(f) 202.416.4050

www.cgdev.org

The Center for Global Development is an independent, nonprofit policy research organization dedicated to reducing global poverty and inequality and to making globalization work for the poor. Use and dissemination of this Working Paper is encouraged; however, reproduced copies may not be used for commercial purposes. Further usage is permitted under the terms of the Creative Commons License.

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors or funders of the Center for Global Development.

Foreword

This paper – and the companion paper by Han Ye ([Working Paper 424](#)) – 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 surveys 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

1 Introduction

The Global Fund for AIDS, TB and Malaria aims to improve people's health across the world by providing funding to domestic institutions committed to reducing the impacts of these diseases. Like many donors, however, it at times has different objectives from the organizations it funds. In particular, the Global Fund would like to ensure that the organizations it funds do all that they can to reduce costs, and hence make more money available to to fund other work. Recipient organizations, on the other hand, may not be intrinsically motivated to reduce costs, since they are not particularly concerned about the other activities the Global Fund might spend that money on. Thus, even when all actors have noble aims, a tension may arise between the funder and the funded.

Economists have long been aware of such a tension when it comes to problems of regulation and procurement. Though these contexts are different in their institutional setup, the fundamental problem remains the same: Asymmetric information. At the most basic level, if the Global Fund (GF) could observe the potential cost-saving actions to be made by the recipient, then it would simply include such behavior as part of the deal. Even if the Global Fund could not observe cost savings directly, it could impute them if it knew of all the other

reasons why costs might vary. The tension thus arises because the agents funded by the Global Fund can take advantage of the extra information they have regarding the challenges they face. Essentially, when the Global Fund faces a large bill, it does not know whether this is because it is unlucky or because it is being taken advantage of.

Over the last thirty years, economists have analyzed how best to deal with this problem. Exemplified by the seminal text produced by Laffont and Tirole (1993), this literature studies contracts that provide incentives to the agent to reduce costs in the presence of such information asymmetries. Whilst this work was recently recognized in the awarding of the Nobel Prize to Jean Tirole, there has at times been disappointment with the extent to which it has been drawn upon by policy makers. In particular, the complexity of the optimal contracts derived has sometimes been seen as a barrier to real world applications.

The aim of this paper is to draw out the insights produced by this literature that are relevant to the Global Fund. In particular, the paper uses the framework of procurement designed by Laffont and Tirole (1993) and developed by others to propose simple contracts that could be used with donor recipients. The idea is both to capture the major lessons produced by this work and understand how contract design may need to be adapted to various contexts the Global Fund faces.

A few papers have used a principal-agent framework to consider aid contracting. Azam and Laffont (2003), and Clist and Verschoor (2014) consider the costs and benefits of contracting on performance, but they contrast ‘conditionality’ with more hands off approaches (e.g. budget support) or donations to alternative targets (e.g. NGOs). In particular, they do not consider the type of contract that is typical for the Global Fund, where payment is conditional on the money having been spent on appropriate inputs. Cordella and Dell’Ariccia (2007) consider such conditioning on inputs, and show how it may distort project choice (because only some inputs are observable), but they do not contrast this with contracting based on outputs.

Output-based contracting has been recently considered through the form of ‘Cash-on-Delivery’ (COD) and ‘Results-based financing’ (RbF) - see, for instance, Birdsall, Savedoff, Mahgoub and Vyborny (2010). This, and other contracting arrangements, have been specifically considered in the framework of Global Fund contracts in Glassman, Over and Fan (2013). This article builds on the ideas sets out in these books by attempting to model some of the key differences between contracting frameworks and relate them to the theoretical literature on incentive contracts.

We begin in the next section by setting out the basic model of procurement used by Laffont and Tirole (1993). For reference, we also briefly describe the optimal contract generated by

the model, though such a contract is likely to be too complex for the Global Fund to use. Section 3 then sets out a number of simple contracts that could be used in practice and uses the model to understand the advantages and disadvantages of each. After providing some tentative conclusions as to which contract might be most useful in practice, Section 4 then explains how a complete contract could be ‘built up’ from elements that resemble those modeled. Section 5 then explores how such contracts might need to be adapted according to the context. Finally, we conclude by synthesizing these findings and generating key questions which will need to be answered in order to move forward with contract design.

2 A basic model of procurement

Let us begin by setting out the basic model of procurement described in Chapter 1 of Laffont and Tirole (1993). The model is extremely simple and is designed to demonstrate the basic intuition behind different kinds of contract design - in Section 5 we will consider how it can be extended to be more realistic.

Before we begin, it should be noted that the framework of Laffont and Tirole (1993), and to a certain extent this paper, is ‘Bayesian’ in the sense that prices are calculated as being optimal given an assumed probability distribution of costs. In practice, this probability distribution of costs would be estimated based on the collection of cost information by the GF.¹ This approach contrasts to a more ‘non-Bayesian approach’ where instead the focus is on simpler pricing rules that may converge over time to some optimum. However, this distinction is somewhat misleading for the purposes of this paper since our focus is on what kind of contracts to use, rather than how to set prices. Since we restrict ourselves to considering simple contracts, our results are not dependent on optimal pricing and would be very similar were we to assume some more ad-hoc pricing strategy.

The model contains a principal, the Global Fund (GF), who is paying an agent, the PR, to complete a project.² We use the pronoun she with respect to GF and he with respect to PR. We presume that the project is indivisible, in that it can either be produced or not, and hence

¹Such a process would draw on experience with previous contracts as well as cost-analysis studies such as Marseille, Giganti, Mwangi, Chisembele-Taylor, Mulenga, Over, Kahn and Stringer (2012); Tagar, Sundaram, Condliffe, Matatiyo, Chimbwandira, Chilima, Mwanamanga, Moyo, Chitah, Nyemazi, Assefa, Pillay, Mayer, Shear, Dain, Hurley, Kumar, McCarthy, Batra, Gwinnell, Diamond and Over (2014); Meyer-Rath and Over (2012).

²PR stands for ‘Principal Recipient’, but we use the abbreviated term here to avoid confusion with the principal within the game.

we temporarily abstract from questions about quantity (this assumption is relaxed in Section 4.1.1). This might be, for instance, the completed construction of a health clinic, where the capacity of the clinic is fixed.

To keep the model simple, we presume that only GF values the project, and in particular that she place a value V on the project being completed.³ Let t be the amount that is transferred to PR in exchange for completing this project. GF's payoff function is therefore $V - t$. We assume that GF observes whether the project has been completed and can make payments conditional on project completion - we consider what happens if these assumptions are relaxed in sections 5.1 and 5.2 respectively.

In order to complete the activity, PR has to pay a cost c . We presume that this total cost is observed by the GF - that is, the GF can identify the costs that are incurred by the PR in order to complete the project the GF desires. In other words, we assume that the auditing technology is sufficiently robust that the PR will not lie over the total costs it reports to the GF. We consider relaxing this assumption in Section 5.2.

We suppose that the cost c is made up of two components, β and e , where

$$c = \beta - e \tag{1}$$

In this equation, the first component β represents the 'innate' cost of completing the project - that is, the part of the cost that is outside of PR's control. For instance, if the overall cost is the wage-bill of a health clinic, this parameter might represent the part of the wage-bill stemming from the availability of healthcare professionals. We presume there is some uncertainty about the value of this variable, such that β is drawn from a distribution with cumulative density function $F(\beta)$, with β taking a value in the range $[\underline{\beta}, \bar{\beta}]$. The size of this range $\bar{\beta} - \underline{\beta}$ therefore represents the uncertainty that exists around PR's innate costs.

The second component of the firm's cost, e , represents the results of effort made by PR to reduce costs. Exerting effort e gives PR a dis-utility of $\psi(e)$, where $\psi(0) = 0$, $\psi'(\cdot) > 0$ and $\psi''(\cdot) > 0$. In the case of the health clinic, this might represent the effort PR goes to in finding the cheapest healthcare professionals available, or in negotiating down their wages. In this case, the dis-utility function $\psi(e)$ might represent the costs involved in searching or negotiating, or the fact that the more expensive professionals might be connected to PR and hence are able to provide him with benefits. PR's payoff function is thus given by the

³We consider the situation where the PR may value the project themselves in Section 5.5.

expression U , where

$$U = t - c - \psi(e) \tag{2}$$

We assume GF makes a take it or leave it offer to PR - that is, we assume GF has all the bargaining power. This assumption is not necessary, but essentially corresponds to the worst case scenario when it comes to dealing with problems of information asymmetry. Were PR to have bargaining power, it would be able to extract rents from the relationship regardless of any information it held, and hence would not need to exploit any asymmetry.

Finally, we assume that PR has a participation constraint - that is, he will not undertake the project if doing so would make him worse off - i.e. we require $U \geq 0$. In section 5.5, we consider relaxing this assumption to allow for the situation where PR is willing to use his own funds to partly finance the project.

