EBA

WORKING PAPER MAY 2023

PAYMENT BY RESULTS IN AID: A REVIEW OF THE EVIDENCE

Geske Dijkstra

Payment by Results in Aid: A Review of the Evidence

Geske Dijkstra

Working Paper, May 2023

to

The Expert Group for Aid Studies (EBA)

The EBA Working Paper Series constitutes shorter overviews, surveys, mappings and analyses that have been undertaken to bring about discussion and advance knowledge of a particular topic. Working Papers are not subject to any formal approval process by the Expert Group. Just as in the EBA reports, authors are solely responsible for the content, conclusions and recommendations.

Please refer to the present report as: Dijkstra, G, *Payment by Results in Aid: A Review of the Evidence*, Working Paper, May 2023, The Expert Group for Aid Studies (EBA), Sweden.

This report can be downloaded free of charge at www.eba.se

Cover design by Julia Demchenko

Geske Dijkstra is emeritus professor of Governance and Global Development in the Department of Public Administration and Sociology (DPAS) of Erasmus University Rotterdam, The Netherlands. Next to her academic work, teaching and research, she has carried out numerous studies and evaluations for development aid agencies such as the Dutch Ministry of Foreign Affairs, the World Bank, the Swedish International Development Agency, the Development Centre of the OECD and the Inter-American Development Bank. These studies and evaluations were on topics like programme aid, social development and poverty reduction, debt relief, budget support, gender, and good governance. She published eight books and many articles in international peer-reviewed journals such as World Development, Journal of Development Studies, Development Policy Review, Evaluation and Program Planning, and Social Indicators Research. She is also associate editor of Feminist Economics.

Acknowledgements

I would like to thank Markus Burman from the EBA for having the idea of writing this report, and for his patience in waiting for the idea to be realized. I am also very grateful to Nicoletta Martelli, then Master student at Erasmus University Rotterdam, for helping with searching and classifying the relevant sources on payment by results. She provided excellent research assistance.

This report benefitted greatly from thoughtful and sharp comments from the Reference Group, consisting of Sara Johansson de Silva, Adam Öjdahl, Markus Burman and Jan Pettersson. It also benefitted from comments from the European and Global Governance group of the Department of Public Administration and Sociology of Erasmus University Rotterdam.

Table of Contents

List of acronyms

BGFA	Beyond the Grid Fund Africa
BGFZ	Beyond the Grid Fund Zambia
CCA	Clean Cooking Alliance
CCT	Conditional Cash Transfer programmes
CHF	Community Health Fund
COD	Cash on Delivery aid
DALY	Disability Adjusted Life Years
DDF	District Development Fund
DfID	Department for International Development
DEVAL	German Institute for Development Evaluation
DIB	Development Impact Bond
DLI	Disbursement Linked Indicators
DLT	District League Table
DPT	Diphtheria, Tetanus, Polio
FCDO	Foreign, Commonwealth and Development Office
GAVI	Global Alliance for Vaccines and Immunisations
GFATM	Global Fund to fight Aids, Tuberculosis and Malaria
GEC	Girls' Education Challenge
GPE	Global Partnership for Education
GPOBA	Global Partnership for Output-Based Aid
GPRBA	Global Partnership for Results-Based Aid
HRITF	Health Result Innovation Trust Fund
IADB	Inter-American Development Bank
ICR	Implementation Completion Report
IDA	International Development Association
IEG	Independent Evaluation Group
IFC	International Finance Corporation
ISS	Immunisation Strengthening Support
KfW	German Development Bank
KOICA	Korean International Cooperation Agency
LICUS	Low-Income Countries Under Stress
MCC	Millennium Challenge Corporation
MECS	Modern Energy Cooking Services

NGO	Non-Governmental Organisation
NORAD	Norwegian Agency for Development Cooperation
OBA	Output-Based Aid
PAF	Performance Assessment Framework
PbR	Payment by Results
PDL	Performance Driven Loan
PES	Payment for Environmental Services
PTR	Pupil Teacher Ratio
RBA	Results-Based Aid
RBF	Results-Based Finance
RCT	Randomized Controlled Trials
REDD+	Reducing Emissions from Deforestation and Forest Degradation
RMNCAH	Reproductive, Maternal, Newborn, Child, Adolescent Health and Nutrition
Sida	Swedish International Development Agency
SMI	Salud Mesoamerica Initiative
SSDP	School Sector Development Plan
UK	United Kingdom
USAID	US Agency for International Development

