Payment by results in international development: Evidence from the first decade

Paul Clist[1]

September 2018
School of International Development
University of East Anglia, UK
Email: paul.clist@uea.ac.uk
Website: https://paulclist.github.io/

This is a pre-print version of an article accepted in Development Policy Review. The published article is available at https://doi.org/10.1111/dpr.12405

[1] Acknowledgements: The article benefited greatly from comments by Robert Chambers, Brendan Whitty, Craig Valters, Arjan Verschoor, Maren Duvendack, Joseph Holden and DFID staff's willingness to share internal reports. This work builds upon work that was funded by DFID UK, but does not necessarily represent their views.

Motivation

Payment by results is a relatively new way of giving development aid, where a recipient's performance against pre-agreed measures determines the amount of aid they receive. It has proved popular, with most bilateral aid donors having at least experimented with the mechanism and the variety of measures stretching from individual health workers being paid for each procedure, to national governments being paid for students' test scores.

Purpose

Given a lack of robust and systematic investigations of the effects of PbR, the paper uses the leading theoretical framework to synthesise the available evidence, asking whether payment by results works as intended. This allows a test of the framework and the evidence against each other.

Approach

I synthesise the evidence from eight projects fully or partially funded by DFID, the recognised world leader on PbR. This represents the best evidence currently available, and is critically analysed using the leading theoretical framework that breaks each agreement into its constituent parts.

Findings

I find no evidence that PbR leads to fundamentally more innovation or autonomy, with the overall range of success and failure broadly similar to other aid projects. This may partly be due to the current use of Payment by Results, with no readily identifiable examples of projects that truly meet the idealised PbR designs. Advocates of PbR may conclude the idea is yet to be tested. I argue PbR does not deal with the fundamental constraints that donors face, and so it is unsurprising that PbR is subject to the normal pressures that affect all aid spending.

Policy Implications

Donors would be sensible to limit themselves to either 'small PbR' (where the costs of PbR are minimised) or genuine 'big PbR' (where projects seek to maximise PBR's benefits). The current evidence shows that projects outside these two categories are worse than traditional forms of aid. Evidence for 'small PbR' is mixed, while there is no evidence for 'big PbR' as it has yet to be tried.

Keywords: DFID, development, payment by results.

1. Introduction

"In essence, we are not arguing that COD [cash on delivery] Aid is worth trying because it creates a better incentive for recipient countries. We are arguing that it is worth trying because it creates a better relationship between funders and recipients"

(Birdsall & Savedoff, 2011, p. vii)

"We are pioneers of 'Results Based Aid' which incentivises partner governments to demonstrably transform peoples' lives."

Justine Greening, then Secretary of State for the UK's Department for International Development (DFID , 2014, p. 3)

Payment by Results (PbR), where aid is disbursed according to results achieved as measured against a pre-agreed tariff, is a relatively new idea that has seen remarkably fast adoption by donors around the world. Eyben (2015, p. 27) dates the first PbR programme to 2008: the World Bank's Health Results Innovation Trust Fund funded by the UK and Norway. Just eight years later, DFID was delivering 19 programmes with a PbR element, with a total multi-year value of £2.2 billion[2], of which approximately 30% was tied to delivery of pre-agreed outputs or outcomes (House of Commons International Development Comittee, 2016, p. 10). DFID, as self-professed pioneers of this approach, are not representative of other donors. However, many other donors have piloted PbR-style programmes, including The World Bank (Program For Results), the Asian Development Bank (Results-Based Lending), NORAD (REDD+, see Angelsen, 2017) and those donors (Agence Française de Développement, DFATD Canada, Denmark's DANIDA and Germany's KfW Development Bank) supporting the results based aid for fiscal decentralisation in Ghana and Tanzania (Janus, 2014).

Proponents of PbR highlight two apparent advantages. First, PbR is meant to solve a problem for donors, by demonstrating impact in a way that allows easy justification of spending on aid to sceptical taxpayers. The period of growth of PbR coincided with the global financial crisis and its aftermath, when aid spending remained buoyant despite high fiscal pressure. This combination is unlikely to be purely circumstantial: donors have experimented with PbR at a time when their budgets have come under pressure. Second, PbR is expected to be a more effective type of aid. Indeed, the hopes for PbR were revolutionary, being conceived by influential think tank the Centre for Global Development (Birdsall & Savedoff, 2011, p. 18) as "a substantial and fundamental change

[2] The number of projects is taken from DFID's own internal list, and was confirmed by the responsible DFID staff. For a comparison of the size, the total budget of DFID in 2016 was £9.9 billion.

in the way some foreign aid programs are conducted." In their discussion of aid more generally, Andrews, Pritchett and Woolcock (2013, p. 241) discuss that type of PbR (cash on delivery aid), with the expectation that greater autonomy would lead to better results as recipients are freed from following prescribed log frames and able to innovatively pursue a mutually agreed goal.

There are dissenting voices. From a theoretical perspective, Paul (2015, p. 313) argued that PbR "will probably not meet its promises" due to difficulties in incentivising all levels involved in influencing an outcome. Influenced by emerging evidence, Chambers (2017, pp. 74-75) has argued: "… forced to achieve targets for payment by results, organizations can face stark choices: abandon participation, go bankrupt, plead for clemency, relax standards of verification, gloss your reports, or lie, or some combination of these." Likewise, BOND (Longhurst & O'Donnell, 2014), an umbrella organisation of British Development NGO's, report instances where PbR has led to lower innovation, higher costs and a lower focus on marginalised groups.

While these sporadic insights are valuable, there now exists enough evidence to present a more systematic overview of how PbR contracts work in practice. Of the PbR projects that DFID has fully or partially funded over the last decade, evidence[3] currently exists for eight. These cover three sectors (education, health and training), a range of recipients (from national governments to small NGOs), a number of different implementing arrangements (from DFID themselves to fund-managed projects) and funders (multi-donor to single donor), and a wide range of budget size (from GAVI's £936m to the Nepali Employment Fund's £13.5m). An obvious, and fair, criticism of this approach is that it is akin to the drunk looking for their keys where the streetlight shines: the evidence that exists may not be representative of the range of experience in PbR projects. For example, no robust evidence exists for the Support for Malaria Control Programme in Tanzania because the programme was terminated "following suspicion of fraud." (DFID, 2015, p. 2). If evidence only exists for successful projects that will skew the conclusions drawn: this synthesis will not be able to detect the existence or direction of such a bias. However, systematic but flawed evidence is preferable to isolated impressions. This synthesis represents the best evidence available from the first decade of PbR's use from its most committed donor.