This thus completes our setting out of the model. We now proceed to describe the optimal contracts when information is symmetric and in the case where it is asymmetric. Since the optimal contract in the latter case is quite complex, the next section then uses the model to analyze simpler contracts that may be used by GF.

2.1 Solution with symmetric information

As a benchmark, let us first consider the situation where there is no information asymmetry. In particular, we assume that GF observes β . Since we have already assumed that GF observes the total cost c , it can back out how much effort PR has made and hence all information is known by both players.

With symmetric information on costs, GF can calculate exactly how much money PR will need to complete the project before PR decides upon his effort level. The optimal level of effort for PR to exert is that which minimizes the net cost $\beta - e + \psi(e)$. Differentiating by e and setting to zero thus gives us the optimal level e^* :

$$\psi'(e^*) = 1 \tag{3}$$

When $e = e^*$, the reduction in costs as a result of cost saving effort is $e^* - \psi(e^*)$. Since this quantity will be used frequently within the paper, let us define the value $k = e^* - \psi(e^*)$ as the maximum amount of net cost-reduction.

One very simple contract that GF can use to ensure first best effort is exerted is one where

the payment received by PR does not depend on the cost. In this way, PR is the residual claimant and gets all the benefits of any cost-reduction he makes. He will therefore choose the optimal amount of cost-reduction. By choosing to pay an amount $\underline{t} = \beta - k$, GF further ensures that PR's participation constraint is always binding. PR therefore receives no rent, and costs have been minimized.

2.2 Asymmetric information

Now let us suppose that GF does not observe β . As far as she is concerned, β may lie anywhere between $\underline{\beta}$ and $\bar{\beta}$, although she does know the cumulative distribution function $F(\beta)$. The problem facing GF is now much more difficult. Although she observes c after effort has been exerted, she cannot disentangle which part of this is made up of β and which part is made up by e . Suppose, for instance, she observes a cost of $\bar{\beta} - k$. This could be the result of bad luck, even if PR was making the optimal amount of effort - it just happened that $\beta = \bar{\beta}$. On the other hand, it could have been the result of idleness on the part of PR - β could have taken the value $\bar{\beta} - k$ and PR made exactly zero effort.

If β is high, this information asymmetry is of relatively little use to PR. In particular, if $c = \bar{\beta}$, GF will be aware that no effort was exerted. However, if β is relatively low, then there is room for PR to cheat. For instance, if $\beta \leq \bar{\beta} - k$, then PR can claim that β took the value $\beta + k$, and that it has exerted effort to reduce this cost. In our earlier example, PR can claim that it has to pay high wages, due to the limited availability of healthcare professionals, when in reality it could have negotiated harder and paid less.

In this situation, we can see that GF cannot implement the contract which worked best in the symmetric information case. If GF asks PR for the value of β and then implements the contract outlined above, PR will lie. In particular, if GF will pay $\bar{t} = \hat{\beta} - e$, where $\hat{\beta}$ is the value of β announced by PR, then PR will always announce $\hat{\beta} = \bar{\beta}$.

The optimal contract for the Global Fund to offer in this situation is derived in Laffont and Tirole (1986). Fundamental to the optimal contract is idea is that the Global Fund can offer the agent a 'menu' of contracts - that is, GF could propose different forms of reimbursement and PR would be free to choose which it preferred. In equilibrium, the Revelation Principle tells us that the optimal mechanism is therefore equivalent to a situation where the agent is incentivized to declare it's true 'type' (in this case, the value of β).

An important idea in the optimal contract is that the agent will receive an *information rent* when β is low. This rent is designed to incentivize the agent to reveal that the innate

costs are low, since otherwise it could gain by pretending that β is high. Since PR must not make a loss when β is high, and PR can always pretend that β is high even when β is low, this rent must be at least as large as the gain PR can get by pretending β is high.

One important insight derived by Laffont and Tirole (1986) is that, when $\beta = \underline{\beta}$, there is no reason to induce anything other than the efficient level of effort. In the optimal mechanism, therefore, the Global Fund will offer a fixed price contract if the PR declares that $\hat{\beta} = \underline{\beta}$, which induces effort e^* .

However, when $\beta > \underline{\beta}$, the Global Fund does have a good reason to induce effort that is below first best. For any given $\hat{\beta}$, the Global Fund wishes to reduce the information rent that it has to give to the agent when β takes any value smaller than $\hat{\beta}$. One way of doing this is to compensate the agent through cost-reimbursement rather than a fixed fee. This is easiest to see at the extreme - if the agent only is paid through cost-reimbursement when $\beta = \hat{\beta}$, an agent with $\beta > \hat{\beta}$ has no incentive to pretend to have $\beta = \hat{\beta}$, since he would not make any profit (he would simply get his lower overall costs reimbursed).

In general therefore, the Global Fund faces a trade-off in choosing a contract for all $\beta > \underline{\beta}$. On the one hand, she wishes to pay such an agent through cost-reimbursement in order to reduce the information rent that she is obliged to give the agent when β is lower. On the other hand, she knows that doing so will reduce the effort that the agent exerts.

We do not derive the optimal contract for the Global Fund here, but its form is demonstrated in Figure 1.⁴ Essentially, the Global Fund offers the agent a choice of contracts along the curve given in the figure. For instance, the agent could choose to have contract at point A, which would be simply a FP contract - \$ 100, for instance. A contract at point C, on the other hand, would involve a relatively small fixed payment - \$ 20, for instance - and a share (strictly less than one) of costs reimbursed - for instance, the GF agrees to reimburse 70% of the money spent by the PR. Contracts in between, such as that at point B, would offer an in-between level of fixed payment with an in-between amount of cost-reimbursement - i.e. a fixed payment of \$70 and 35% of costs reimbursed.

As the agent's innate costs β increases he will be more tempted to pick a contract to the right of this curve.

Although this contract is the optimal one for the Global Fund to offer, we presume it is not practicable for the Global Fund to use. In particular, we consider there are two dimensions along which it may be too complex. First, the PR is given a continuum of options to choose

⁴See Laffont and Tirole (1986, pp. 63-70) for a formal derivation.

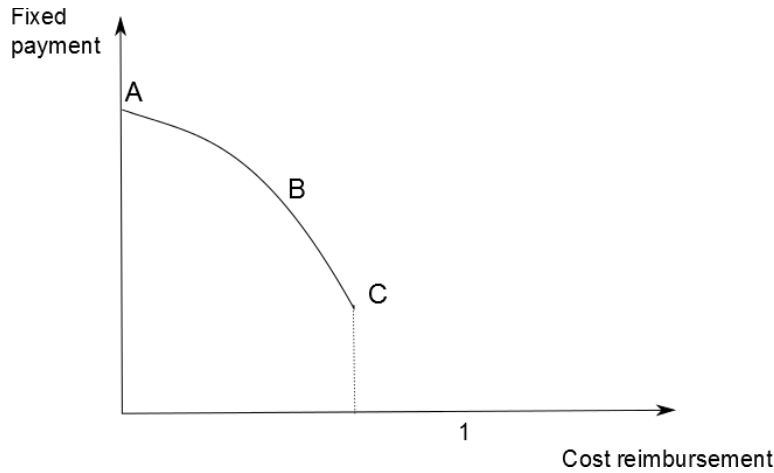


Figure 1: Optimal contract for the Global Fund

from. It is unlikely in such a situation that PR will choose optimally, and indeed the complexity of being given so many options may risk confusion. Second, deriving the optimal curve is a complex process for the Global Fund, with it depending on estimating the distribution of innate costs as well as PR's cost-saving functions. If GF makes mistakes in deriving the curve, even an optimizing PR will probably end up choosing only one of a few points on the curve, and much of the benefits of the contract will be lost. Finally, it is also worth noting that this is the optimal contract in the simple case where there is only one good of uniform quality, and the contract will add further degrees of complexity once we allow for more complicated situations as we do in Section 5.

Given this complexity, we instead proceed to analyze a set of simpler contracts that we believe might be practicable for the Global Fund. As we will discuss in the following section, some of these contracts have been shown to capture much of the gains of the optimal contract under reasonable assumptions, and hence we are probably not throwing much away by disregarding such complex contracts.

3 Simple contracting solutions

This section outlines five contracts which we believe could form the basis of a deal between the Global Fund and PRs. These contracts are as follows:

1. Cost-reimbursement (CR)

2. Fixed price (FP)
3. Linear cost-sharing (LCS)
4. Fixed price / cost-reduction (FPCR)
5. Linear cost-sharing / cost-reduction (LCSCR)

These contracts have been chosen based on their simplicity and the extent to which they have been used in practice by other organizations facing similar incentive problems.⁵ In each case, the payment to PR takes the following form:

$$t = \underline{t} + \alpha c \tag{4}$$

where \underline{t} and α are parameters fixed by GF.