Foreword by EBA

Official International Development Assistance is a sector characterised by asymmetric information. Every relation in this sector includes some type of contract, often in the form of explicit agreements between a donor and a recipient where the parties agree on the donor paying for some inputs supposed to lead to the delivery of certain expected results or specified services or goods. At the same time, the party who pays does not really know the level of performance by the contractor, and the expected results may in themselves be difficult to measure. This becomes even more difficult when feedback from the intended beneficiaries is hard to obtain, which is often the case in development cooperation. Aid effectiveness could, a bit exaggerated, be described as an issue of whether contracts are properly designed and properly delivered on.

A set of performance-based contracts, collectively referred to as Payment by Results, implies that – at least part of – the resources are transferred after some pre-agreed results have been achieved. Given the challenges mentioned, this is likely not a panacea for aid effectiveness. But in which sectors or under what circumstances is this a viable method? What do we know about the effectiveness of Payment by Results?

In this working paper, Professor Emerita Geske Dijkstra reviews the available literature on Payment by Results in development cooperation, primarily in the health and education sectors. It is our hope that the report will find its audience among policy makers and managers in development cooperation and persons interested in aid effectiveness.

The EBA Working Paper Series constitutes shorter overviews, surveys, mappings, and analyses that have been undertaken to bring about discussion and advance knowledge of a particular topic. Working Papers are not subject to any formal approval process by the Expert Group. Just as in the EBA reports, authors are solely responsible for the content, conclusions, and recommendations.

Stockholm, May 2023

Jan Pettersson, Managing Director

Sammanfattning

Syfte och fokus

Denna studie sammanställer och analyserar befintlig akademisk litteratur och litteratur som beställts och publicerats av givare om det som kallas payment by results (PbR) inom internationellt bistånd, på svenska resultatbaserad finansiering. Idén med resultatbaserad finansiering är att givaren överför hela eller delar av finansieringen *efter* att överenskomna resultat har uppnåtts. Kontraktet kan ses som en *Principal-Agent*-relation där biståndsgivaren, ”principalen”, vill uppnå vissa resultat och mottagaren eller den implementerande organisationen, ”agenten”, får betalt först när resultat har observerats.

Studien undersöker särskilt den typ av resultatbaserad finansiering som kallas ”Results Based Aid” (RBA) där principalen är en givare och agenten ett mottagarland, samt ”Results Based Finance” (RBF) där principalen kan vara en givare, men också nationella och subnationella regeringar eller icke-statliga organisationer och den implementerande agenten lokala eller regionala regeringar, organisationer eller olika tjänsteleverantörer. Studien täcker de flesta sektorer där man hittills tillämpat resultatbaserad finansiering inom biståndet (hälsa, utbildning, kapacitetsbyggnad, energi, vattenfrågor) med undantag för miljöområdet.

Genomförande och metod

En litteratursökning gav totalt 867 resultat. Efter en genomgång enligt fastställda inkluderingskriterier återstod 48 studier.

Analysen utgår från ett teoretiskt ramverk med inriktning på förväntad nytta, potentiella risker och kostnader förknippade med resultatbaserad finansiering. Studien undersöker både insatsernas effektivitet och de uppnådda resultatens kvalitet. Dessutom undersöks oavsiktliga resultat som manipulation och felrapportering, men även anpassning till mottagarsidans system, transparens, ansvarsutkrävande, innovation, rättvis fördelning, kostnadseffektivitet och insatsens träffsäkerhet. Dessutom analyseras om principalens, det vill säga biståndsgivarens, beteende kan undergräva den förväntade nyttan med resultatbaserad finansiering.

