Social Science & Medicine 75 (2012) 377–383

Contents lists available at SciVerse ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Why performance-based contracting failed in Uganda – An “open-box” evaluation of a complex health system intervention

Freddie Ssengooba [a,b,*], Barbara McPake [c], Natasha Palmer [b]

[a] Department of Health Policy, Planning and Management, School of Public Health, Makerere University, Uganda
[b] Department of Global Health and Development, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
[c] Institute for International Health and Development, Queen Margaret University, Edinburgh, Musselburgh EH21 6UU, UK

ARTICLE INFO

Article history:
Available online 20 April 2012

Keywords:
Performance-based contracting
Health program evaluation
Health system
Case study
Uganda

ABSTRACT

Performance-based contracting (PBC) is a tool that links rewards to attainment of measurable performance targets. Significant problems remain in the methods used to evaluate this tool. The primary focus of evaluations on the effects of PBC (black-box) and less attention to how these effects arise (open-box) generates suboptimal policy learning. A black-box impact evaluation of PBC pilot by the Development Research Group of the World Bank (DRG) and the Ministry of Health (MOH) concluded that PBC was ineffective.

This paper reports a theory-based case study intended to clarify how and why PBC failed to achieve its objectives. To explain the observed PBC implementation and responses of participants, this case study employed two related theories i.e. complex adaptive system and expectancy theory respectively.

A prospective study trailed the implementation of PBC (2003–2006) while collecting experiences of participants at district and hospital levels.

Significant problems were encountered in the implementation of PBC that reflected its inadequate design. As problems were encountered, hasty adaptations resulted in a *de facto* intervention distinct from the one implied at the design stage. For example, inadequate time was allowed for the selection of service targets by the health centres yet they got ‘locked-in’ to these poor choices. The learning curve and workload among performance auditors weakened the validity of audit results. Above all, financial shortfalls led to delays, short-cuts and uncertainty about the size and payment of bonuses.

The lesson for those intending to implement similar interventions is that PBC should not be attempted ‘on the cheap’. It requires a plan to boost local institutional and technical capacities of implementers. It also requires careful consideration of the responses of multiple actors – both insiders and outsiders to the intended change process. Given the costs and complexity of PBC implementation, strengthening conventional approaches that are better attuned to low income contexts (financing resource inputs and systems management) remains a viable policy option towards improving health service delivery.

© 2012 Elsevier Ltd. All rights reserved.

Introduction

Performance-based contracting (PBC) has been promoted in both high and low income health systems as a means of delivering improved performance from providers – both institutions and individual health workers, and hence better outcomes from health service provision (Conservatives.com, 2011; Loevinsohn & Harding, 2005).

PBC is a tool that links monetary or material rewards to measurable actions or achievements in relation to predetermined performance targets (Eichler & Levine, 2009 page 6). For example, bonus payments are expected to improve performance by tying some or all of agents’ compensation to the achievement of specified objectives (Mooney & Ryan, 1993). In comparison to traditional public management approaches that finance resource inputs and systems management, tools like PBC are hypothesised to give more flexibility to the agent to achieve the desired results by allowing for creativity and innovation in the production of results (Loevinsohn & Harding, 2005).

Whereas studies have been undertaken to test the effectiveness of PBC in different contexts (Basinga et al. 2011; Lundberg, Marek, &

* Corresponding author. Suite 323 MakSPH Building, Mulago Hospital Complex, P.O. Box 7072, Kampala, Uganda. Tel.: +256 414 543 872; fax: +256 414 540633.
E-mail addresses: Sengooba@musph.ac.ug, freddie.ssengooba@gmail.com (F. Ssengooba).

0277-9536/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.socscimed.2012.02.050

Pariyo, 2007), the evidence suffers from methodological problems prompting scholars such as Galvin (2006) to conclude that "*the methodologies used in these evaluations along with the overwhelming preponderance of positive results, makes it difficult to distinguish proven findings from enthusiasm and marketing*"(page 126). Similarly, Eichler et al. (2009) observe that evaluation studies have lacked a comprehensive assessment of the causes of poor performance. The inadequacy of impact evaluations has become increasingly evident, and clients of these studies – such as decision-makers at national and global level have particularly criticised the blinded or black-box approach used in generating results (Astbury & Leeuw, 2010; Eldridge & Palmer, 2009; Kalk, 2008; Kelly, McDaid, Ludbrook, & Powell, 2005). Black-box evaluations here refer to evaluation approaches that primarily focus on the magnitude of effects from public health interventions in contrast to explaining how and why effects come about (open-box). The black-box approach provides inadequate advice as to how to design and implement similar programs successfully in other settings (Astbury & Leeuw, 2010; Pawson & Tilley, 2004).

The main criticisms of available evidence in relation to PBC schemes target three major weaknesses: 1) limited consideration of the broader health system, 2) unrealistic assumptions that implementation fits a linear program logic, and 3) unrealistic assumptions such as short and simple causal chains between the intervention and its effects.