We proceed to outline the details of each contract and then discuss the advantages and disadvantages of each. We begin with the most simple contracts and gradually get more complex. Finally, we conclude the section by summarizing the contracts and describing the conditions under which each may be well suited.

3.1 Cost-reimbursement

In our model, a cost-reimbursement (CR) contract corresponds to a contract where GF makes a payment $t = c$. In terms of equation 4, it corresponds to a value of $\alpha = 1$ and $\underline{t} = 0$. This is perhaps the contract that most closely resembles the traditional funding model used by the Global Fund. In particular, PR is only given money to reimburse costs that it can prove it has spent.

Besides its simplicity, the main advantage of this model of financing is that PR never makes any ‘profit’ from the money he receives from GF. When β happens to be low, the Global Fund can take advantage of that by spending less money on the project. Moreover, PR will always participate, as he will never make a loss when all his costs are reimbursed.

The key disadvantage with this contract however is that PR will never exert any effort to reduce costs, since all the benefits of any cost reduction accrue to GF. Hence $c = \beta$ for all β . The expected cost for the Global Fund under this contract is therefore $\mathbb{E}[\beta]$.

⁵See, for instance, the examples given in Reichelstein (1992), Chalkey and Malcomson (1998), Gasmi, Laffont and Sharkey (1999) and Rogerson (2003).

Contracts such as cost-reimbursement are often described as *low-powered* incentive contracts, because the agent is given little incentive to reduce costs.⁶

3.2 Fixed Price

A Fixed Price (FP) contract is possibly as simple as the CR contract. In this scheme, GF pays PR a fixed amount \underline{t} and does not reimburse any of his costs - i.e. $\alpha = 0$. This is perhaps close to the type of model that the GF theoretically implements in Rwanda, or the DFID scheme for education in Ethiopia.⁷ In each case, the donor pays a fixed amount that does not vary according to how much they actually spent.

If PR accepts the contract, he will then be the residual claimant for any cost savings made. Hence he will exert the optimal amount of effort e^* , and the realized cost of the project will be $\beta - k$ (where $k = e^* - \psi(e^*)$). PR's expected payoff will thus be $\underline{t} - \beta + k$, and hence he will only accept the contract under the condition that $\beta \leq \underline{t} + k$.

In designing this contract, GF has one parameter to set - the amount \underline{t} . If the project is extremely valuable to the Global Fund and she wants to ensure it is undertaken under any condition, she will set $\underline{t} = \bar{\beta} - k$. In this way, PR will accept the contract no matter the value of β . However, if the project is not so valuable to GF, she may want to pay a price below this value in order to reduce the payment paid when the project is accepted. In this case, she will set \underline{t} to maximize her expected payoff $(V - \underline{t})F(\underline{t} + k)$. Differentiating by \underline{t} and setting to zero gives us:

$$(V - \underline{t})f(\underline{t} + k) = F(\underline{t} + k) \tag{5}$$

If β is distributed uniformly, then this gives us that the optimal value of \underline{t} is:

$$\underline{t}^* = \min \left\{ \frac{\beta - k + V}{2}, \bar{\beta} \right\} \tag{6}$$

Note however that this assumes that the GF will not come back with a higher offer if the PR

⁶This is the same notion of 'power' as considered in the recent article by Geruso and McGuire (2014). There is also some parallel between their notion of 'fit' and the amount of cost-reimbursement, since both describe the correlation between the cost of the contract (in our case to the GF) and the realized payments by the agent (in our case the PR).

⁷In practice, the weak verification and response to outputs may mean that the GF model in Rwanda operates more like budget support, which relies on the agent's willingness to implement the project without incentives.

rejects. Since in reality such commitment may not be possible, most likely a FP contract will need to set $\underline{t} = \bar{\beta} - k$.

The main advantage of a fixed price contract is that it always induces the first-best amount of cost-reduction. It is also very simple, and in theory means that GF does not have to collect information on the costs incurred by PR.⁸

The main disadvantage of a fixed price contract is that PR may make a large ‘profit’ at the expense of GF. In particular, if innate costs are low, PR will make money just because it happened to get lucky. GF may reduce this profit by reducing the fixed payment, but this risks the project not being undertaken even when it is valuable.

To understand when the fixed price contract will be better than cost-reimbursement, consider the simple situation when the project is very valuable and hence $\underline{t}^* = \bar{\beta} - k$. In this case, the expected cost to GF is precisely $\bar{\beta} - k$. This will be lower than the expected cost of the CR contract when k (the potential for cost savings) is large compared to $\bar{\beta} - \mathbb{E}[\beta]$ (the degree of uncertainty over innate costs).

Contracts such as fixed-price are often described as *high-powered* incentive contracts, because the agent is given a large incentive to reduce costs.

3.3 Linear cost-sharing

The contracts so far considered in some sense represent two extremes - on the one hand, costs are reimbursed in their entirety, and on the other, no costs are reimbursed. A linear cost-sharing rule represents a compromise between these two contracts, where the proportion of costs reimbursed, α , may lie between zero and one. Since this is clearly not enough for PR to cover all of his costs, he also receives a fixed payment \underline{t} .

We can immediately see that the effort exerted by PR will be in between zero and the first best. In particular, PR will minimize his residual costs which are $(1 - \alpha)(\beta - e) + \psi(e)$, giving $\psi'(e) = 1 - \alpha$. Let us label this value of e as $e(\alpha)$. In the case of linear cost-sharing contracts, a more high-powered contract is therefore one with a lower value of α .

GF now has two parameters to set - α and \underline{t} . Gasmi et al. (1999) derive the optimal contract when \underline{t} is set sufficiently high that the project will always be undertaken, which is probably the most relevant situation given that it may be politically infeasible for the GF to

⁸GF may still wish to collect cost information in order to improve other contracts, both in other countries and with the same PR in the future. This latter point is discussed in section 4.4.

make a take-it-or-leave-it offer.⁹ Indeed, they show that such a contract generally captures a large amount of the gains of the optimal contract. In this case, setting \underline{t} sufficiently high that the project will always be undertaken is less costly than in the fixed contract case, because the value of \underline{t} required is lower (due to the cost sharing). In particular, we will have $\underline{t} = (1 - \alpha)(\bar{\beta} - e(\alpha) + \psi(e(\alpha)))$.

The expected cost to GF of this contract is then:

$$\mathbb{E}[t] = (1 - \alpha)(\bar{\beta} - e(\alpha)) + \psi(e(\alpha)) + \alpha(\mathbb{E}[\beta] - e(\alpha)) \quad (7)$$

$$= (1 - \alpha)\bar{\beta} + \alpha\mathbb{E}[\beta] - e(\alpha) + \psi(e(\alpha)) \quad (8)$$

Differentiating this equation by α and setting to zero then gives the interior solution for α :

$$\alpha^* = \psi''(e(\alpha^*)) (\bar{\beta} - \mathbb{E}[\beta]) \quad (9)$$

If $\alpha^* > 1$, then the optimal LCS contract is a CR contract. Note that we will never have $\alpha^* \leq 0$, and hence a FP contract is never strictly optimal. This is because the cost to GF of having $\alpha = \epsilon$ for small ϵ is second order, since at $\alpha = 0$ exerted effort is first-best, whereas the cost-savings are first order. Note however, that α^* may still be fairly close to zero, and hence the Global Fund might prefer an FP contract to an LCS contract with small α just because the former contract is simpler.

Suppose that the cost of exerting effort e - the function $\psi(e)$ - is quadratic, and in particular is given by the following function:

$$\psi(e) = \frac{e^2}{4k} \quad (10)$$

Under this assumption, the definition of α^* becomes

$$\alpha^* = \frac{\bar{\beta} - \mathbb{E}[\beta]}{2k} \quad (11)$$

Setting this contract optimally may not be realistic, since the Global Fund may not feel confident in estimating a cost-reduction function $\psi(e)$. But from this equation we can learn

⁹For instance, in countries where the GF has already funded ARTs for patients, the GF may have a moral obligation to continue funding such treatment.

the broader point that the optimal amount of cost-sharing is increasing in the innate cost uncertainty and decreasing in the potential costs savings. As we previously saw, cost-recovery will be attractive when the innate cost uncertainty is much larger than the potential cost savings - in particular, the above equation says, if β is distributed symmetrically, cost-recovery will be optimal when the range of possible innate costs $\bar{\beta} - \underline{\beta}$ is more than four times the potential first-best cost savings. On the other hand, if potential cost-savings are very large compared to innate cost uncertainty, then a fixed-price contract will be close to the optimal linear cost-sharing contract. Linear-cost sharing rules with $0 < \alpha < 1$ are likely to be most useful therefore when innate cost uncertainty is of a similar magnitude to the potential cost savings - i.e. where the range of potential innate costs $\bar{\beta} - \underline{\beta}$ are between one and three times the amount of potential cost savings k .