Studiernas kvalitet

Precis som i tidigare litteratursammanställningar om resultatbaserad finansiering fokuserar flertalet studier på hälsosektorn. Majoriteten av de kvantitativa effektutvärderingarna har gjorts där. När resultatbaserad finansiering använts inom hälsoområdet har det ofta handlat om Results Based Finance (RBF), vilket underlättar rigorös effektutvärdering med randomisering. I sådana studier kan implementerande agenter placeras i försöks- respektive kontrollgrupper slumpmässigt. Trots det kan inte alla studier om RBF inom hälsosektorn beskrivas som rigorösa och de kvantitativa studierna lider ofta av en eller flera begränsningar. De använder sällan oberoende verifierade data eller undersöker resultat-manipulation och otillbörlig påverkan. Ofta bedöms inte heller effekten på andra variabler än de som insatsen direkt sökt påverka. En annan brist är att studierna inte bedömer om ”cherry picking” förekommit genom att insatsen till exempel inriktat sig på målgrupper som är lätta att nå. Studierna bedömer inte heller kostnadseffektivitet. Detta gör att man har stor nytta av kvalitativa studier. Inom övriga studerade biståndssektorer har resultatbaserad finansiering ofta utgjorts av Results Based Aid (RBA). Men eftersom RBA bygger på ett samarbete där den nationella regeringen är motpart är det svårt att åstadkomma en kontrafaktisk utvärderings-situation. Detta gäller också de typer av resultatbaserad finansiering som använt kommersiella aktörer för att tillhandahålla tjänster eller produkter för fattiga och utsatta, då dessa företag vanligtvis inte får biståndspengar eller subventioner.

De flesta studier undersöker minst ett effektivitetsmått, men eftersom det finns begränsad evidens om kvaliteten på resultat och andra potentiella fördelar, kostnader och risker, innebär det att även när det funnits evidens av god kvalitet gällande effektivitet är det svårt att helt fastslå värdet av resultatbaserad finansiering.

Övergripande resultat

Evidensen för att resultatbaserad finansiering är en effektiv biståndsform på nivån output, outcome och impact är blandad och motsägelsefull. Det gäller särskilt när man tar hänsyn till studiernas kvalitet. I den mån det finns evidens för att insatser med resultatbaserad finansiering ökar transparens, ansvarsutkrävande och innovation så är den inte entydig. Insatser som uttryckligt riktats till fattiga och utsatta är ofta framgångsrika i att åstadkomma rättvis fördelning av resurser. Men riskerna och de

möjliga kostnaderna med resultatbaserad finansiering är reella: de studier som rapporterat om oavsiktliga effekter pekar ofta på en eller flera sådana. De program där man inte säkerställt att rätt målgrupper faktiskt nås har dessutom ofta haft negativa effekter på jämlik och rättvis fördelning på grund av ”cherry picking” (av målgrupper). Det fåtal studier som bedömer kostnadseffektivitet drar olika slutsatser.

Slutsatser och lärdomar

Sammantaget dras slutsatsen att resultatbaserad finansiering varit effektivt i vissa fall och under vissa omständigheter, men att effektivitet och kostnadseffektivitet inte har säkerställts på generell nivå. Detta beror på typ av finansieringsmetod, sektor, risknivå, (negativa) oväntade effekter och kostnadsökningar. Det innebär att effektiviteten inte bara beror på hur insatserna är designade och i vilket sammanhang de genomförs – vilket också tidigare syntesstudier visat – utan även på vilken typ av resultatbaserad finansiering som använts och inom vilken sektor.

För programdesignen tycks framgångsfaktorer vara att berörda intressenter involveras, flexibilitet vid utformandet av (utbetalningskopplade) indikatorer, att de uppfattas som rättvisa och att de kan verifieras på ett oberoende sätt. Storleken på den finansiering som utgår när resultaten uppnåtts, liksom de finansiella incitamentens omfattning, har också betydelse. Resultatbaserad finansiering förefaller svårare att tillämpa i miljöer med låg kapacitet och små resurser. Detta gäller dock för alla typer av bistånd. Dessutom måste principalen, det vill säga biståndsgivaren, vara beredd att hålla inne pengar vid utebliven prestation. Detta förefaller svårare vid RBA än vid RBF och i synnerhet för bilaterala givare.