The purpose of this paper is to increase understanding of a PBC intervention, its mechanisms and potential effects. The specific objectives of the paper are: 1) to outline an evaluation approach that supports the design and implementation of PBC and 2) to illustrate the empirical dynamics in PBC implementation. The paper starts with a brief overview of the PBC pilot in Uganda and concludes with some guidance for practitioners and policy makers who seek to expand PBC approaches within dynamic health systems. This paper draws from the doctoral research of the first author (Ssengooba, 2010).

PBC pilot design in Uganda

Between 2003 and 2006, a PBC pilot was designed, implemented and its impact evaluated by the World Bank's Development Research Group (DRG) in partnership with the Uganda Ministry of Health (MOH). The primary aim of the pilot was to assess whether performance-based contracts with bonus payments would induce higher service production (performance) for a pre-defined set of health services (Lundberg et al., 2007; Morgan, 2010) among private-not-for-profit (PNFP) health centres. Performance was defined as a percentage increase in the selected service outputs for a calendar period relative to outputs for a similar period in the previous year. Service targets included increases in: 1) outpatient visits (10%); 2) child births in a health facility (5%); 3) children fully immunized at 12 months from birth (10%); 4) new acceptors of modern family planning methods (5%); 5) pregnant women attending antenatal care (10%) and 6) children under-5 years treated for malaria (10%). The participating health centres were required to choose only three service targets from the 6-item menu above.

Employing an orthodox black-box design, the participating health centres were randomly assigned to one of three groups:

1. The control group (public and PNFP centres) continued with the pre-existing arrangements – these received government grants as before and used the grant according to the usual prescriptive guidelines of Government.
2. The autonomy group (all PNFP) received grants as before but prescriptive guidelines about the internal allocation of grants were suspended as a way to widen the decision space of managers.
3. The bonus group (all PNFP) received the grant as before, operated under the same conditions as the autonomy group but members of the group were eligible to receive a monetary bonus if they achieved or exceeded their contract targets. The maximum bonus was equivalent to 11 percent of the total annual grant amount provided by the government.

The pilot covered five (out of 56) districts of Uganda at the time and 120 health centres participated. The service outputs of these centres were measured every six months for three years. More details about the pilot are published in (Lundberg, 2010; Lundberg et al., 2007; Morgan, 2010).

Pilot evaluation

The black-box evaluation published in 2007 used standard statistical techniques – difference-in-difference analysis, to compare changes in service outputs between the three groups and concluded that:

> *[…] The performance-based bonus scheme did not have a discernible impact on the production of health care services provided by PNFP facilities. […] it appears that granting autonomy in financial decision-making had a positive impact on health care production (Lundberg et al., 2007 page 2)."* The evaluation was unable to provide explanations for these results but authors offered some speculations about possible reasons for pilot failure:

> *"Why has the performance bonus not worked? One can imagine a number of possible explanations. First, perhaps the performance bonuses were not large enough. […] Second, the performance bonus was paid to the facility and not to the individual providers directly. […] Third, it is possible that the performance-based contract was too difficult to manage. […] Finally, it is possible that the experiment had not had long enough time to take effect. (Lundberg et al., 2007 page 31)."*

These points are entirely based on speculation, illustrating the problems of orthodox evaluation approaches. The evidence collected in our own evaluation shows that none of them captures the underlying causes of failure, although we provide some support for the difficulty of managing the PBC pilot.

Methods

Our evaluation used an "open-box" approach to provide explanations of the PBC effects.

A theory-based and prospective case study design was used. The study was embedded in the concurrent implementation of the pilot and its processes, with the aim of opening the black-box and building plausible explanations derived from empirical observations. We used multi-disciplinary methods, grounded in systems thinking as advised by several authors (Annerstedt, 2010; De Savigny & Taghreed, 2009; Rosenfield, 1992). Our approach was a 'theory-based evaluation' (Grol, Bosch, Hulscher, Eccles, & Wensing, 2007).

We identified expectancy and complexity theories as promising approaches to address the main domains in PBC (see Table 1). Complexity theory predicts nonlinear routes to program impact and highlights the evolution and adaptation of the intervention to its context (Leykum et al., 2007; Plsek & Wilson, 2001). Following from this prediction, we examined how the implementation of the pilot was affected by the evolving responses of pilot participants and external dynamics like financial flows. Although many theories

Table 1
Selected theories and their use in guiding research questions.