3.4 Fixed-price / Cost-reimbursement menu (FPCR)

An important insight in the derivation of the optimal contract by Laffont and Tirole (1986) is that menus of contracts can be a useful way to reduce the effects of asymmetric information - they allow the Global Fund to retain relatively high-powered incentives whilst reducing the expected information rent. Although the optimal contract involved a continuum of menus, Wilson (1989) shows that, if instead there are a finite number of contracts n on the menu, the value of additional one decreases rapidly in n . In practice, this means that a large portion of the gains of using menus is captured by moving from one option (i.e. no choice) to a menu with two options.

Given this, the simplest possible menu is where PR is offered a choice between an FP contract (with a given \underline{t}) and a CR contract. In this case, the only parameter that GF has to set is the price \underline{t} given in the fixed price contract. Note now that the trade-off when setting this value is different from that considered when only an FP contract was offered. There is no longer any risk that PR will refuse to implement the project, since he can always choose the CR contract. Instead, the trade-off is between the cost of having to pay a higher transfer and the benefit of getting such a contract chosen more frequently, and hence getting more expected cost reduction.

Let β^i be the value of β where PR is indifferent between a FP contract and a CR contract. Since he makes no profit under a CR contract, and he exerts effort so that the net cost-saving

is k under a FP contract, β^i is given by the following expression:

$$\beta^i = \underline{t} + k \tag{12}$$

If $\beta > \beta^i$, PR will pick a CR contract, since it would make a loss under the FP contract. On the other hand, if $\beta < \beta^i$, PR will make a profit under the FP contract and hence choose that.

Given this, GF will pick \underline{t} to minimize the following expected cost:

$$F(\beta^i)\underline{t} + (1 - F)F(\beta^i)\mathbb{E}[\beta|\beta \geq \beta^i] \tag{13}$$

Differentiating and setting to zero gives us a solution for the value of β^i at which the PR is indifferent between the two contracts:

$$kf(\beta^i) = F(\beta^i) \tag{14}$$

The solution for β^i could be calculated using a computer if a distribution function was specified. Rogerson (2003) shows that the answer is particularly simple if we assume that β is distributed according to the uniform distribution. In particular, this gives us $\beta^i = \underline{\beta} + k$, which corresponds to a fixed price of $\underline{\beta}$. Since we cannot have $\beta^i > \bar{\beta}$, the optimal value of \underline{t} is given according to the following equation:

$$\underline{t} = \min\{\underline{\beta}, \bar{\beta} - k\} \tag{15}$$

Note that, if $\underline{t} = \bar{\beta} - k$, then PR will choose the FP contract no matter the value of β , and hence in this case an FPCR menu of contracts is equivalent to a simple FP contract.

Note that, holding the expected cost constant, the price that the GF should offer in an FPCR contract is decreasing in the level of cost uncertainty. This is because, for a given offered price, greater uncertainty will not change the proportion of values of β for which the PR will choose the FP contract, and hence the expected amount of cost reduction will stay constant. However, greater cost uncertainty means that there is a higher probability that the actual cost is a large amount below the price offered, and hence there are more values of β where a CR contract would be cheaper for the GF. Looking at it another way, high cost-uncertainty generally favors CR contracts over FP ones, and hence the GF should set a lower price which will mean that the PR picks the FP option less often.

Rogerson (2003) gives a simple example of how such a contract might work in practice.

Suppose that GF believes the innate cost β to be distributed uniformly between 90 and 110. Then, if the potential efficiency gains of offering a fixed price contract, k , are less than 20, the GF should offer a fixed price contract of 90. This will be accepted by the agent if the cost, given the efficiency gains $(\beta - k)$, is less than 90, and the agent will choose cost reimbursement otherwise. On the other hand, if GF believes that the potential efficiency gains are greater than 20, she should offer a fixed price contract of $110 - k$, such that the agent will certainly take it.

Despite being much less complex than the optimal contract outlined in Section 2.2, Rogerson (2003) shows that the optimal FPCR contract can perform almost as well. In particular, if β is uniformly distributed and the cost of effort function $\psi(e)$ is quadratic, then the optimal FPCR contract captures at least three-quarters of the efficiency gains achieved by the fully optimal contract.

The FPCR's simplicity also gives it another advantage over other contracts, which is that it performs well when the Global Fund is uncertain in the 'knightian' sense. That is, it may be the case that the Global Fund does not feel comfortable estimating a particular effort saving function $\psi(e)$. Without doing so, it obviously cannot calculate a contract that minimizes the expected cost, but it may instead aim to minimize the maximum cost (over all possible functions ψ). This is the case considered by Garrett (2014). He shows that, if all the principal knows is the minimal net cost saving \underline{k} , and she wishes to minimize the worst-case expected payment, then the best contract is in fact an FPCR contract with price set as follows:

$$\underline{t} = \min\{\underline{\beta}, \bar{\beta} - \underline{k}\} \tag{16}$$

The basic rationale behind such a result may be the real reason an organization like the Global Fund might prefer simple contracts - that calculating an optimal amount of cost-sharing α will involve making judgments on the PR's ability to save costs, and hence will make GF worse off if these judgments are wrong. In a 'risk-averse' (or technically, an uncertainty averse) setting such as a large bureaucratic organization, minimizing the cost of the 'worst-case' scenario may be very appealing.

3.5 Linear cost-sharing / Cost-reimbursement menu (LCSCR)

In the same way that an FP contract is essentially a limiting case of an LCS contract, we can consider a more general menu of two contracts where PR chooses between a CR contract and

a particular LCS contract. In this case, GF sets parameters α and \underline{t} for an LCS contract, and offers PR the choice between this and complete cost-reimbursement.

Why would such a generalization be necessary when Rogerson (2003) has shown that FPCR contracts can be very effective? The reason, as shown by Chu and Sappington (2007), is that FPCR contracts can perform badly when we move away from Rogerson's assumptions. In particular, when innate costs β are substantially more likely to be high than low, but low is still possible, then FPCR contracts perform badly.

To see the intuition behind this, consider a case where most of the time innate costs were between 100 and 110, but there is a small chance it might take a value between 50 and 100. Suppose furthermore that k is around 10. In this case, the optimal FPCR contract may well be to offer a choice between a CR contract and a FP contract with a price between 50 and 100, since offering a higher price will leave a lot of profit to those with low innate costs. As a result, since it is relatively unlikely that innate costs will be within the lower range, most of the time PR will choose the CR contract. Hence, most of the time, no efficiency savings will be made.

Chu and Sappington (2007) show that, in this situation, LCSCR contracts can do better. In particular, they consider the case where the distribution of β has cumulative density function $F(\beta) = \left(\frac{\beta - \underline{\beta}}{\bar{\beta} - \underline{\beta}}\right)^\delta$, with $\delta \in [0, \infty)$. Note that when $\delta = 1$ this is the uniform distribution, and higher values of δ mean that higher values of β are relatively more likely. They further assume costs of exerting effort are quadratic in the same way as previously, i.e. with $\psi(e)$ given by equation (10).

Chu and Sappington (2007) derive the optimal LCSCR contract and consider empirically when it outperforms FPCR. In particular, they find that the optimal α is given according to the following expression:

$$\alpha^* = \min \left\{ \frac{\delta}{\delta + 2}, \frac{\bar{\beta} - \underline{\beta}}{2k(\delta + 1)} \right\} \quad (17)$$

In this case, the fixed payment \underline{t} is given according to the expression:

$$\underline{t} = (1 - \alpha^*)(\underline{\beta} + (1 + \alpha^*)k\delta) + k(1 - \alpha^*)^2 \quad (18)$$


They then show that such a contract substantially outperforms an FPCR contract when δ is greater than 1 and $\frac{\bar{\beta} - \underline{\beta}}{k}$ is greater than 1.

In reality, deriving a LCS contract with exactly these values of \underline{t} and α may be too technically challenging for the Global Fund, but the broader point is that the GF might find that a LCSCR contract preferable to an FPCR contract when innate costs are larger than potential cost reductions and higher innate costs are more likely than lower innate costs.

3.6 Which contract is most appropriate?

In this section we have considered five simple contracts that could be adapted to the Global Fund's needs. Before we proceed to consider how such contracts would need to be adapted, it is useful to understand which settings might favour each type of basic contract. This is outlined in Figure 3.6.