Resultatbaserad finansiering tycks mest lämpligt att använda om målbilden för principal och agent inte överensstämmer i utgångsläget. Biståndsformen verkar vara särskilt framgångsrik i att förmå kommersiella agenter att leverera varor eller tjänster till fattiga och utsatta målgrupper. Ofta vet vi emellertid inte om dessa programupplägg har oavsiktliga effekter (till exempel på resultatens kvalitet eller på angränsande områden), eller om de är kostnadseffektiva i jämförelse med andra alternativa insatser. Om incitament och motiv skiljer sig åt mellan principal och agent kommer också riskerna för manipulation, felrapportering eller otillbörlig påverkan på resultatindikatorer att öka.

De största utmaningarna med resultatbaserad finansiering rör tre avvägningar i utformningen av resultatindikatorer. För det första är det svårt att hitta indikatorer som agenten kan påverka i tillräckligt hög grad, som stimulerar autonomi och innovation och leder till övergripande mål. Exempelvis kan incitament för att få barn till skolan i ett låginkomstland (vilket agenten kan påverka) inte självklart kopplas till bättre kunskaper hos elever. På samma sätt är det inte säkert att indikatorer i den nedre delen av resultatkedjan (på stegen mellan aktivitet och slutresultat), till exempel angående antal besökare på en förlossningsklinik, skapar incitament för innovation för minskad mödradödlighet. För det andra finns det en avvägning mellan förutsägbar utbetalning och incitament för god prestation. Givare tenderar att välja lättåtkomliga och närliggande indikatorer på process- eller outputnivå för att säkerställa utbetalning, men detta stimulerar troligen inte till bättre prestationer. För det tredje: ju starkare de ekonomiska incitamenten blir, desto större är risken för felrapportering eller annan otillbörlig påverkan på indikatorerna.

Dessa avvägningar varierar beroende på sektor, typer av agent/genomförare och kontext. Den första och andra avvägningen är särskilt allvarlig i sektorer med lång resultatkedja, till exempel inom hälsa och utbildning, medan den är mindre allvarlig inom andra sektorer, som vatten eller energi. Dessutom är den första avvägningen troligen mer allvarlig inom utbildning än hälsa, med tanke på den högre säkerheten om samband mellan kortsiktiga och långsiktiga resultat inom hälsa än inom utbildning. Det andra dilemmat är allvarligare om det finns ett större ömsesidigt beroende mellan principal och agent, till exempel i relationen mellan givar- och mottagarland. Dilemmat är mindre starkt för civilsamhällesorganisationer och frånvarande om agenten är en kommersiell aktör. Den tredje avvägningen är allvarligare när agenten ska engagera sig i multipla eller komplexa uppgifter, och när resultaten är svåra att mäta. Dessa två omständigheter gäller för hälsa och utbildning mer än övriga sektorer.

Det verkar alltså som att resultatbaserad finansiering kan ha ett värde om det kan mobilisera privata resurser för att leverera varor och tjänster till fattiga och utsatta inom sektorer som energi eller vattenförsörjning, där resultatkedjan är kort, resultaten konkreta och kan mätas förhållandevis lätt. I andra sektorer och situationer är mervärdet av resultatbaserad finansiering mer osäkert. Om principal och agent kan förväntas ha samma mål (exempelvis att förbättra inlärningen hos barn i grundskolan) finns det inget behov av att utbetalningar görs beroende av kostsamma och osäkra definitioner och indikatormätningar. Resultatbaserad finansiering kan vara

värdefullt i specifika sammanhang, till exempel om det finns ett klart behov av att förbättra servicen i något avseende för utsatta grupper. I flertalet fall är dock genomföraren i behov av resurser för att uppnå de mål som samtidigt är gemensamma med principalen. Flexibilitet och innovation hos agenten kan då främjas genom kärn- eller budgetstöd.

Sammantaget tycks drivkraften bakom resultatbaserad finansiering inom sektorer som hälsa och utbildning vara mer föranledd av givarsidans behov att ”visa resultat” än överväganden om biståndseffektivitet.

Summary

Aim and scope

This review assesses available literature on payment by results (PbR) in development cooperation. The basic idea of PbR is that the donor transfers – at least part of – the resources after some pre-agreed results have been achieved. The aid contract can be seen as a Principal-Agent relationship, in which the donor is the "principal" who wants to achieve certain objectives, and the recipient is the implementing agency, or the "agent", who will be rewarded when certain results have been achieved. This report looks in particular at Results Based Aid (RBA), where the principal is a donor and the agent is a recipient country, and Results Based Finance (RBF), where the principals can be donors but also national or sub-national governments or NGOs, and implementing agents can be sub-national governments, private sector organisations or (public or private) service providers, and individuals. It covers almost all sectors in which PbR has been applied (health, education, governance, energy, water), with the exception of the environment (agriculture and forestry).