Theory lens	Key tenets of theory	Research questions
1. Theory of complex adaptive systems (Leykum et al., 2007; Plsek & Wilson, 2001)	Non-linearity of implementation activities due to adaptability of agents/actors based on: 1) Experiences/learning; 2) Context/external factors; 3) Inter-dependencies.	Identification of the gap between *de jure* and *de facto* intervention: 1) What historical events/precedents are important for pilot implementation? 2) What changes to implementation plans were made and why? 3) What emergent behaviours or adaptations took place? 4) What role did the context or external factors to the pilot have?
2. Expectancy theory (Lawler, 1971; Lawler, 1989)	Essential perceptions needed to link the pilot bonus to actions required of hospital staff to improve performance: 1) Perceptions of internal capacity to improve performance; 2) Perceptions linking efforts and bonus/rewards; 3) Perceptions about performance auditing system.	Explore changes in perceptions about enablers and constraints to achieving performance targets; Explore the explanations of participants regarding issues of: 1) How adequate is the capacity of participants to achieve targets? 2) How valuable is the bonus to them? 3) How well is performance audited? 4) How well is communication shared?

seek to identify the links between a bonus and the motivation to perform a task, expectancy theory (Lawler, 1971, 1989) has been widely used to explain the underlying mechanisms that link external incentives to pre-specified tasks (Bokhour et al., 2006; Galvin, 2006; Teleki, Damberg, Pham, & Berry, 2006). Expectancy theory suggests that agents must hold at least four beliefs if a promise of a bonus is to spur performance improvements (see Table 1). Inspired by this theory, our methods included an assessment of the communication of pilot expectations and performance feedback – both necessary to understand the presence or not of the required beliefs.

Data collection and analysis

This case study focused on the implementation activities of the pilot among all the seven hospitals that participated – two of these eligible for the bonus payments (intervention arm). Participant observations during PBC implementation activities and in-depth interviews with pilot participants were used as research strategies to understand the pilot design, decision processes and the rationale for participants' responses. Study participants were purposively selected on the basis of direct involvement in the pilot. These included hospital managers, district health directors, pilot implementers and performance auditors. The findings in this paper are extracted from a total of 28 in-depth interviews with these groups. The first set of interviews in mid 2005 (see Fig. 1) covered three management team members in each of two hospitals in the intervention arm of the original study and with one hospital manager in each of five hospital in the control arm of that study. In the two intervention hospitals, 11 follow up interviews (2 sets each after 6–8 months) were done with hospital managers, or their replacement in the same post for the period April 2005–June 2006. These were typically interviewed 2–3 months after their performance audit, feedback and bonus payment events. One interview each was also done with a pilot implementer and an auditor in mid 2005 and repeated with the same persons 12 months later. Two District Health Officers (out of 5) that hosted the original study were also interviewed towards the end of the pilot in June 2006.

Interview questions explored in-depth how and why the participants were reacting to targets, performance auditing, feedback and bonuses. Additional qualitative data were collected through participant observations at four pilot-related meetings and from the review of quarterly progress reports for the period January 2005 to August 2006. Fig. 1 shows the time sequence of the main pilot events (upper side) and the timing for research events for this study (lower side) between 2003 and 2006.

Thematic analysis was undertaken by using a deductive analytical template derived from complexity and expectancy theories to construct plausible explanations of participant responses to the pilot. Descriptions of the implementation of the pilot and plausible explanations of actions and reactions by pilot participants were built from the data by combining findings about the context, participant actions and the underlying mechanisms or explanations of behaviour (Pawson & Tilley, 2004). Repeated

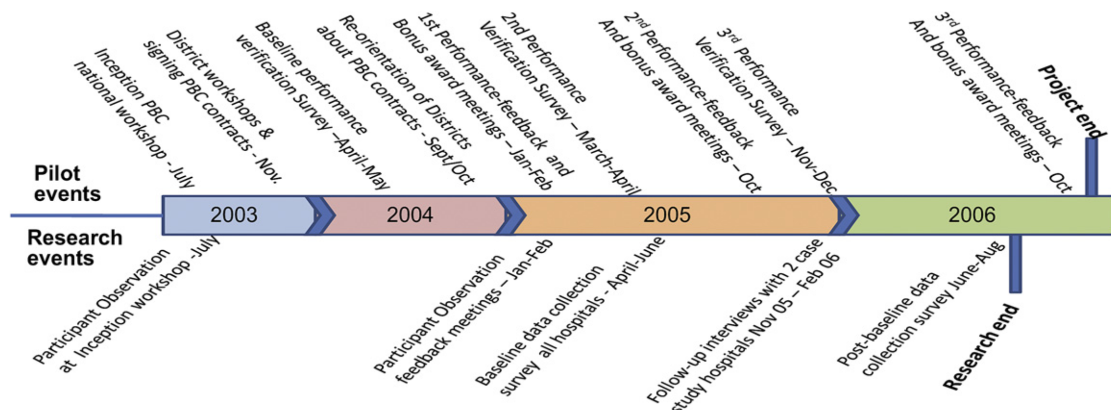

Fig. 1. Timelines for PBC pilot and research events for the period 2003 to 2006.

interviews with the participants enabled the research team to refine or validate the explanations of the mechanism by respondents. Ethical approval for the main study from which this paper is derived (Ssengooba, 2010) was obtained from 1) London School of Hygiene and Tropical Medicine, 2) Makerere University School of Public Health and 3) the Uganda National Council of Science and Technology.