Single contracts	<p>CR</p> <p>Works well if:</p> <ul style="list-style-type: none"> - Small potential for cost reduction 	<p>LCS</p> <p>Works well if:</p> <ul style="list-style-type: none"> - Innate cost uncertainty not too large - Potential for cost reduction not too large 	<p>FP</p> <p>Works well if:</p> <ul style="list-style-type: none"> - Small innate cost uncertainty
	Menu of two contracts		<p>LCSCR</p> <p>Works well if:</p> <ul style="list-style-type: none"> - Innate costs are more likely to be at the high end of the distribution than the low end and large innate cost uncertainty



Power of incentives

Figure 2: Conditions under which each contract works well

In some sense, all the contracts are versions of the LCSCR contract described. The FPCR contract is a version of the LCSCR where $\alpha = 0$, whilst the single contracts are versions of

the menus where one of the items on the menu will never be chosen by the PR. Thus, one strategy to picking a contract is to recommend the LCSCR contract and then allow the choice of parameters to include these extremes.

However, in practice, such a method may not be pragmatic since the LCSCR contract is also the most complex. There may be little value in using time and effort, as well as potentially political capital, in getting the Global Fund to accept the use of such a complex contract if it turns out that one of the simpler contracts will generally suffice. Thus a more sensible strategy may be to pick the simplest contract that is likely to generate most of the potential gains.

It is not clear in an abstract sense what makes a contract ‘simple’. In particular, the pragmatic problems with using a menu of contracts are unclear. Arguments can be made both that offering menus is unrealistic and that doing so is trivial.

On the one hand, the Global Fund may initially find the concept of offering a menu of contracts a strange one, since this is not how it typically proceeds. Objections may be raised along at least three lines. First, offering a menu of contracts involves more work for the Global Fund, because it has to design two contracts rather than one. Second, offering a menu of contracts only works if the recipient believes the offer is credible, and it is not clear that the Global Fund could bind itself to not going back on her original offer. Third, the value of offering a menu of contracts relies on the recipient choosing well, and in practice recipients may be bad at making such decisions when they involve multiple actors and a lot of uncertainty.

On the other hand, since both of the menus suggested in this paper involve a cost-reimbursement option, in practice the idea of a menu may appear more alien than it really is. Since one option in the menu is essentially the status quo, a menu involves the Global Fund suggesting a new type of contract, and then letting the recipient decide whether it wants to do things the old way or the new way. In reality, this may be a much more credible offer than forcing the recipient to agree to a single FP or LCS contract, and is typically how the GF has proceeded when offering experimental results-based financing contracts. Indeed, seen this way, it is clear that this does not involve much more work than designing a single new contract. Finally, the risk of the PR ‘choosing badly’ has to be set against the risk of the Global Fund ‘designing badly’. If the Global Fund were to mis-estimate the possible cost-savings or the range of innate costs, offering the status quo as an option ensures that the fund’s programme in a country is not derailed as a result.

Given this, it overall seems reasonable to argue that offering a menu may be ‘simpler’ than designing a LCS contract, and hence an FPCR contract should be the place to start.

As previously discussed, this contract also generally has good properties in terms of achieving cost-reduction and minimizing the maximal cost. If there is reasonable evidence to suggest that such a contract will not perform well (i.e. we are in a situation as described in Section 3.5), then it may be necessary to consider an LCSCR contract or an LCS contract as an alternative.

4 Building a real contract

The previous section described various simple contracts and provided some recommendations as to which might be best for the Global Fund in different circumstances. Note, however, that the model through which the contracts were considered was very simple. In particular, there was only one good (i.e. there were no questions of quantity) and quality or user pricing were not issues. In reality, of course, Global Fund contracts are much more complex. In this section we will discuss how we can ‘build up’ a more complex contract from the basic elements described in the previous section.

4.1 Separating a contract into parts

In the previous section we considered different contracts where the Global Fund was procuring a single good from the PR. In reality, the GF procures from the same PR varying quantities of a large number of different goods and services of varying qualities. How should we adapt the contracts from the previous section to such a context? The first step is to think about how to split up the overall contract into smaller parts. These individual parts then are closer to the single good considered in the previous section, and we can then think about how best to contract for each of them.

4.1.1 Multiple units of the same good

In most circumstances, the Global Fund does not contract on indivisible single items, but goods where the quantity produced can be varied. Such quantities can include the number of patients treated, the prevalence rates of diseases or the amount of drugs distributed. How should the contracts suggested above be adapted to contracts where a quantity q is required?

A simple case to analyse is one where average non-fixed costs, and therefore marginal costs, are constant. In particular, rather than producing just one good at a cost $c = \beta - e$, assume that the agent produces a variable quantity q at a cost $C = (\beta - e)q$. In this simple case, it

is not complicated to adapt the contracts above. The payment made by the Global Fund can now be described as:

$$t = \underline{t}q + \alpha C \tag{19}$$

It is straightforward to show that the choice of contract to use is not related to the quantity produced. In other words, we should choose the form of contract to use on exactly the same criteria as if there were one good, and then apply this to every unit.

If marginal costs are not constant, both uncertainty about innate costs and the potential for cost reductions might vary as function of quantity. A typical situation might be one where the Global Fund is more certain of the costs involved in producing a small quantity of the good than a large quantity. For instance, if the country has recently expanded treatment from 30 to 40% of HIV infected persons at a cost of \$X million, the Global Fund may estimate that expanding treatment to 50% will cost a similar amount. On the other hand, it would probably be less sure about the cost of expanding treatment to 60% and very uncertain as to the cost of expanding treatment to 90%.

Since we saw in the previous section that the optimal contract is dependent on the amount of cost uncertainty, the Global Fund may want to offer a different contract for the first unit of prevalence reduction than others. In particular, a contract may look something like the following:

- For the first q_1 units, payment will be according to payment scheme 1
- For units more than q_1 , payment will be according to payment scheme 2

Since the incentive scheme should be more high powered when cost uncertainty is lower, it may be that payment scheme 1 is an FP contract and payment scheme 2 is CR. Alternatively, all payment schemes might be FPCR contracts, but with the prices offered in the FP of the menu being lower in scheme 2 than in 1, following the logic discussed after equation 15. How many blocks the contract should be split up into will obviously depend on the complexity of the contract desired.

In the example above, for payment scheme 1 to have a different amount of cost reimbursement (α) from payment scheme 2 would require that costs can be separated across the units produced. For instance, GF cannot offer cost-reimbursement for only units above q_1 if it cannot monitor whether costs were spent on the production of the first q_1 units or units after that. Whether or not this is possible will depend on the good produced. If it is a fairly simple

output, such as the training of healthcare professionals, it may be possible to attribute costs incurred to which professionals were trained. However, if the good is an outcome variable that is the result of a more complex process, such as the prevalence rate of a disease, such a decomposition is unlikely to be possible. In this scenario, the share of costs reimbursed will have to be constant across all quantities, though the fixed price can vary.

4.1.2 Multiple goods and variable quality

Suppose that GF wishes PR to produce quantities of two different products, q_1 and q_2 . It may be that the costs spent on producing the two goods are entirely separable - i.e. GF can identify that C_1 was spent on producing q_1 and C_2 was spent on producing q_2 . This would be the case, for instance, with the purchase of drugs, where antimalarial drugs are clearly only intended to treat malaria. In this situation, we can simply think of GF as having two contracts with PR, and the two can be decided upon entirely independently. For example, GF could use an FP contract for the first good, and a CR contract for the second.

The case becomes somewhat more complex if costs are not attributable across goods. Healthcare professionals, for instance, may work to treat both malaria and HIV/AIDS. In this case, as with the non-constant cost function considered, GF is obviously constrained to reimbursing a single fraction α of these non-attributable costs.

In practice, practitioners may attribute joint costs across goods through ‘step-down accounting’ or using arbitrary rules. However, this does not in reality allow for different fractions of the joint-costs to be reimbursed, as instead the cost-reimbursement will simply always be some average of the two fractions. For instance, suppose that half of the cost of healthcare professional is attributed to malaria, and half to HIV/AIDS. If malaria costs are reimbursed, but not HIV/AIDS costs, then in practice this is simply equivalent to a rule where half of malaria costs are reimbursed, and half of HIV/AIDS costs are reimbursed, and hence the effective α is the same across the two goods. If the proportion of joint costs is relatively small, it may therefore be simplest to divide these costs according to some rule across two different contracts.

If quality is verifiable by the Global Fund, the problem is essentially equivalent to the multiple goods scenario. For instance, suppose that the PR treats some number q of patients, and it can either treat them well or badly. Then we can consider two contracts, one for treating patients well and one for treating them badly. Most likely, costs will not be attributable, and hence GF will have to use the same α for both, though \underline{t} can vary.

4.1.3 Varying reimbursement across inputs

The basic model in the previous section essentially aggregates all the PR's cost into a single cost variable which the GF observes. In practice, it is normally possible to split the cost into various components, particularly if the output is complex. For instance, treating patients may involve building a clinic and paying staff to work in the clinic. In contracting on patient treatment, it may be useful for the GF to treat these two costs differently.