The evidence presented here is analysed with the leading theoretical contribution in mind: Clist's (2016) MAP framework, where each PbR contract can be analysed as comprising of the Measure (i.e. the tariff that determines the payment), the Agent (i.e. the recipient that will get paid) and the

[3] The inclusion criterion is deliberately low: projects need merely to have had some available evaluation documents. This precludes projects that were cancelled, cases where the only information was whether payment was made and a number of cases where evidence is forthcoming.

Principal (i.e. the donor). This enables the inclusion of established theoretical insights from contract theory, behavioural economics and development economics. It may be that ideas from very different settings are simply not relevant in typical PbR contracts, but the theoretically-informed framework should draw attention to likely issues.

The synthesis finds virtually no evidence that these PbR projects are *fundamentally* different from other aid projects. The effects on innovation and autonomy are neutral, with instances of innovative practice outnumbered by risk averse recipients taking a safety-first approach. Where PbR has succeeded, it is mainly in highlighting specific bottlenecks within a broader project. I call this use of PbR 'small', with no ready examples of the more transformative type of 'big' PbR that proponents described. I argue that blaming donors for not implementing the 'pure' idea is simply naïve: if PbR is to simultaneously solve a problem for donors and deliver more effective aid, it needs to recognise the practical and political constraints of donors that affect all types of aid. At present, PbR seems likely to settle into a practice of being an added-extra, used by donors seeking a way to exert more influence as part of a traditional aid relationship.

In section 2, I present an overview of the case studies included and the framework used to synthesise them. Sections 3-5 analyse the evidence according to the specific Measure, Agent and Principle respectively. The discussion and conclusion follow in sections 6 and 7.

2. Overview of the Eight Case Studies

As shown in

Table 1, robust evidence is available on eight projects which together have total program costs of around £1.7 billion. The majority of this is represented by the GAVI program, with a large spread in program costs. This variety is evident elsewhere, with a range of sectors (four health, three education and one employment), implementing partners (recipient governments, NGOs or private companies) and contractual arrangements. This enables some exploration of the likely conditions for PbR success or failure, and a judgement of whether the MAP framework captures the important determinants of these.

The evaluations are as varied as the projects. All meet basic standards by providing descriptions of outcomes under specific conditions, but several weaknesses are common. Often the evaluation confounds extra finances with the PbR mechanism, only captures performance by the incentivised measure or is unable to separate details of the design with the specific conditions of the project. These weaknesses are not universal. Where evaluations avoid these pitfalls, they shed light on the fundamental characteristics of PbR. Even when prone to these weaknesses, the evaluations are able

to provide specific insights on an aspect of the framework. By aggregating these disparate evaluations, we can achieve a more complete picture.

Table 1: Eight Case Studies

Short Project Name, with dates and *total* programme costs	Measure	Agent	Principal
GAVI (2010-16, £936m)	National vaccination rates	National Governments	Managed Fund
Health Results Innovation Trust Fund (2010-22, £114.25 from DFID)	A variety of indicators, focused on numbers of procedures administered	Health Facilities	Managed Fund
Reproductive Health in Pakistan (2012-17, £38.5m)	3 indicators of quantity of health services provided	NGOs	DFID
Results Based Financing in Health, Uganda (2009-16, £100.5m)	Several composite clinical performance indicators	Private not-for-profit providers	DFID
Girls Education Challenge Fund stage I (2011-17, £355m)	Learning (and attendance) outcomes	NGOs	Managed Fund
Results Based Aid in Ethiopia Education (2012-15, £27.4m)	Exam sitters and passers	National Government	DFID
Results Based Aid in Rwandan Education (2011-15, £97.5m)	Exam sitters and teachers proficient in English	National Government	DFID
The Employment Fund in Nepal (2010-15, £13.5m)	Employed Trainees, after 3 and 6 months	A range of providers (including public and private)	DFID

Of the health projects, the largest is the *Gavi Alliance* (formerly the Global Alliance for Vaccines and Immunisation): a multi-year attempt to increase vaccination rates in developing countries. The programme contains many elements, with the most relevant part the payments to national governments based on their measured vaccination rates: attempts to incentivise and reward high vaccination rates. While vaccination rates have gone up, Dykstra et al (2015) show that improvements weren't caused by Gavi incentives and Glassman and Sandefur (2015) show that much of the measured improvements were due to over reporting of vaccination rates.

The next largest health programme is the *Health Results Innovation Trust Fund* (HRITF), which has invested US $420 million since 2008 in PbR programs in the health sector across 32 countries, with total financing of $2.4 billion (mostly from the International Development Association). Of the 29

expected impact evaluations, Kandpal (2016) reviews the seven that are currently available. Whilst far from a complete picture of HRITF, the seven evaluations include many useful attempts to evaluate not just PbR projects but the PbR mechanism itself. HRITF projects have in common working in the health sector and a focus on paying individual health workers for improvements in the number or quality of measured procedures. They differ in the country they operate in, and the specific contractual details. An overview of the current evidence is that most outcome indicators have shown steady improvements, but the available impact evaluations have shown rather more mixed results (HDD, 2016).

Delivering Reproductive Health Results (DRHR) Through Non-State Providers In Pakistan was designed to strengthen the delivery and use of rural health services. The NGOs delivering the project used a social franchise model, and were paid according to performance against three key performance indicators. Witter et al (2016, p. 79) conducted an independent evaluation, finding that "on most indicators, performance (change since baseline) was either comparable to or worse than in the control areas." The management response to this was surprise (DFID, 2016), given previous positive signals from programme data, achievements against its logframe targets, and an independent mid-term review.

Results Based Financing Programme for Northern Uganda aimed to improve the health of post-conflict communities by paying private not-for-profit providers against a range of outcome measures. Valadez et al (2015) evaluated the project, using a nearby region as a control, which was funded by a more traditional input-based financing agreement. This means estimates of the effectiveness of the PbR mechanism are confounded by the region, furthermore PbR was not randomly assigned. However, typical evaluations of PbR confound the effects of additional funding and the contracting mechanism, and so this evaluation at least attempts to separate these by calculating a difference-in-difference estimate. While the quality of care is a concern across the board, the evaluation finds the PbR region achieved 50% of the available performance points, with the more traditionally financed control regions achieving only 20%: the PbR mechanism is associated with performance 2.5 times that of the more traditional mechanism.