Results

Four themes are used here to illustrate the main findings relevant to how and why the PBC pilot generated its effects. For each of these themes, the descriptive explanations of the mechanisms address the pilot as planned and designed (*de jure*), the actual enactments of pilot activities in practice (*de facto*) and the implications of the differences between these two for the pilot effectiveness.

The selection of targets and start-up activities

Expectancy theory suggests that hospital teams – who were ultimately responsible for the achievement of service targets, would require a good understanding of the pilot project, what they might gain from it, and how they could achieve the gains possible. The *de jure* design of the pilot reflected this, and required wide participation and communication of pilot expectations to hospital managers. The selection of three service targets out of a 6-item menu represented a strategic decision for bonus eligible health centres. An appropriate approach to this strategic choice would involve analysis of prior trends in service outputs, community needs, alternative providers and the assessment of centres' capacity to succeed in addressing the selected targets (Gibson, Martin, & Singer, 2005).

In reality (*de facto*), two significant events resulted in a problematic start. Firstly, performance targets were selected hastily during the district meetings organized to generate pilot awareness in 2003. Without prior warning about the pilot or the requirement to select service targets, the invited managers from each health centre were asked to select the targets and sign a contractual agreement for the pilot. Secondly, pilot activities stalled for nearly one year soon after the performance contracts were signed. Progress reports in the early phase of the pilot indicated that the stalling of activities was partly due to delays in the disbursement of funds from the World Bank to implementers in Uganda. The implication of these two events was inadequate communication of pilot expectations within the participating centres and a loss of institutional memory about the pilot contracts. When pilot activities resumed in late 2004, some of persons who had selected targets had forgotten them and some had left their posts without sharing sufficient information about the pilot or its contracts.

> *"[…] Back then, myself and a doctor that used to work here, we both attended the meeting at the district and we chose these targets. […] All along I had kept silent about this (PBC pilot) but when the bonus came, I reminded the staff about the study and announced the bonus." (Hospital Manager, Facility 01)*

> *"… In other cases, the managers were new. They did not know and did not add any extra effort. In some units, we found new management teams with no information at all about the pilot. […] I found out that some people who had been sensitized before (in 2003) had moved out of the health units. Therefore, fresh sensitization was done" (Member, Pilot Implementation Team).*

Partial ignorance and inconsistency about the selected targets remained a problem among those responsible for the achievement of pilot targets. When asked a direct question such as *"which service targets did your hospital choose?"* most hospital managers tried to search for and read their answers from their copies of the contract. In one instance (see extract below) the manager discovered during our interview (mid 2005) that the service targets in the contract were different from what he thought they were.

> *"We selected […] I think, outpatient visits – (looks up the file and reads from file) ooh no!, […] (smile of surprise) I wish we had selected outpatient visits. […] Back then, we selected (maternal) deliveries, immunization and malaria treatment for under-5s (children). […] I wish we had selected outpatient visits, deliveries and immunization of under 5s" (Hospital Manager, facility 01).*

After a lack of response to pilot targets during the initial period (2003–2005), all participating health centres were advised by the pilot leaders at Ministry of Health (MOH) to retain the same service targets for the second year (2005–2006). In effect, the participating centres were locked-in to the choices they made in haste and whose context had changed significantly. For example, between August 2003 when the first contracts were signed and mid 2005, when the second contracts were due, community-based programmes to treat malaria at the community level had been initiated by the MOH across the country. The distribution of mosquito nets had been expanded. These prevention programmes were led and supported by the same health centres that had elected to increase facility-based malaria treatment as one of the contract targets. When asked about how they planned to achieve the pilot targets within the new context, a manager outlined the conflicts in the performance expectations:

> *" […] for the one of malaria (target), I am not so sure (of success). We have […] (stops and laughs). There are many things we are improving down there in the communities. […] For malaria, there is this home-based management of fevers (a new programme) that we did not factor in (at the start of pilot). We thought the malaria will always be there but it was not to be. So, I really don't know how we can treat 10 percent more malaria at this hospital" (Hospital Manager, Facility 06).*

We conclude from these findings that health centres were given insufficient opportunity to select service targets strategically and this, combined with a dormant period that created individual and institutional memory loss about the pilot resulted in a lack of purposeful actions to achieve the pilot targets. Lack of success in relation to the malaria treatment target was specifically explained by parallel interventions that reduced malaria incidence and recourse to the health centres for treatment.

Performance audit

Service outputs were validated every six months by a performance audit. According to the *de jure* design performance audit should have played a vital role in supporting positive response to the pilot. A key tenet of expectancy theory is that valid measurement of performance and the provision of corresponding feedback is a pre-condition for a response involving increased effort. Performance audit may also serve to encourage improved keeping of performance records (Goddard, Mannion, & Smith, 2000). According to the pilot design, performance was to be audited using aggregated service output data reported every month to the district authorities and to the Ministry of Health.