In order to understand when subcosts should be treated differently, recall from the previous section that a contract's suitability depended on the nature of the costs. If costs were highly uncertain, lower-powered schemes such as CR were more appropriate. On the other hand, if there was a large potential for cost-reduction, high-powered schemes such as FP were better. Since these are properties of the cost variable, it may be best to treat different costs separately.

For instance, if the costs of building a clinic are very uncertain and there is little potential for the PR to reduce costs, then it may be best for the GF to reimburse all of these costs. On the other hand, if staff costs are relatively well known and there is little potential for cost reduction, then the GF should simply pay a fixed price to cover this cost.

Examples of varying contracts across inputs can be seen in certain examples of healthcare provision. This is essentially what the Global Fund is doing in the Solomon Islands - it is paying the cost of certain items (e.g. drugs), but then for the remaining items it is paying a fixed amount linked to outputs rather than any costs.

A potential disadvantage with varying contracts across inputs is that it may induce the PR to inefficiently substitute between inputs. For instance, if costs of building clinics are reimbursed, but not staff costs, then the PR will be tempted to over-spend on the clinic if this can reduce staff costs - for instance, by installing expensive capital rather than simply employing an extra person. The GF should therefore be wary of varying contracts across inputs when those inputs are highly substitutable.

4.2 Determining quantities

The previous subsection outlined how a contract might depend on the quantity demanded by the global fund. This quantity is a further parameter that needs to be determined by the GF. To a certain extent, determining this quantity can be considered separately from the type of incentive scheme used, in the following way.

If the GF has budget flexibility, quantity should be determined such that the marginal value of an extra unit is equal to the marginal cost to the Global Fund. Since the Global

Fund's international budget is fixed, this marginal cost is essentially the opportunity cost of not spending the money elsewhere. This is clearly easier to set in the case of the FP contract than the CR contract, since in the latter case only the average cost is observed. The GF will then be forced to make some assumption as to how average cost relates to marginal cost.¹⁰

If the GF has no budget flexibility, then quantity will simply be determined by the budget. In this case quantity demanded will be that which produces $t = \bar{t}$, where \bar{t} is the budget allocated to this particular contract. The politics of the Global Fund may mean that the money available to a country has to remain constant, independent of the costs revealed. However, this may not necessarily be true within a country - for instance, it may be possible to give more money to combating Malaria, and less to HIV, depending on the cost effectiveness. This may thus allow for flexibility across different products or regions within a countries budget.

4.3 User fees and government resources

So far we have assumed that PR does not receive any revenue from users or the government for the goods that it provides. In some circumstances this assumption may not be valid - patients may be expected to contribute towards medical costs. In other circumstances, users may need to be subsidized. How would contracts need to change as a result?

Considering pricing is essentially the flip-side to considering quantity. If the market clears, quantity will be determined by price and quality - and hence, if the Global Fund can specify two of these, the third is a result. As with quantity, the pricing decision should be made independently of the power of incentives. The two problems are essentially separable for the Global Fund.

Once prices (or subsidies) have been chosen, this simply needs to be reflected in the total amount of revenue received by PR. Suppose, for example, that PR charges a price p for each

¹⁰Since the Global Fund faces uncertainty regarding the shape of the agent's cost function, one might be tempted to appeal to the use of something like a 'Vogelsang-Finsinger' mechanism, originally outlined in Vogelsang and Finsinger (1979). Such a mechanism devolves the setting of price and quantity to the agent under some constraints based on previous years' costs, with the attraction being that prices converge towards 'Ramsey pricing' even when the Global Fund knows nothing about the shape of the cost (or demand) function. Ramsey pricing is attractive because it is the optimal price that covers the firms cost when prices are constrained to be constant as a function of quantity produced. Such a constraint is very relevant in the case of regulated utilities which do not receive a transfer from government, because the firm receives its revenue from selling to many customers and therefore cannot change its price as a function of the quantity sold. However, it is not clear that this is very relevant for the situation the Global Fund finds itself in, as there is no reason why it should restrict itself to paying a constant price per unit produced - at the very least a lump sum payment (to cover fixed costs) in addition to a per-unit price should be possible.

product sold. Then the transfer from the Global Fund needs to be $t = \underline{t}p + \alpha C - pq$, with α set as before and \underline{t} adjusted appropriately.

One important exception to this rule is when quality may be unobserved by the Global Fund - this is considered in section 5.3.

A similar logic holds if the government has resources that it could also invest in the project GF would like to fund. If the government is willing to invest an amount, then this should be taken away from the amount reimbursed by the Global Fund.

4.4 Dynamic considerations

So far, we have considered the contract as completely static. In reality, GF contracts last for several years and may be updated over time. How does this change our previous analysis? In this subsection, we consider two dynamic considerations. First, how should the GF use any information it gathers? Second, if a menu of contracts is used, when does the PR need to decide which contract it chooses?

4.4.1 Updating with information

As outlined in the previous section, the key problem for the GF is that the PR knows more about its costs. Over time, the PR will learn more information. How should this change the contract?

If the GF initially offered a simple contract without a menu, then it should use any information it gathers to improve this contract. For instance, it may be that the GF initially offered cost-reimbursement because it was very uncertain about cost levels, but then garnered new information during this contract. At this point, GF may wish to switch to offering a FP contract, possibly as part of a menu.

However, if the initial contract that was offered was a menu, then, ex ante, it may be preferable for the GF to commit not to use this information. The logic is as follows. Consider a contract with two periods. If GF commits to not using this information in the second period, then we will have the same result as before - PR will choose the FP contract when $\beta \leq \underline{t}_1 + k$. However, if GF is going to use this information in the future, by, for instance, offering a lower value of \underline{t}_2 , then PR will not choose the FP contract when β is just below $\underline{t}_1 + k$. As a result, to get a similar division of β 's as before, GF will need to increase \underline{t}_1 .¹¹

¹¹This problem is known as the 'ratchet effect' - see Section 5.4 for more information.

It is worth pointing out that this logic only holds to the extent that PR is not myopic. If PR discounts the future heavily - because, for example, there is a large amount of rotation in the people staffing the bureaucracy - then this argument will not apply. In the extreme, where PR is completely myopic, GF can use any information it garners in period 1 without loss.

4.4.2 Timing of menu choice

If the GF opts to offer the PR a menu of contracts, then a practical question that arises is when the PR must decide between these contracts. In reality, it may be that as the contract is being written up, the PR is not aware of their costs, and that this only is revealed to it later. In this case, can menu choice be postponed?

For a menu of contracts to be effective, it is important that the PR makes it's choice after it has learnt about any information that will impact this choice. In theory, therefore, there is nothing preventing this decision from being made at the end of the contracts duration. In reality, such a contract might appear odd, since activities are typically pre-financed by the Global Fund. However, one could imagine an FPCR contract that pays the same amount of money but offers the PR a choice between either:

1. Showing that they spent the money producing some output $q < \underline{q}$
2. Show that they produced at least a target amount \underline{q}

In this way, there is no variation in the amount of money that the PR will receive. Instead, the PR faces a choice as to whether to show receipts or instead to reach a target. The advantage to the GF of this contract over a standard CR contract is that it increases the likelihood the target will be met, since the PR has an extra incentive to reach the target \underline{q} .

5 Adapting the contracts to context

The previous section described how, from the simple model of Section 2, we can build a complex contract that can cover the range of activities the GF might wish to include. However, we have so far abstracted from how such a contract might vary according to the context, except for variation in the extent that costs are uncertain or subject to reduction. In this section, we explore how various aspects of the country or sector context beyond the cost function might influence contract choice.

5.1 Unverifiable or mismeasured output

The basic model above assumed that output could be verified. This was an important assumption when it came to implementing any contract besides a CR contract, since the PR might otherwise overstate production in order to claim a higher transfer.¹² If instead output is not verified, it will not be possible for the GF to implement any contract besides CR.

An intermediate situation may occur when output can be verified but is measured with noise. To consider this situation, let us adapt our model as follows. Suppose that if PR produces a quantity q , GF observes a quantity $\tilde{q} = q + \epsilon$, where ϵ is some noise parameter with $\mathbb{E}[\epsilon] = 0$. As a result, rather than PR receiving a transfer $t = \underline{t}q + \alpha C$ as before, he receives a transfer $t = \underline{t}\tilde{q} + \alpha C$.

How does this affect PR's behaviour? The first point to note is that, for a given α , the amount of effort exerted reducing costs will not change. In particular, we will still have $\psi'(e) = 1 - \alpha$, independent of ϵ . Moreover, if PR is risk-neutral, no other aspect of his behavior will change, and hence everything will follow through as before. Hence, if PR is risk-neutral, GF can ignore potential noise when designing the contract.