The largest education project included is the *Girls' Education Challenge* (GEC): a flagship DFID programme with a budget of £355 million for the period 2013-2017, aiming to improve the education outcomes of girls in 18 countries. Of the 37 projects, 15 include an element of PbR where a fraction of a contract with the implementing NGOs was linked to learning outcomes (measured using literacy and numeracy tests) or attendance, often against a control group. While a synthesis of the GEC is not yet possible as the endline data is not yet available, emerging lessons from publically

available documents (e.g. Holden & Patch, 2017) and private impact evaluations allow an assessment of progress. It is worth noting that the final assessment of PbR in the GEC will be flawed as treatment and control groups were not randomly assigned, and the evaluations are of variable quality[4] as they were commissioned by the implementing NGO. A current overview of the evidence is that PbR has successfully increased a focus on learning outcomes where used, and that recipient NGOs have been highly motivated by the monetary incentive. However, this generally hasn't led to greater innovation, but instead a safety-first approach, as risk-averse NGOs seek to limit the chances of failure.

The two *Results Based Aid* (RBA) projects in education, in Ethiopia and Rwanda, are essentially twin programmes. They differ in several details, but were designed in tandem to pay governments for improvements in quantity and quality of education with both projects paying the recipient government on the basis of the number of students sitting key exams. In Ethiopia this was augmented by paying for the extra number of exam passers, while in Rwanda (on the initiative of the Rwandan government) it was augmented by paying for improvements in the English level of teachers. The two projects also diverged in that while the payments in both cases went to the central government, in the Ethiopian case the money was passed on to specific regions. Both projects were independently evaluated, and performance in Ethiopia was not "reasonably attributable to the RBA pilot" (Cambridge Education, 2015, p. iii), and in Rwanda "[t]he quantitative evidence is unanimous in finding that RBA had no consistent effect on completion results" (Upper Quartile, 2015, p. viii). The second of these evaluations is a particularly strong signal, as out-of-sample predictions from the underlying model are remarkably accurate.

The *Employment Fund* in Nepal is the smallest project included here, with one of the most positive evaluation findings. The project aimed to increase employment by providing skills training to young people, and part of the project was a PbR outcome-based payment to organisations who find their trainees work. Chakravarty, Lundberg, Nikolov and Zenker (2016) evaluated the entire project (not just the PbR element) against a control of no intervention, and found a 15-16 percentage point increase in non-farm employment. Average monthly income increased by 921 NRs (c.12 USD) against a baseline of 1272 NRs (c.17 USD): about a 72% increase for the combined 2010-2012 cohorts. The effects were larger for women than for men, with no obvious evidence of cherry picking.

[4] The evaluations are not publicly available, but this is my judgement having had access to all of the currently available evaluations.

3. Measure

The MAP framework (see Clist & Verschoor, 2014; Clist, 2016; Holden & Patch, 2017) is summarised in Table 2, with the three elements dealt with in these next three sections, starting with the most important: "[t]he principle factor that should determine whether PbR is used, and the strength of incentives, is the quality of the performance measure." (Clist & Dercon, 2014, p. 1). That quality is determined by the extent to which the measure captures something we really care about, even after it is incentivised (i.e. the signal/noise ratio). Other factors include the ease of gaming, the likelihood of asymmetric information and the ease of verification.

Table 2: MAP framework overview

MAP element:	PbR is more attractive if:
Measure	There is a low signal/noise ratio The measure is difficult to game There is little/no asymmetric information The amount at risk is enough to incentivise the agent Verification is relatively straight forward and cost effective
Agent (Recipient)	The agent has low risk and loss aversion The agent has smaller present bias The agent has higher control over the measure The agent and principal are not normally aligned
Principal (Donor)	The principal is able to contract over a long-enough period The principal is able to enforce the contract

Looking at the eight case studies, we can identify cases where well-chosen measures appear to have focused minds on achieving a sensible goal. With the Girls Education Challenge Fund, Holden and Patch (2017, p. 6) argue "[t]he overall focus on learning outcomes and their rigorous measurement was broadly seen as very positive, a 'step change' for some organisations." In the Zambian Health Results Innovation Trust Fund project one health worker explained: their "attitude has really changed, people used to come late for work, now everyone is on time. We were doing shortcuts, but now we are doing full procedures." (Evans, 2016) Of course a greater focus on achieving the measure may not be sufficient, but it is normally necessary.

Most of the eight cases studies include an attempt to assess any cherry picking or gaming, where the measure is artificially inflated either by focusing on the easiest people to help, or by more brazen means. Generally, there is no evidence of any problems. Regarding Health Results Innovation Trust Fund's Zimbabwean project, Kandpal (2016, p. 12) states "none of the non-incentivized services investigated showed a decline in the number of cases treated, as would be expected if task shifting

affected these services." This is a particularly strong piece of evidence, given good measurement of other tasks that could have been neglected. (However, the results on the actual indicators aren't particularly positively affected, and so it remains unclear how changing the size of the incentives would have affected both the PbR measures and other activities.) Evidence from the Employment Fund in Nepal (Chakravarty et al, 2016, p. 6) suggested the measure itself discouraged cherry picking as it included greater payments for disadvantaged groups. So far, there has been little ability to accurately document cases of perverse incentives, despite common fears. For those on the recipient side of the contract in the Girls Education Challenge Fund, Holden and Patch (2017, p. 36) provide a useful insight: "Respondents perceived there were perverse incentives from PbR, particularly to prioritise the short-term over the long-term. They claimed their projects did not respond to these incentives, although sometimes felt headquarters pressure to do so." Few recipients will claim to have been affected by such incentives, but the current evidence does not provide any clear examples.