De facto, auditing activities were modified in several ways, partly to circumvent the potential for data manipulation in the aggregated monthly reports of service outputs. Our findings show that:

a) Instead of using the aggregated monthly reports, primary clinical registers were used for performance audit. Primary registers are multi-purpose records at health centres in which clinical details of individual clients are recorded on a daily basis. The extraction of the three contract-relevant service outputs from multi-purpose primary registers escalated the workload for the auditors. They also relied on their understanding of clinical recording practices – an area where they had little competence. Although they received some training, they did not have clinical experience that would have allowed comprehension of the shorthand and recording practices in the primary registers. Several narratives indicate the evolving competence of the auditors in this respect.

"Capacity of the (auditors) to make correct recording improved the numbers (service outputs) in the second (audit) round. For some health centres, it was improving their handwriting so that we (audit team) can understand the diagnosis for tallying. [...] Clinicians use a lot of shorthand in the registers. Some used "PF" (plasmodium falciparum) for a diagnosis of malaria. But now we (audit team) are more used to these abbreviations" (Member, Performance Audit Team (February 2006)).

b) Although not essential for the pilot, the auditing team was concurrently tasked with collecting data for the impact evaluation of the pilot. For the impact evaluation, 39 different service outputs were extracted from different hospital registers (World Bank & Ministry of Health, 2004 page 3&4) at the same time as the performance audit for the three contract-relevant services. From a practical standpoint, 2–3 auditors were required to look for and manually tally 42 different service categories from typically 13,000 to 16,000 clinical entries for each audit period for hospitals. Activities for impact assessment also included surveys of clients (at exit) and households, and the collection of facility-level data about medicine stocks, staffing and equipment.

"Overall the most problematic data collection was the facility service module. [...] The workload of abstracting data from the registers consumed a lot of time; hospitals took 3 days instead of one day while small facilities took 2 days. The workload at the facilities required an extra assistant to cover exit interviews." (Pilot Update Report June 2004).

The two main implications from these findings about performance auditing were:

- The heavy workload for extracting data for the impact evaluation over-shadowed the focus of the audit team on the three services in the pilot contract.
- The reliability of measured outputs could not be trusted. The auditors' capacity to identify and tally contract-relevant services improved over time – a situation that would bias the audited changes in performance.

We conclude that changes in performance observed in audited activities arose partly from the improved capacity of the auditing team to extract data from clinical registers. Where positive change was measured, this partly arose from the auditors' growth in competence.

Performance feedback

In principle, and according to expectancy theory, provision of feedback to the health centres about their level of achievement motivates them to improve or sustain effort. On its own, performance feedback can be effective in improving providers' performance (McNamara, 2006; World Bank, 2003). As expected from complexity theory, several adaptations to providing feedback during the pilot life were observed:

a) Engaging stakeholders in performance feedback

With the intention of improving effective performance accountability and generating district-level leadership for performance management, district political leaders and officials in charge of planning, finance, administration and health departments were invited to pilot feedback meetings early in 2005, and attended in significant numbers. Performance audit results for each centre were presented at the meeting. However, this practice of inviting district leaders to the pilot feedback meetings was suspended for the rest of the pilot life due to budget constraints.

b) Responses in the control group

Observation of feedback meetings illustrated some unintended consequences in the pilot mechanisms. Performance feedback was provided at a district-level meeting where all groups participating in the pilot were invited. Managers from successful health centres belonging to the bonus group were congratulated by the chief guest (district leader) who also handed over a financial bonus in the form of a bank cheque ranging from $150 to $9000 for small and large health centres respectively. During the same meeting, control group participants who had also achieved targets were simply congratulated by handshake, consistent with the study design's randomization where controls do not receive bonus payments. Control group members contested the design and considered the bonuses unfair. It was observed that the cooperation (data access) among the control group declined over the period of the pilot. District health managers and pilot leaders noticed the problem:

"What I hear is that the people are not happy. [...] I remember, one manager complaining that even if they performed better, the bonus is only paid to the same YYY hospital every time. The control groups have performed even better than those getting the bonuses. YYY hospital got 10 million (shillings) and yet ZZZ hospital is doing far much better. It's frustrating and lost meaning – "done better but no bonus"" (District Health Official)

"I think the lack of cooperation (low response rate among control group) was because some units were not included in the bonus group. [...] They did not appreciate the issue of randomization. All they needed was to be in the bonus group. [...] It was an oversight. [...] It would have been a good idea that all participating (facilities) got something (benefit)" (Member, Pilot implementation team).

We conclude that the randomised design could have undermined health service provision in affected districts. The perceived unfairness was likely to lead to reduced cooperation by participants in the non-bonus group, potentially reducing their performance and preventing the pilot data from providing meaningful statistical comparisons.