If PR is risk-averse, however, noise may affect his decision as to whether to accept a contract. In particular, the fixed payment part that he gets will now be riskier, and hence less valuable. In a simple contract therefore, GF will need to increase \underline{t} in order to ensure it is accepted no matter what the value of β . This thus makes it more attractive for GF to use contracts with a lower power of incentives.

In terms of the optimal FPCR contract, let us suppose that PR discounts the fixed payment \underline{t} by an amount $\lambda < 1$ given that it is now risky. PR will choose the CR contract whenever $\lambda \underline{t} < \beta - e^* + \psi(e^*)$, and will pick the FP contract otherwise. Let $\beta^i = \lambda \underline{t} + e^* - \psi(e^*)$. Hence the expected cost to GF is

$$F(\beta^i)\underline{t} + (1 - F)F(\beta^i)\mathbb{E}[\beta|\beta \geq \beta^i] \quad (20)$$

Differentiating and setting to zero gives us a solution for β^i :

$$kf(\beta^i) = F(\beta^i) \quad (21)$$

¹²The indivisibility of the single unit of output in the basic model makes this point somewhat obscure, but this point is clear in the case where the PR produces multiple units - the PR would simply claim that he had produced the maximum number of units which would be consistent with his incurred costs.

Hence β^i is unchanged. The fixed payment \underline{t} must therefore be increased by a factor of $\frac{1}{\lambda}$ to compensate for the fact that this is now riskier.

5.2 Unverifiable inputs and cost-padding

We have so far assumed that the GF observes the cost incurred by the PR in producing the output. In reality, this may not be the case. Weak accounting systems may mean expenditures go unrecorded, or records can be created of expenditures that were not really made. Even if there is a close mapping between expenditures and receipts, it may be difficult for the GF to know whether a given input was used for the project or for something else. For instance, if a potential input is the purchasing of new vehicles, it may be difficult to verify whether the vehicles that were bought were simply inputs for the output the GF desires, or whether they were in fact inputs for something else. This problem is described as ‘cost-padding’ in the terminology of Laffont and Tirole (1993, Chapter 12). Cost-padding is typically prevented by auditing, which is a process verifying that claimed costs were indeed incurred for the stated project. However, auditing may be imperfect, particularly when there is the possibility of corruption. In such a situation, PR will face incentives to pad costs in order to increase the profit he makes from the contract with GF.

When it is difficult to verify that money has been spent on inputs, this pushes against using low powered contracts such as Cost Reimbursement. This is because it is necessary to verify what the money has been spent on in order to reimburse costs. The alternative, high powered contracts such as FP, can place more weight on the verification of outputs.

In some senses, the risks of cost-padding are similar to the potential for cost-reduction through effort. A key difference, however, is that PR will behave differently in the two cases when choosing from a menu. To see this, suppose that innate costs are β . Without the possibility of cost-padding, PR will prefer a FP contract with transfer $\underline{t} = \beta - k + \epsilon$ to a CR contract for any $\epsilon > 0$, since it allows PR to make a profit. However, if PR can pad costs up to some amount κ , it will prefer the CR contract as long as $\kappa > \epsilon$. GF will therefore have to offer a higher price in the FP contract in order to persuade PR to choose the FP contract over the CR contract.

The risk of cost-padding therefore provides a strong reason not to provide a menu with a CR option in it. If the Global Fund is convinced that cost-padding may be a serious problem, it may therefore be optimal to force the PR to accept a higher-powered incentive contract. This may well implicitly be behind the Global Fund’s logic in offering Rwanda a choice in the

contract used, but not the Solomon Islands.

5.3 Unverifiable quality

Consider a scenario where quality is not verifiable, or is expensive to verify. Typically, this will lead to GF preferring to use lower-powered incentives. This is because PR now has two potential ways to reduce costs. One is to increase cost-reducing effort, which is what GF would like PR to do. The other, however, is to reduce quality. The higher the incentives PR has to reduce costs, the more he will be tempted to cut back on quality. If quality is not verifiable, GF is unable to prevent PR from doing this.

Suppose, however, that even though quality cannot be measured by GF, it can be measured by service users. This may be more pertinent for aspects of amenity quality (e.g. was the doctor there when they should be) than aspects of technical quality (e.g. did the doctor prescribe the right drugs), since the latter aspects are typically difficult for patients to ascertain. If service users observe quality, it is likely that quality will affect user demand for the product, particularly if users can find an alternative supplier. If the quantity consumed is partly demand driven, then GF can implicitly measure quality through monitoring consumption. GF can then incentivize PR to produce higher quality by providing bonuses for producing higher quantity (at a given price). In this way, GF may still be able to use relatively high-powered incentives even when she does not observe quality directly.

5.4 Limited commitment

5.4.1 Limited commitment to pay

We have so far abstracted from when the transfer is made to the PR relative to when the costs are incurred. Since the transfer is dependent on the costs incurred and/or the quantity produced, the most natural ordering would perhaps be for the GF to make the transfer after they have received this info. However, typically in practice the GF makes transfers to the PR before costs are incurred, as the PR is likely to be credit constrained. Then, if part of the money is deemed to have not been spent appropriately, it is ‘clawed back’ by the GF, typically through making smaller future transfers.

If there was no problem for the GF to reclaim money from the PR ex-post, the timing of the transfer would be irrelevant. In practice, however, this process is often difficult, since the GF has little power to force the PR to make transfers. Moreover, since discrepancies may

arise for innocent reasons - i.e. receipts are lost, or output is mismeasured - the GF may be tempted to give PR ‘the benefit of the doubt’.

Clearly if the PR anticipates that it will be able to keep money to which it is not entitled this will alter his behaviour ex-ante. If he is under a CR contract, and he believes he can keep a transfer t whilst only showing receipts for a share θt , he may be tempted only to spend θt on the project. If he is under a FP contract, and he believes he can keep a transfer t whilst only producing a share θ of the target quantity, then he may aim to only produce this smaller share. In each case, therefore, production will be reduced by the amount of ‘wiggle room’ the PR presumes he has.

Without building a full model of the process through which the GF decides how much the PR can keep, it is difficult to see whether one type of contract may be more vulnerable to problems of paying funds up-front. An important factor is likely to be how much potential error lies in the verification of the object upon which the transfers are conditional. If the PR provides receipts for costs significantly below costs claimed, this may be more likely tolerated by the GF when she believes there are large problems in providing receipts. Equally, the GF may tolerate reported production below target when there is a lot of noise in the monitoring of production. Overall, therefore, if the GF must finance the project ex-ante, this is another reason why contracts should be dependent on the component that can be most precisely verified.

A further factor to consider may be potential errors in calculation when dealing with contracts that depend on output. Even if output can be perfectly measured, the PR may miss a target if costs were above expectation. This cannot arise in the model above because we assume that the PR knows their cost when they accept the contract. In practice, however, they may make a mistake and accept a contract that only pays conditional on making a target, only to find out later that such a target was not possible. In this circumstance, the GF may be tempted to let the PR keep the money anyway. As a result, higher-powered incentive contracts may be more vulnerable to problems related to ex-ante financing.

5.4.2 Limited commitment to not use information

In Section 4.4.1 we discussed that the GF should sometimes commit not to change the contract based on information revealed when initially offering the PR a menu. It may, however, be impossible for GF to make such a commitment. Since GF is not technically writing contracts that are enforceable in a court of law, it may be unable to tie its hands in this way. Equally,

contracting costs may mean it can only write a contract for a few years in advance. Knowing this, how should it change the incentive scheme offered in period 1?

This problem is often described as the *ratchet effect*. Once GF learns about PR's innate costs, it can and, due to pressures from its donor constituents, will use this information in period 2 to make a lower transfer. Hence, by exerting more effort to reduce costs now in order to keep a residual payment or earn a bonus, PR knows that it's also going to be forced to make more effort in the future, in order to avoid financial loss. As a result, the benefit of high-powered contracts in period 1 will be reduced, and GF may be tempted to use lower-powered contracts. In an FPCR contract or LCSCR contract, GF will have to offer a higher price in order to persuade PR to accept the higher powered contract.¹³

An alternative problem that may be faced is if bilateral renegotiation can take place - that is, a change in the contract which benefits both parties. In the second period, it will be optimal for PR and GF to switch to a FP contract, whereby gains from cost reduction are made and the resulting benefits can be shared between the two parties. However, anticipating that an FP contract will take place in the second period, PR now has an incentive to increase it's realized costs in the first period. Again, therefore, high powered incentives in the first period will be less effective.

Glassman et al. (2013, pp. 61-62) argue that these commitment problems may not be very serious in practice, because PRs are likely to behave in ways that maximize their short-term payoffs, without too much regard for future contracts.