However, the more prevalent risk with PbR measures is much more subtle and difficult to measure than fraud, gaming or perverse incentives. The widespread risk is of poor quality of measure (a low signal/noise ratio), i.e. a low alignment between the measure and what we really care about, especially after the measure has been incentivised. While the majority of cases don't allow for an assessment of the quality of the measure (as the only record of success is the incentivised measure), where such a judgement is possible we find several measures that do not robustly capture success. The problems with the initial GAVI measures have been robustly demonstrated (see Glassman & Sandefur, 2015 and the references therein): GAVI initially used reliable self-reported administrative data for vaccination rates from each country, but once that data was incentivised recipient countries tended to over-report. A more accurate picture of vaccination rates can be recovered by triangulating vaccination rates from a non-incentivised source (the Demographic and Health Survey), which shows the PbR measure was simply disbursing too much money for the progress achieved.

In other cases, the chosen measures simply don't capture the underlying goal very well. This has been a problem in the implementation of several PbR projects, occasionally meaning the original design is altered. For example, Holden and Patch (2016, p. 6) state that "… learning and attendance had an equal focus on the programme, but due to measurement issues, projects in 2014 were given the choice to remove attendance as a PbR outcome." In most cases however, poor quality measures are retained. Both of the education RBA projects included here (Rwanda and Ethiopia) use exam sitters to capture educational quality, which has since been criticised (e.g. Upper Quartile, 2015, p. 47), as students can *sit* exams without learning anything. In both cases, the problems with the

measure run deeper. In Rwanda the test of teacher's English was used for part of the payment. However, it was not comparable across years (a different test was taken in different conditions), and so while it was paid out upon it had no robust ability to measure improvements in the level of English (Upper Quartile, 2015, p. 25).

In Ethiopia, another set of problems is apparent, this time due to a measure of exam passers, in an attempt to measure changes in quality. DFID (2016) states that "[i]nvestigation by the Project Completion Review (PCR) team and DFID Ethiopia advisers suggests that the pass rate for 2014/15 was due to a change by the National Exam Agency (NEA) in the statistical process for calculating pass rates." In other words, a change by the exam agency led to differences in pass rates that was unrelated to educational quality: this led to greater payments for improvements in educational quality. There is a larger issue. Exams in Ethiopia are effectively 'graded on a curve', and so pass rates should theoretically be consistent regardless of the quality of each cohort. In essence a measure that should not be able to change was selected in order to incentivise improvements. The measure then *did* improve markedly, but this cannot be related to a large improvement in actual quality.

A less-widely-discussed problem with the GAVI measures is that they seem to have little effect on non-performing countries[5]. This is apparent from Khatib-Othman (2016, p. 3):

"This brings into question the benefit of Gavi's current PBF [Performance Based Finance: a form of PbR] approach. Trends in performance payment eligibility for 2014 – 2016 show that the intention of incentivising improved coverage and data quality has not been realised. Instead, the PBF approach has largely served as a reward to countries with over 90% DTP3 coverage for maintaining high coverage."

This argument receives support from Dykstra et al (2015), who find that GAVI had no robust positive effect on immunisation rates (their regression discontinuity design focuses on countries near the GNI threshold of $1,000, and so can't provide insights on countries far from that threshold). This is a perennial problem: appropriately matching the incentive to different levels of cost and interest for different parties so that a broad spectrum of agents is incentivised.

An issue that was not incorporated into the MAP framework, but is now clear, is the extent to which measures will fail to incentivise recipients simply because they are too complex relative to the

[5] At this stage any attempt to understand the reason for this is speculative: more data is expected soon. However, it is consistent with 'cherry picking' i.e. where the easiest to change programs do respond to incentives (by continuing with improvements) but the hardest-to-change cases are left unaffected.

incentive size. This is clear from agreements with individuals (see the HRITF agreements in Afghanistan and Cameroon discussed in Kandpal, 2016, p. 7), NGOs (Holden & Patch, 2017, p. 6) and Governments (see the education RBA agreements with Ethiopia and Rwanda discussed respectively in Cambridge Education, 2015 and Upper Quartile, 2015). Even apparently simple measures are not considered worth the investment to really understand if the attached payment is sufficiently small.

The question of the size of the incentives on offer is a delicate balance. It is clear from the theory that only a good measure can bear large incentives without problems (Clist & Dercon, 2014, p. 1). Beyond that there is a balance between having incentives that are large enough to incentivise a recipient and small enough to be both value for money and not encourage gaming, fraud or ultimately unproductive activities (Clist & Verschoor, 2014, pp. 22-27). A common theme for projects with poor performance is low-powered incentives (e.g. both RBA education projects). In line with theoretical predictions, it appears that NGOs perceive incentives to be higher-powered than governments, as they themselves are smaller and more risk averse. As such, the 10% PbR element was felt to be sufficiently large for NGOs in the GEC (see Holden & Patch, 2017), but the 100% PbR element of multimillion pound agreements were too small for some recipient governments (e.g. on GAVI, see Khatib-Othman, 2016, p. 3). An interesting case comes from the HRITF:

"Indeed, the strongest evidence for sustained impacts from RBF [results based financing, a form of payment by results] comes from the Misiones province, where the increase in incentives was substantial—threefold. It may also be the case that the signaling effect of an incentive introduced by a health system in an environment that previously did not incentivize individual services may be somewhat more effective at changing behavior than the income effect of the relatively small incentive amount offered." (Khandpal, 2016, pp. 14-15)

The last aspect of the measure is that the verification process needs to be reasonably straightforward and cost effective. The practicality of verification in typical contexts are often questioned; Holden and Patch (2017, p. 6) concur "[t]here were also concerns around the complexity of assessment and the capacity of evaluators to enumerate them properly." The experience is in contrast to how verification was envisaged. Birdsall and Savedoff (2011, p. 59) argued that "reporting and verification also provide incentives to improve *education* data" (emphasis added). More recently, Barder et al. (2014) argued that verification should be cheaper than alternative systems (input tracking) and will lead to benefits of better information: "The focus on results need not be more expensive in terms of staff or money than the detailed tracking of inputs which it replaces, and because it focuses on outcomes it may provide much more useful information." The optimism that verification costs are lower and offset by benefits in information appears naïve in the

face of current reality. There is no evidence so far that verification strengthens *standard* data gathering procedures: they are typically standalone efforts in order to have the necessary confidence to pay out upon a contract. Where standard data has been used (e.g. GAVI, Ethiopia RBA, Rwanda RBA) there is evidence that the quality of the data declined (respectively Sandefur & Glassman, 2015; Cambridge Education, 2015; Upper Quartile, 2015), as would be expected (Clist, 2016 and references therein). It is of course possible that in time some of the verification costs can be reduced (as better measures are identified) or that those costs will be offset by standard data sources improving in quality, but at the moment verification is often *felt* to be a substantial cost with few redeeming benefits.