Bonus payments and financing problems

Financial shortfalls were both a major constraint and a driver of adaptations with the aim of delivering planned outcomes with fewer resources than budgeted. As the main architect of the pilot, the DRG secured the funds for impact assessment and partial funding for pilot implementation activities. The financing of the pilot was itself subject to performance-based reimbursement of costs to the pilot implementation team. Two aspects illustrate the *de facto* implications of financial shortfalls.

a) Uncertainty of bonus funds:

Despite the centrality of the bonuses in the pilot as *the* intervention, the financing of bonus payments was not initially mobilised for the pilot. The Ugandan Ministry of Health (MOH) was expected to cover bonuses as counterpart funding. MOH funding did not materialize with the effect that bonus payments were uncertain and untimely.

> *"[…] This is the problem. It was assumed that MOH will take care of some of these costs but this is wrong.[…] In theory, they (MOH) accepted the study to be done, but due to resource constraints, their commitment to pay the bonuses was lost[…] It's possible that the Bank team did not do good thinking and ground work. Can you imagine? No budget for the critical parts of the study – the bonuses and facilitation of monitoring […]" (Member, Pilot Implementation Team)*

b) External influences on the pilot design

The United States Agency for International Development (USAID) country office agreed to finance the bonus payments and costs of the performance feedback meetings through one of their non-governmental (NGO) contractors in Uganda. For the first round of bonus payments in early 2005, few health centres achieved their targets, requiring only limited funding. Given this and the recognition of the need to rekindle interest of participants after pilot activities had stalled, the NGO contractor advised payment of higher bonuses than stipulated in the pilot contract. Bonuses paid in this first wave were doubled. By the time of the second bonus payment in late 2005, many pilot participants were eligible. The total bonus amount owed then increased 3-fold and this exhausted the NGO contractor's budget, necessitating financial restrictions. As a consequence of the restrictions, performance feedback meetings were altered from a participatory forum of multiple stakeholders to only include the bonus group facilities' personnel.

We conclude that the initial doubling of the bonuses had the potential to enhance the power of incentives to stimulate effort. However, the lack of funds for bonus payments introduced uncertainty and undermined the impact of the intervention.

Discussion and conclusions

There is a growing volume of literature presenting successful case studies of performance-based contracting (Basinga et al. 2011; Loevinsohn & Harding, 2005; Meessen et al. 2006, 2007; Soeters et al. 2006; Soeters et al. 2003; Yanagisawa et al. 2004). The orthodox evaluation of the Ugandan case study has not received significant attention, for example it has not been published in a peer-reviewed journal by December 2011. This is in marked contrast to the experiences of Rwanda and Cambodia that have been viewed as successful and have generated many peer-reviewed publications between them. The case exemplifies well-known publication bias in favour of success stories, and the consequent neglect of the lessons that can be learnt from more limited achievement of objectives, but also shows that a particular type of intervention can be promoted as overwhelmingly successful when its performance in other contexts has generated mixed effects (Bokhour et al., 2006; Rusa et al., 2009; Eldridge & Palmer, 2009).

The existing literature is also relatively silent on the mechanisms of effect underlying the successful case studies. It demonstrates that it is possible to achieve desired outcomes through performance-based contracting but does not help those aiming to emulate such success by establishing what features of design and context were supportive of those outcomes; or what mechanisms of effect were fired in the process. For example Soeters et al. (2006) and Basinga et al. (2011) report improved service provision without providing explanations of how those were achieved by the performance contracting interventions. A black-box approach to evaluation, together with "hands off" implementation – where the health providers identify their own innovations to achieve targets or performance auditors to verify results have been advanced by Soeters et al. (2006), as a means to optimise provider autonomy and entrepreneurial spirit. However the hands off approach limits the learning that would support replication of successful performance-based schemes because of its failure to clarify mechanisms and their contexts. This literature also illustrates a heavy reliance on external technical assistance to undertake major roles that local institutions would need to effectively shoulder for sustainable performance-based contracting initiatives.

What emerges from an open-box evaluation design is that the main reasons for the failure of the intervention were unrealistic design of the intervention, ill-considered adaptations made hastily as the inadequacies of the design revealed themselves and poor anticipation of the responses of institutions and individuals both inside and outside the change process. Key factors were the underfinancing of the initiative, the underestimation of the technical and institutional capacity requirements for successful implementation, the overloading of the implementation team with additional research activities, and the failure to consider important actors who influence outcomes but are not directly included in the change process.