5.5 Valuing of 'profit'

We have so far treated the PR as equivalent to a firm. The difference between GF's transfers and costs are marked as 'profits' that are of no direct value to the Global Fund. Alternatively, suppose that the surplus $t - C$, which we have so far discussed as profit, goes towards some activity valued by the Global Fund. This may be the case if, for instance, it goes into the general budget of the Ministry of Health and is spent on other activities that improve healthcare. In this case, GF will be much less concerned with reducing the 'information rent', and as a result should offer higher powered contracts.

For surplus to be valuable to the GF in this way, it has to be that any marginal increase in

¹³Note that this doesn't follow directly from the model laid out in this paper, since if PR chooses a CR contract he makes no effort and hence GF will learn the true innate cost. However, the logic would apply if there was uncertainty about the amount of potential cost reduction or cost-padding, and hence it would be straightforward to extend the model in this way.

profits results in a marginal net increase in spending on activities the GF cares about. This is most likely to take place if the PR shares similar objectives to the GF. If the PR would rather spend this surplus on other activities, a requirement that it is spent on healthcare may have little impact if budgets are fungible. Moreover, if the PR is forced to spend the money on an activity which it values relatively little, this reduces their incentives to accrue surplus and hence make cost savings. Such a requirement may therefore undermine the incentive effects of high-powered contracts.

6 Conclusions

In this paper, we have laid out some simple contracts that can be used by the Global Fund to incentivize cost reduction in a context of asymmetric information. We have argued that perhaps the best contract to use in practice may be an FPCR contract, whereby the PR is offered the choice between cost-recovery (i.e. the status quo) and a fixed transfer - i.e. payment on production rather than cost. We have then described how these simple contracts can be part of more complex ones, and when situations may favour different contracts.

A selection of the results of these comparisons is given in Table 6. The ‘power’ column describes how more of a particular factor effects the power of incentives the Global Fund should choose. So, for instance, higher cost uncertainty or PR with objectives closer to those of the GF means that a fixed price contract is more likely to be preferred to a cost-recovery contract. The ‘price’ column then describes how each factor affects the optimal fixed transfer to set in a FPCR or LCSCR contract. Hence, higher cost uncertainty means that the fixed transfer should be lower, whilst a PR with objectives closer to the GF’s means that it should be higher. Finally, the ‘section’ column references the section of this paper in which the reader can find the result.

In general, we have seen that an FPCR contract can be adapted to many situations through adjustment of the price offered. However, in two situations we have seen an FPCR contract may perform badly. First, when innate costs are more likely to be towards the higher end of the cost distribution, the cost-recovery option is generally chosen and hence little cost-reduction is achieved. In this case, a linear cost sharing contract, preferably combined in a menu with a cost recovery option, can perform better. Second, when weak auditing or corruption may allow the PR to lie about costs incurred, it may be best not to give the PR the option of cost-recovery. In this case, a fixed price contract or linear-cost sharing with a small share of

	Power	Price	Section
Cost uncertainty	↓	↓	3.6
Potential cost savings	↑	(↓)	3.6
Output mismeasurement	↓	↑	5.1
Unverifiable inputs and cost-padding	↑	↑	5.2
Unverifiable quality	↓	↓	5.3
Limited commitment	↓	↑	5.4
PR shares GF objectives	↑	↑	5.5

Table 1: Summary of factors

costs reimbursed will be preferable.

An important point to bear in mind is that there is no reason for the Global Fund to use the same type of contract for all activities within a given country. For some activities, it may be most appropriate to remain with a cost-reimbursement style strategy. For others, the Global Fund may wish to offer a fixed price contract and not require evidence of how much money was spent. Even within a single activity, the Global Fund could choose to reimburse some costs based on PR expenditure, and other costs based on a fixed price. The PR could be offered a menu of contracts (i.e. FPCR) for each activity, and then decide differently depending on the activity.

A second point to note is that offering a menu does not imply that the PR must make the choice when the contract is written up, or that pre-financing is not possible. Currently, a typical contract sets a target outcome, and then reimburses all the PR's costs that are made attempting to reach this outcome, up to a fixed amount. The PR could be offered a choice between this contract, and one where he simply receives the fixed amount on condition that he reaches the target. The PR would not need to decide until he knows whether or not he has reached the target, and would therefore not bear any extra risk - he could always take the business as usual option. But the possibility of receiving the money without strings attached may make him more likely to reach the target, particularly if he anticipates cost savings that can be made along the way.

References

Azam, Jean-Paul and Jean-Jacques Laffont, "Contracting for aid," *Journal of development economics*, 2003, 70 (1), 25–58.

- Birdsall, Nancy, William D Savedoff, Ayah Mahgoub, and Katherine Vyborny,** *Cash on delivery: a new approach to foreign aid*, CGD Books, 2010.
- Chalkley, Martin and James M Malcomson,** “Contracting for health services when patient demand does not reflect quality,” *Journal of health economics*, 1998, 17 (1), 1–19.
- Chu, Leon Yang and David EM Sappington,** “Simple cost-sharing contracts,” *The American Economic Review*, 2007, 97 (1), 419–428.
- Clist, Paul and Arjan Verschoor,** “The conceptual basis of payment by results,” *School of International Development, University of East Anglia*. http://r4d.dfid.gov.uk/pdf/outputs/misc_infocomm/61214-The_Conceptual_Basis_of_Payment_by_Results_FinalReport_P1.pdf, 2014.
- Cordella, Tito and Giovanni Dell’Ariccia,** “Budget Support Versus Project Aid: A Theoretical Appraisal*,” *The Economic Journal*, 2007, 117 (523), 1260–1279.
- Garrett, Daniel F.,** “Robustness of simple menus of contracts in cost-based procurement,” *Games and Economic Behavior*, 2014, 87 (0), 631 – 641.
- Gasmi, F., J. J. Laffont, and W. W. Sharkey,** “Empirical Evaluation of Regulatory Regimes in Local Telecommunications Markets,” *Journal of Economics & Management Strategy*, 1999, 8 (1), 61–93.
- Geruso, Michael and Thomas G. McGuire,** “Tradeoffs in the Design of Health Plan Payment Systems: Fit, Power and Balance,” Working Paper 20359, National Bureau of Economic Research July 2014.
- Glassman, Amanda, Mead Over, and Victoria Fan,** *More Health for the Money: Putting Incentives to Work for the Global Fund and Its Partners*, CGD Books, 2013.
- Laffont, Jean-Jacques and Jean Tirole,** “Using Cost Observation to Regulate Firms,” *Journal of Political Economy*, 1986, 94 (3), pp. 614–641.
- and –, *A theory of incentives in procurement and regulation*, Cambridge, Mass ; London: MIT Press, 1993.
- Marseille, Elliot, Mark J. Giganti, Albert Mwango, Angela Chisembele-Taylor, Lloyd Mulenga, Mead Over, James G. Kahn, and Jeffrey S. A. Stringer,** “Taking

- ART to Scale: Determinants of the Cost and Cost-Effectiveness of Antiretroviral Therapy in 45 Clinical Sites in Zambia,” *PLoS ONE*, 12 2012, 7 (12), e51993.
- Meyer-Rath, Gesine and Mead Over**, “HIV Treatment as Prevention: Modelling the Cost of Antiretroviral Treatment—State of the Art and Future Directions,” *PLoS Med*, 07 2012, 9 (7), e1001247.
- Reichelstein, Stefan**, “Constructing incentive schemes for government contracts: An application of agency theory,” *Accounting Review*, 1992, pp. 712–731.
- Rogerson, William P**, “Simple menus of contracts in cost-based procurement and regulation,” *American Economic Review*, 2003, pp. 919–926.
- Silverman, Rachel, Mead Over, and Sebastian Bauhoff**, *Aligning Incentives, Accelerating Impact: Next Generation Financing Models for Global Health*, CGD Books, 2015.
- Tagar, Elya, Maaya Sundaram, Kate Condliffe, Blackson Matatiyo, Frank Chimbwandira, Ben Chilima, Robert Mwanamanga, Crispin Moyo, Bona Mukosha Chitah, Jean Pierre Nyemazi, Yibeltal Assefa, Yogan Pillay, Sam Mayer, Lauren Shear, Mary Dain, Raphael Hurley, Ritu Kumar, Thomas McCarthy, Parul Batra, Dan Gwinnell, Samantha Diamond, and Mead Over**, “Multi-Country Analysis of Treatment Costs for HIV/AIDS (MATCH): Facility-Level ART Unit Cost Analysis in Ethiopia, Malawi, Rwanda, South Africa and Zambia,” *PLoS ONE*, 11 2014, 9 (11), e108304.
- Vogelsang, Ingo and Jörg Finsinger**, “A regulatory adjustment process for optimal pricing by multiproduct monopoly firms,” *The Bell journal of economics*, 1979, pp. 157–171.
- Wilson, Robert**, “Efficient and Competitive Rationing,” *Econometrica*, 1989, 57 (1), pp. 1–40.