To summarise briefly, the evidence shows some cases where the incentivised measure increased attention and performance. In several cases measures were of poor quality: predictably there is little evidence of success in such cases. An obvious policy implication is merely to choose good measures, but there should be a pragmatic understanding that many of the measures were judged as good at the time, often by PbR's leading proponents. A good measure is hard to find, but the evidence concurs with the theoretical prediction that it is needed in order for PbR to be better than alternatives.

4. Agent

The agent (recipient government or implementing organisation) is the area with the least amount of evidence from current PbR projects, mainly as it has the smallest amount of variation. On risk aversion, there is some evidence that the NGO's dealing with the Girls Education Challenge fund took fewer risks because of the PbR contract (Holden & Patch, 2017, p. 7). On recipient control, there is some evidence of cases where PbR failed because the recipient had limited ability to affect the outcome (e.g. the Afghanistan project in the HRITF discussed in Kandpal, 2016). On alignment, this 'first generation' of PbR contracts has tended to select recipients which were felt to be more aligned, and so there is little variation that could generate evidence of the effect of different levels of alignment. Holden and Patch (2017, p. 36) state that in GEC, "[p]roject staff are generally very motivated to achieve outcomes, and this is not linked to the payment incentive for those on PbR projects." Likewise in Rwanda, the government was felt to already be focused on increasing enrolment (Upper Quartile, 2015). By contrast, the evidence base provides few ready examples of cases where PbR aimed to incentivise recipients that had fundamentally different objectives to the donor. On the time horizon of recipients, there is an indication that some recipients were overly focused on the short term. On each of these aspects, theory predicts that these will influence the

effectiveness of PbR projects, but there simply isn't sufficient variation to examine their effect in practice.

One characteristic of an agent that *has* generated evidence is around the effect of PbR on motivation (a point not really foreseen in the MAP framework). This is more of an issue where those incentivised are individual staff, and so most of the evidence comes from the HRITF which mainly used supply side incentives in the health sector. The most negative effect was found in the DRC, as design problems and poor decision making caused an average 34% drop in take home pay for health workers, and 42% lower earnings for facilities (Kandpal, 2016, p. 9). This is representative of general difficulties of implementation, where agents do not always respond in the envisaged fashion. The early evidence from the HRITF shows that "RBF mechanisms are not always easy to implement and have been associated with implementation failures that result in less effective programs." (Kandpal, 2016, p. 15) The logic of PbR is of course that these unexpected responses could be positive or negative, and in negative cases the donor disburses lower funds. However, it appears that PbR has, so far, been associated with consistent implementation difficulties for more complicated projects.

Returning to motivation, another negative effect was found in the HRITF's Zimbabwean project with staff reporting a greater likelihood of burnout in PbR areas. A more positive effect was found in Afghanistan, where there was a perceived boost in motivation. An evaluation in Zambia "found large gains in health worker satisfaction and staff motivation" (Kandpal, 2016, p. 12). Evans (2016) argues (with a specific focus on Zambia) that this works not through pecuniary interest but rather in simply being recognised in a context where workers feel undervalued. A last example of positive motivation comes from the final year project completion report of the Ethiopia RBA, where incentives were passed on to the school level in some regions: "there appeared to be broad support for this 'reward for performance'. It was cited as being a positive motivating factor for teachers and school administrations, and in some cases regional bureaus" (DFID, 2016, p. 13) The current evidence on motivation is thus *suggestive*: there are some instances where PbR schemes have seen increases in motivation, and others where problems in the measure undermined motivation.

5. Principal

The principal (i.e. the donor) will clearly affect the effectiveness of any aid program, but there are a number of specific ways that they may do so in a PbR project. One important aspect will be whether they are able to withhold aid from non-performing recipients, as if they are not then there is little incentive for recipients to expend extra resources in order to meet these targets. Amongst others, Svensson (2003) showed that aid donors typically found it difficult to withhold aid in ex ante

conditional aid agreements, and so it is possible that assumptions they will be able to do so in PbR contracts are misplaced. The current evidence is mixed. From the Girls Education Challenge Fund (Holden & Patch, 2016, p.8) we see that NGOs didn't doubt the ability to withhold aid:

"Project staff generally understood the PbR risk, and saw the threat as credible that DFID would be willing to hold back PbR lost on the downside. Head offices were more concerned about the PbR downside risk than local offices, and in some cases this put significant pressure on organisations, as one stated it was a 'sword hanging over our heads'."

This experience seems more typical with Fund-managed programmes, with the final destination of non-disbursed funds often somewhat unclear when the agreement involves DFID field offices, who naturally don't wish to lose control of unspent funds. For example, with RBA in Rwanda "it is not clear how the unspent funds are used." (Upper Quartile, 2015, p. 44).

One of the claimed benefits of PbR (Birdsall & Savedoff, 2011, pp. 21-22) is that it enables recipients greater flexibility and autonomy to achieve the targets in different ways. Previous evidence has questioned whether this link genuinely exists, e.g. Honig (2014) found autonomy was not higher in World Bank's projects when PbR contracts are used. The emerging evidence from the HRITF is useful here: "[a] common theme in the results from Argentina, Afghanistan, Cameroon, Zambia and Zimbabwe is that RBF schemes effectively improve autonomy at the facility level" (Kandpal, 2016, p. 13) and "[e]xamples of institutionalizing RBF in the context of Burundi or Rwanda provide strong evidence on the need for facility autonomy. They also show the importance of PFM reforms in ensuring that RBF moves beyond the program stage." (WHO, 2016, p. 4)

Outside of the experience with the HRITF, there is scant evidence that PbR has allowed for greater autonomy. The main limiting factor here appears to be DFID's own systems. Holden and Patch (2016, p. 7) report "…the process for making changes on the GEC [Girls Education Challenge Fund] in terms of milestones, outputs and budget amendments, was felt by some to be too time-consuming and cumbersome and a barrier to adaptation." This appears to be a major factor in why there was not a higher level of changes and adaption amongst GEC projects that had PbR compared to those that didn't.