Development interventions operate in contexts, which by definition are constrained in terms of institutional capacities and in practice, are constrained in terms of technical capacities. Ad hoc adaptations in these circumstances tend to have more than marginal implications for programme implementation and impact. They are not simply management failures. While management failures seem apparent in the case study reported, it is the lack of contextual understanding embedded in the initial design of the pilot that stands out. Unrealistic expectations of the capacities of Ministry of Health, pilot implementation teams and institutions targeted for performance improvement are reflected in all our key findings reported above. Selecting just a few of the mechanisms of effect identified by the study, it was unrealistic to expect the Ministry of Health's counterpart funding to materialise; to expect the pilot implementation team to cope with the data collection burden of either *de jure* or *de facto* interventions; to expect strategic identification of appropriate targets within the programming dynamics for curative and preventive services; or to expect health centre staff from the control group to be unaffected by their exclusion from the main benefit of the pilot.

Some of the mechanisms uncovered by this evaluation should have biased the study in favour of a conclusion of higher statistical effectiveness. The auditing team's increasing ability to read the information in primary registers will have inaccurately measured increases in performance; and the control group's disaffection should have resulted in misattribution of growing performance differences to the motivating effect of bonuses (rather than the demotivating effect of being excluded from them). Other mechanisms give clear explanations of the causes of failure, apparently sufficient to counterbalance these sources of superficially improved performance.

The case study demonstrates how the impacts of complex interventions are contingent on multiple micro level implementation details. We cannot conclude that performance-based contracting cannot deliver intended public health outcomes – only that it did not in this case, and that the reasons are embedded in the complexity of the intervention, the difficulties of making it work

and most importantly, the underestimation of these difficulties by the programme designers. A verdict that the intervention was effective (or not) hides the *de facto* from the *de jure* intervention. Only the *de facto* intervention can be deemed to have 'failed', but orthodox black-box evaluations of programme impacts are incapable of describing the *de facto* intervention. Adaptations to pilot implementation and response behaviours to the prevailing realities generated an evolving set of actions different from the assumptions of orthodox evaluation i.e. fidelity to established protocols (Gonzalez et al. 2004).

The lessons for those aiming to implement similar interventions are that PBC should not be attempted 'on the cheap', requires a plan to match institutional and technical capacities required of implementers to those that can be marshalled, and requires careful consideration of the likely responses of multiple actors – both insiders and outsiders to the intended change process. Given the complexity of PBC and scarce institutional and technical capacities for successful and sustainable implementation of performance-based contracting, traditional approaches to delivering public health goals (financing resource inputs and systems management) in environments such as Uganda should be strengthened. PBC case studies that have relied on imported technical and institutional capacities may mislead about the promise of this approach.

Acknowledgements

This work would not have been possible without the support and cooperation of the hospitals managers and the district officials that provided the bulk of the information and interviews. Additional gratitude go to the pilot implementation team and leadership at Ministry of Health-Uganda and World Bank Research Group that steering the pilot that provided the opportunity for this work. The Health Systems Development Programme (2002–2006) at London School of Hygiene and Tropical Medicine funded SF and BM, courtesy of a grant from the Department for International Development (DFID) UK. Many thanks go to Nicola Lord for all the administrative support to this study.