A common theme across the projects is that current PbR projects have been subject to both the expectations of PbR projects to be innovative and the standard procedures of more traditional aid modalities. Occasionally, these are augmented by new financial procedures, due to the contractual nature of PbR agreements. These dual requirements have tended to undermine possible gains in autonomy and innovation. To rehearse the arguments, one hope for PbR is that it enables recipients

of aid monies to innovate and discover through trial and error the most successful way of delivering the contracted results (Birdsall & Savedoff, 2011). Current evidence shows that autonomy (e.g. in the HRITF) is the exception rather than the rule. Holden and Patch (2017, p. 7) are somewhat typical in the discussion of the Girls Education Challenge Fund: "a consistent view emerging from the study is that PbR did not incentivise innovation, and more likely had the opposite effect, leading organisations to be more risk-averse"

Closely related to incentivising innovation is the length of the contract – a donor that is able to contract over a longer time horizon is predicted by the theory to see greater innovation (as the rewards for successful innovation are captured for longer, and the feedback loop works a greater number of times allowing successful adaption). Here too, current evidence is not positive in terms of the design of projects. The main reasons given for non-impact in Ethiopia's education RBA were the relatively small incentive of the project, especially in comparison to its complexity and *duration* (Cambridge Education, 2015, p. v). A similar story is found in the Rwandan RBA: "While RBA was perceived to be a small amount of money, it is possible that the reason it did not receive a greater response was more due to the short length of the agreement" (Upper Quartile, 2015, p. 46) Also in the education sector, Holden and Patch (2017, p. 7) found in the Girls Education Challenge Fund that "[p]roject staff perceived potential perverse incentives from PbR, particularly to prioritise the short-term over the long-term." Presumably, a longer agreement would ease this pressure. It is worth noting findings in the broader literature where incentives worked in the short run but not the long run (Muralidharan & Sundararaman, 2011; Olken, Wong, & Onishi, 2014). Unfortunately, the case studies identified are only able to offer examples where non-impact has been related to short time horizons, but there simply aren't robust examples of longer term agreements to see whether this is related to more successful PbR outcomes.

One consequence of the difficulty of enabling adaption is that the quality of the original plans have a greater weight in determining the effectiveness of a PbR project. This relates to a discussion by Clist (2016, p. 309), who argues donors may need *more* information when designing a successful PbR project than for a more traditional project. While in theory PbR means a greater ability to innovate, in practice current PbR often has time horizons that are too short and incentives of the wrong level.

The evidence emerging from HRITF is also worth examining a little further here, as it emphasises the need of PbR to successfully identify (and incentivise) bottle necks in order to really achieve underlying goals. Kandpal (2016, p. 13) discusses the case of Afghanistan, where insufficient attention to demand side factors explain the failure of supply side incentives to work as planned. Here, it appears PbR successfully incentivised the recipient, but the wrong constraint was targeted,

and so the recipient was unable to achieve the desired goals. Even with a PbR contract that was able to provide a degree of autonomy, the recipient was hamstrung by a project design that didn't target the binding constraint. This is not a criticism of the original design work – it is often not obvious where such constraints are. Furthermore, the difficulty of PbR contracts here is that PbR tariffs must be predictable: if the initial design misses the binding constraint then it is difficult to adapt the contract to target it at a later date.

6. Discussion: Big and Small PbR

Two different approaches to PbR are beginning to emerge from the case studies analysed. These 'sweet spots' can be thought of as big and small versions of PbR, and are described below.

Table 3: Two Types of PbR

	'Big' PbR	'Small' PbR
Measure	High powered incentives Excellent quality measure Should be Outcome focused Complexity allowed Costly verification allowed	Low powered incentives Fair quality measure Could be output focused Very simple measure Requires cheap verification
Agent	Mostly Governments or Private Sector Needs reduced input tracking Needs low cost to change to plans	Mostly NGOs or Private Sector Standard procedures less harmful
Principal	Longer term agreement required, with multiple pay-outs on one measure Requires good design of measure Requires strong ability to withhold	Shorter time horizon allowed, changes less damaging Requires good design of intervention Ability to withhold less crucial

Table 3 provides a simplification of two sensible combinations of different factors that are both logically coherent and have emerging empirical reasons for considering. 'Big' PbR most closely resembles what has been discussed as 'Cash on Delivery Aid' (Birdsall & Savedof, 2011). Some of the elements are necessary to ensure there is a possibility for success of this kind of PbR, whereas others are secondary issues that may affect the level of success but do not preclude it. Required elements include an excellent measure of something we really care about, which is incentivised for a long enough period of time[6] and at a sufficient level, with enough administrative space to search for the

[6] A longer time horizon means a greater degree of reward once an agent (aid recipient/implementer) discovers a successful approach, and so incentivises investment in discovering that approach (Clist & Verschoor, 2014, p. 20). Furthermore, multiple payments are needed in a 'big PbR' project to act as feedback to the agent so they can refine their approach (see Birdsall & Savedoff, 2011).

right approach. Secondary issues here include a measure that is difficult to explain or a costly verification process. A measure that is very complex does not undermine PbR here as the incentive to truly understand the process is there, and enough time is available to respond to feedback. Likewise, a costly or difficult verification process doesn't necessarily undermine this type of PbR as the cost is offset by less tracking of inputs and/or the lower cost of changing approaches.

The evidence base so far does not provide ready examples of big PbR projects, with a question hanging over whether donors are able to design and implement this kind of PbR given their constraints. The constraints appear to be inherent to donors and affect other types of aid too, be it in the difficulty of withholding aid (Svensson, 2003), the institutional incentives to disburse and the difficulty of delegation (Easterly, 2003), and the volitility of aid which reveals a difficulty in long term consistency (Hudson & Mosley, 2008). Most projects that are close to 'big PbR' fall down on these familiar elements: the time horizon or the permitted autonomy for the recipient, with evidence often pointing to the ways in which these projects don't fit the model as a reason for a lack of success. In this way, much of the evidence for this kind of PbR remains negative, meaning projects that lack these characteristics are found not to work, and the *absence* of these characteristics are often pointed to as crucial.