References

Annerstedt, M. (2010). Transdisciplinarity as an Inference technique to achieve a better understanding in the health and environmental sciences. *International Journal of Environmental Research and Public Health, 7*(6), 2692–2707.
Astbury, B., & Leeuw, F. (2010). Unpacking black boxes: mechanisms and theory building in evaluations. *American Journal of Evaluation, 31*, 363. doi:10.1177/1098214010371972.
Basinga, P., Gertler, P. J., et al. (2011). Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. *Lancet, 377*(9775), 1421–1428.
Bokhour, B. G., Burgess, J. F., Jr., Hook, J. M., White, B., Berlowitz, D., Guldin, M. R., et al. (2006). Incentive implementation in physician practices: a qualitative study of practice executive perspectives on pay for performance. *Medical Care Research Review, 63*(1 Suppl.), 73S–95S.
Conservatives.com. (2011). *One World Conservatism: A Conservative agenda for international development*. Policy Green Paper Number 11 London. Conservatives.com.
De Savigny, D., & Taghreed, A. (Eds.). (2009). *Systems thinking for health systems strengthening. Alliance for health policy and systems research*. Geneva: World Health Organization (WHO).
Eichler, R., Levine, R., & Performance-based Incentive Working Group. (Eds.). (2009). *Performance incentives for global Health: Potential and Pitfalls*. Washington DC: Center for Global Development.
Eldridge, C., & Palmer, N. (2009). Performance-based payment: some reflections on the discourse, evidence and unanswered questions. *Health Policy and Planning*, Accessed 07.02.09.
Galvin, R. S. (2006). Evaluating the performance of Pay-for-Performance. *Medical Care Research and Review, 63*(Suppl. 1 Feb), 126S–130S.
Gibson, J. L., Martin, D. K., & Singer, P. A. (2005). Priority setting in hospitals: fairness, inclusiveness, and the problem of institutional power differences. *Social Science & Medicine, 61*, 2355–2362.
Goddard, M., Mannion, R., & Smith, P. (2000). Enhancing performance in health care: the theoretical perspective on agency and the role of information. *Health Economics, 9*, 95–107.
Gonzalez, C. P., Barrera, M., et al. (2004). The cultural adaptation of prevention interventions: resolving tensions between fidelity and fit. *Prevention Science, 5*, 41–45.
Grol, R. P. T. M., Bosch, M. C., Hulscher, M. E. J. L., Eccles, M. P., & Wensing, M. (2007). Planning and studying improvement in patient care: the use of theoretical perspectives. *The Milbank Quarterly, 85*(1), 93–138.
Kalk, A. (2008). Rwanda's health system: some concerns. *The Lancet, 372*(9651), 1729.
Kelly, M., McDaid, D., Ludbrook, A., & Powell, J. (2005). *Economic appraisal of public health interventions*. Briefing Paper. London: Health Development Agency, National Health Service UK.
Lawler, E. E. (1971). *Pay and organizational effectiveness: A psychological view*. New York: McGraw-Hill.
Lawler, E. E. (1989). Pay for performance: making it work. *Compensation Benefits Review, 21*, 55–60.
Leykum, L. K., Pugh, J., Lawrence, V., Parchman, M., Noel, P. H., Cornell, J., et al. (2007). Organizational interventions employing principles of complexity science have improved outcomes for patients with Type II diabetes. *Implementation Science, 2*(28), 2–28.
Loevinsohn, B., & Harding, A. (2005). Buying results? Contracting for health service delivery in developing countries. *The Lancet, 366*(9486), 676–681.
Lundberg, M. (2010). Client satisfaction and the perceived quality of primary health care in Uganda. In S. Amin, J. Das, & G. Goldstein (Eds.), *Are you being served? New tools for measuring service delivery* (pp. 313–342). Washington DC: The World Bank.
Lundberg, M., Marek, T., & Pariyo, G. (2007). *Contracting for primary health care in Uganda*. Kampala: Ministry of Health.
McNamara, P. (2006). Provider-specific report cards: a tool for health sector accountability in developing countries. *Health Policy Planning, 21*(2), 101–109.
Meessen, B., Musango, L., et al. (2006). Reviewing institutions of rural health centres: the Performance Initiative in Butare, Rwanda. *Tropical Medicine & International Health, 11*(8), 1303–1317.
Meessen, B., Kashala, J.-P. I., et al. (2007). Output-based payment to boost staff productivity in public health center: contracting in Kabutare district, Rwanda. *Bulletin of the World Health Organization, 85*, 108–115.
Mooney, G., & Ryan, M. (1993). Agency in health care: getting beyond first principles. *Journal of Health Economics, 12*, 125–135.
Morgan, L. (2010). *Some days are better than others: Lessons learned from Uganda's first results-based financing pilot result-based financing*. World Bank. www.rbfhealth.org.
Pawson, R., & Tilley, N. (2004). How to construct realistic data. In R. Pawson, & N. Tilley (Eds.), *Realistic evaluation*. London: SAGE Publication Ltd.
Plsek, P. E., & Wilson, T. (2001). Complexity, leadership, and management in healthcare organisations. *British Medical Journal, 323*, 746–749.
Rosenfield, P. L. (1992). The potential of transdisciplinary research for sustaining and extending linkages between the health and social sciences. *Social Science & Medicine, 35*, 1343–1357.
Rusa, L., Schneidman, M., et al. (2009). Rwanda: performance-based financing in the public sector. Performance incentives for global health: potential and Pitfalls. In R. Eichler, R. Levine, & Performance-based Incentive Working Group. (Eds.). Washington DC: Center for Global Development.
Soeters, R., & Griffiths, F. (2003). Improving government health services through contract management: a case from Cambodia. *Health Policy Plan, 18*(1), 74–83.
Soeters, R., Habineza, C., et al. (2006). Performance-based financing and changing the district health system: experience from Rwanda. *Bulletin of the World Health Organization, 84*(11), 843–889.
Ssengooba, F. (2010). *Performance-based contracting: A case study of non-profit hospitals in Uganda Department of Global Health* (p. 253). London: University of London, PhD.
Teleki, S. S., Damberg, C. L., Pham, C., & Berry, S. H. (2006). Will financial incentives stimulate quality improvement? Reactions from frontline physicians. *American Journal of Medical Quality, 21*(6), 367–374.
Yanagisawa, S., Mey, V., et al. (2004). Comparison of health-seeking behaviour between poor and better-off people after health sector reform in Cambodia. *Public Health, 118*(1), 21–30.
World Bank. (2003). The frameworks for service provision. In World Bank. (Ed.), *World Development Report 2004: Making services work for poor people* (pp. 46–53), Washington DC.
World Bank, & Ministry of Health. (2004). *Contracting for health service Delivery: Health facility Questionnaire, Wave-1 2004*. Kampala: Ministry of Health.