By contrast **'**Small PbR' has several examples, including several projects from the Results Based Financing Programme for Northern Uganda, the Health Results Innovation Trust Fund and Nepal's Employment Fund. These kinds of projects essentially see PbR as a small element of their overall project, and mostly use measures that capture something indicative of success rather than measuring success itself. The evidence for these projects is often difficult to assess as they are bound up in the programme as a whole. Where PbR is well designed and a genuine bottle neck has been identified, PbR appears to bring greater attention and focus on the results and adds value to the project. There is also some, limited, evidence of an increase in motivation, but in general the mechanism appears to be one of higher attention. Where the programme is poorly designed, PbR is unable to overcome that difficulty and only adds costs to the programme. In essence, with 'small PbR' its success or failure is bound up in the overall project, and relatively poor measures are less problematic than they would be in 'big PbR' as they use low-powered incentives. Therefore, requirements for 'small PbR' are that the *project* is well designed, the PbR element is well targeted and that the additional costs of PbR (including the verification, difficulty of explaining to the recipient, risks of gaming, and management) are small.

7. Conclusion

For its proponents, Payment by Results has revolutionary potential. They envisage a situation where recipients innovate more due to greater autonomy, donors see their aid budgets go further through more effective aid and taxpayers receive reassurance that their money is well spent. After the first decade of PbR's use by DFID, its most committed supporter, this article asks whether such hopes have been achieved.

Using the MAP framework, the clearest warning from the theory is of the difficulty of finding a good measure (Clist, 2016). This chimes with the available evidence: measured results were often achieved and paid out upon, while the underlying goal of the donor was unmet. For Gavi, this was a subtle process of the incentivised data source starting to over report progress. There are starker examples. In Rwanda, payment was made on the basis of measured improvements by non-comparable teachers tacking non-comparable tests under non-comparable conditions. In Pakistan, an independent evaluation found a project that appeared to be performing well had worse scores than control groups. In Ethiopia, the goal of incentivising improvements in the quality of education was undermined by the use of norm-referenced exams, meaning the results bore no relation to actual quality improvements and could not increase by definition. Strangely, large improvements in quality were then reported, and paid out against. These may be mere teething problems of a new modality, but it is clear that good measures are hard to find.

Where PbR has been effective, it has been in 'small PbR' cases, with individual employees or NGOs being successfully incentivised and rewarded for high performance. An obvious corollary of this may be that proponents point towards the discrepancy between the ideal PbR conditions (see table 2) and current practice. The evidence examined here does not provide any genuine examples of 'big PbR', the kind of revolutionary change that excites proponents. One obvious failing is that DFID do not contract over a single measure for a long enough period, with multiple feedback loops, and so the ability to reap the reward for high performance is too limited to justify heavy investment in innovation and adaption. In other words, PbR typically does not resemble Cash on Delivery Aid (Birdsall & Savedoff, 2011), and so it may be unsurprising that results are not more positive.

The criticism is accurate: in the first decade of use, the leading proponent of PbR has mainly implemented 'small PbR', with the most positive PbR results found when its use has been incidental to the programme design. It has helped to deliver change where the PbR measure has targeted a binding constraint, but misfired when the incentives have been too low, complicated, or misguided. However, such criticism is also naïve: PbR is bound by the same political constraints as other forms

of aid. Current PbR practice differs from the idealised PbR contracts in many ways, including difficulties in withholding aid, being inconsistent over longer time horizons and flawed yet inflexible project designs. However, these realities are far from unique to PbR, and are well known weaknesses of aid more generally. If donors could circumvent these constraints, then aid could be more effective across the board. Instead, PbR currently seems destined for a familiar mix of success and failure, its main addition being occasional illusory success.

First submitted May 2018

Final draft accepted September 2018

Bibliography

Andrews, M., Pritchett, L., & Woolcock, M. (2013). Escaping capability traps through problem driven iterative adaptation (PDIA). *World Development, 51*, 234–244.

Angelsen, A. (2017). REDD+ as result-based aid: General lessons and bilateral agreements of Norway. *Review of Development Economics, 21*(2), 237–264.

Barder, O. P., & Talbot, T. (2014). 12 Principles for Payment by Results (PbR) in the Real World. *Mimeo, available at https://www.cgdev.org/blog/12-principles-payment-results-pbr-real-world-0*.

Birdsall, N., & Savedoff, W. (2011). *Cash on Delivery: A New Approach to Foreign Aid.* available at https://www.cgdev.org/publication/9781933286600-cash-delivery-new-approach-foreign-aid: Center for Global Development: Washington, D.C.,.

Cambridge Education. (2015). Evaluation of the Pilot Project of Results-Based Aid in the Education Sector in Ethiopia. *available at: http://iati.dfid.gov.uk/iati_documents/5608531.pdf*.

Cashin, C. F., & Hashemi, T. (2015). Verification Of Performance In Results-Based Financing (Rbf): The Case Of Afghanistan, Health, Nutrition and Population. *HNP discussion paper, available at: http://documents.worldbank.org/curated/en/561511468187139014/Verification-of-performance-in-Results-Based-Financing-RBF-the-case-of-Afghanistan* .

Chakravarty, S. L., & Zenker, J. (2016). The Role of Training Programs for Youth Employment in Nepal: Impact Evaluation Report on the Employment Fund. *World Bank Policy Research Working Paper , 7656*.

Chambers, R. (2017). *Can We Know Better? Reflections for Development.* Rugby, UK: Practical Action Publishing.

Clist, P. (2016). Payment by Results in Development Aid: All That Glitters Is Not Gold. *The World Bank Research Observer, 31*(2), 290–313.

Clist, P., & Dercon, S. (2014). 12 Principles for Payment By Results (PbR) In International Development. *Mimeo, available at https://assets.publishing.service.gov.uk/media/57a089d2e5274a27b20002a5/clist-dercon-PbR.pdf* .

Clist, P., & Verschoor, A. (2014). The Conceptual Basis of Payment by Results. *Mimeo, available at https://assets.publishing.service.gov.uk/media/57a089bb40f0b64974000230/61214-The_Conceptual_Basis_of_Payment_by_Results_FinalReport_P1.pdf* .

DFID . (2014). *Designing and Delivering Payment by Results Programmes: A DFID Smart Guide.* available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/352519/Designing-Delivering-PbR-Programmes.pdf .

DFID. (2015). *Annual Review of Support for Malaria Control Programmes.* Retrieved from http://iati.dfid.gov.uk/iati_documents/5167673.odt

DFID. (2016). *DFID Management Response: Delivering Reproductive Health Results (DRHR) through non-state providers in Pakistan.* Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/563590/Man-resp-delivering-reproductive-health-results-pakistan.pdf

DFID. (2016). *Project Completion Report of Pilot Project of Results Based Aid in the Education Sector in Ethiopia.* Retrieved from http://iati.dfid.gov.uk/iati_documents/5419380.odt

Dykstra, S., Kenny, C., Sandefur, J., & Glassman, A. (2015). The Impact of GAVI on Vaccination Rates: Regression Discontinuity Evidence. *Center for Global Development Working Paper, 394*.

Engineer, C. D., & Peters, D. (2016). Effectiveness of a pay-for-performance intervention to improve maternal and child health services in Afghanistan: a cluster-randomized trial. *International Journal of Epidemiology, 45*(2), 451–459.

Evans, A. (2016). Results based financing in Zambia“ an informal, unpublished annex. *mimeo, available at: https://www.researchgate.net/publication/308985858*.

Eyben, R. (2015). Uncovering the politics of evidence and results. In I. G. Rosalind Eyben, *The Politics of Evidence and Results: Playing the game* (pp. 19–38). Rugby, UK: Practical Action Publishing.

Gertler, P. G., & Martinez, S. (2014). Rewarding provider performance to enable a healthy start to life: evidence from Argentina's Plan Nacer. *World Bank Policy Research Paper, 6884*.

HDD. (2016). Update on the latest evidence emerging from Results based financing in health. HDD team meeting presentation, internal.

Holden, J., & Patch, J. (2017). The experience of Payment by Results (PbR) on the Girlsâ€™ Education Challenge (GEC) programme Does skin in the game improve the level of play? *Mimeo, available at: http://foresight.associates/wp-content/uploads/2017/01/2017.01.19-Skin-in-the-game-PbR-on-the-GEC.-Final.pdf* .

Honig, D. (2014). Navigation by Judgment: Organizational Autonomy and Country Context in the Delivery of Foreign Aid. *Mimeo*.

House of Commons International Development Comittee. (2016). *UK aid: allocation of resources: interim report: Government Response to the Committee's Third Report Session of 2015-16.* House of Commons. Retrieved from https://publications.parliament.uk/pa/cm201617/cmselect/cmintdev/256/256.pdf

Hudson, J., & Mosley, P. (2008). Aid Volatility, Policy and Development. *World Development, 36*(10), 2082–2102.

Janus, H. (2014). Real Innovation or Second-Bset Solution? First experiences from results-based aid for fiscal decentralisation in Ghana and Tanzania. *DIE discussion paper, 3*.

Kandpal, E. (2016). Completed Impact Evaluations and Emerging Lessons from the Health Results Innovation Trust Fund Learning Portfolio. *available at: https://www.rbfhealth.org/sites/rbf/files/IE%20and%20emerging%20lessons_Eeshani%20Kandpal.pdf*.

Khatib-Othman, H. (2016). Country Programmes: Strategic Issues, Report to the [GAVI] Board 7-8 December 2016, Appendix B. *available at http://www.gavi.org/about/governance/gavi-board/minutes/2016/7-dec/minutes/07a---country-programmes---strategic-issues/*.

Kutzin, J. (2016). RBF: from program to entry point for strategic purchasing. *Presentation at the Health Financing Technical Network Meeting, slides available at http://www.who.int/health_financing/events/session3-results-based-financing-and-strategic-purchasing.pdf* .

Longhurst, R., & O'Donnell, M. (2014). *Payment by Results: what it means for UK NGOs.* BOND.

Mumssen, Y. J., & Kumar, G. (2010). *Output-based aid: lessons learned and best practices. Directions in development; finance.* Washington, DC: World Bank.

Muralidharan, K., & Sundararaman, V. (2011). Performance Pay: Experimental Evidence from India. *Journal of Political Economy, 119*(1), 39–77.

Olken, B., Wong, S., & Onishi, J. (2014). Should Aid Reward Performance? Evidence from a field experiment on health and education in Indonesia. *American Economic Journal: Applied Economics, 6*(4), 1–34.

on Output Based Aid, G. P. (2016). Annual Report. *available at: https://www.gpoba.org/sites/gpoba/files/GPOBA_AnnualReportFY2016_0.pdf*.

Paul, E. (2015). Performance-Based Aid: Why It Will Probably Not Meet Its Promises. *Development Policy Review, 33*(3), 313–232.

Perakis, R., & Savedoff, W. (2015). Does Results-Based Aid Change Anything? Pecuniary Interests, Attention, Accountability and Discretion in Four Case Studies. *CGD Policy Paper, 52*.

Sandefur, J., & Glassman, A. (2015). The political economy of bad data: evidence from African survey and administrative statistics. *The Journal of Development Studies, 51*(2), 116–132.

Savedoff, W. (2010). Basic Economics of Results-Based Financing in Health. *Mimeo, available at http://www.focusintl.com/RBM082-RBF%20Economics.pdf*.

Svensson, J. (2003). Why conditional aid does not work and what can be done about it? *Journal of Development Economics, 70*(2), 381–402.

Upper Quartile. (2015). Evaluation of Results Based Aid in Rwandan Education. *available at: http://iati.dfid.gov.uk/iati_documents/5549076.pdf*.

Valadez, J., Jeffery, C., T, B., W, V., & Pagano, M. (2015). Final Impact Assessment of the Results-Based Financing Programme for Northern Uganda. *available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/607579/Evaluation-of-Results-Based-Financing-Programme-for-Northern-Uganda.pdf* .

WHO. (2016). *Sharing and debating country experiences on health financing: Fiscal sustainability and transition, public finance management, and results-based financing.* Meeting Brief.

WHO. (2016). Sharing and debating country experiences on health financing: Fiscal sustainability and transition, public finance management, and results-based financing. *Meeting Brief*. Retrieved from http://www.who.int/entity/health_financing/events/sharing-and-debating-health-financing-challenges-meeting-summary-13022017.pdf

Witter, S., Zaman, R., M, S., & Mistry, R. (2016). Evaluation of Delivering Reproductive Health Results (DRHR) through non state providers. *MSI/PSI Impact Evaluation Report, Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/533669/Delivering-Reproductive-Health-Results-Non-State-Providers-Pakistan1.pdf* .