Methodology

An extensive search in Web of Science, Scopus, Google Scholar, and websites of relevant donor organisations led to 867 results. After applying exclusion and inclusion criteria, 48 sources on PbR remained. These include 20 reviews (including nine literature reviews and 11 discussions of multiple cases for which literature has been used), and 28 primary studies, of which 21 discuss a single case and seven discuss multiple cases.

The report develops a theoretical framework of expected benefits and possible risks and costs of PbR, which is used when discussing the sources. This means the report examines effectiveness (at input, output, outcome, and/or impact level, plus quality), unintended effects like gaming, manipulation and distortions, extent of alignment (of principal and agent, and of multiple principals), transparency and accountability, innovation, targeting and equity, and cost effectiveness. In addition, the report assesses whether the behaviour of the principal does not hinder expected benefits.

Quality of included studies

In line with other reviews of PbR, most studies are on the health sector. Most quantitative evaluations of effectiveness also proved to be on the health sector. PbR in health is mostly RBF, which facilitates high-standard evaluations of effectiveness like randomized controlled trials, where implementing agents can be assigned randomly to treatment and control groups. Nevertheless, not all studies on RBF in health proved to be rigourous, and quantitative studies usually suffer from one or more limitations: they do not use data that have been verified independently, they do not examine whether gaming or manipulation has occurred, they do not assess the effects on results that were not incentivized, they do not assess whether "cherry picking" has occurred with negative effects on equity, and they do not assess cost effectiveness. This means that qualitative studies also have value. In other sectors, PbR is often RBA. Since this is a contract with the national government, applying rigorous counterfactuals is more difficult. The same holds for forms of PbR that incentivize commercial actors to provide services for poor and vulnerable populations, since private sector agents usually do not receive aid or subsidies.

Most sources report on at least one measure of effectiveness, but there is much less evidence on quality and all other possible benefits, costs and risks (in Tables 7–10, all columns on the right side of "quality"). This means that even with good quality evidence on effectiveness for outputs, outcomes or impact, the value added of PbR is not certain.

Results

The evidence on the effectiveness of PbR for outputs, outcomes and impact is mixed, especially when taking into account the quality of studies. To the extent there is evidence on improved transparency and accountability, and on innovation, it is mixed as well. Schemes that explicitly target poor and vulnerable groups were often successful in improving equity. But the risks and costs involved in PbR are real: the studies that do report on unintended effects often report one or more of these effects. In addition, schemes that do not apply explicit targeting often have negative effects on equity due to "cherry picking". And while there is little information on cost effectiveness, the few studies that do report on it have mixed results as well.

Conclusions

It can be concluded that PbR appears to have been effective in some cases and circumstances, but that effectiveness, let alone cost effectiveness (or value added) of PbR in general is by no means certain. This is due to possible risks, to negative effects, and extra costs. However, these risks and costs vary by type of PbR and by sector. This means that the effectiveness of PbR not only depends on design and context, as concluded by many other reviews, but also on type and sector.

In terms of design, key factors appear to be the involvement of all stakeholders, flexibility in setting indicators, a perception of fairness of indicators, and independent verification. The amount of extra funding and the size of the incentives also matter, while accompanying technical assistance often increased effectiveness. In terms of context, it can be concluded that PbR is more challenging in low capacity and low resources environments. But this holds for all interventions. In addition, agents at sub-national level must have some level of autonomy. Furthermore, the principal must be willing to withhold the money in case of non-performance. This seems more difficult in RBA than in RBF, and in particular for bilateral donors.

PbR seems to be most appropriate if the objectives of principal and agent are not fully aligned at the outset. In particular, it appears to be successful in inducing (commercial) agents to deliver goods and services to poor and vulnerable groups. However, we often do not know whether these programmes have unintended effects (for example, on quality or on other service areas) or whether they are cost effective in comparison with other interventions. The initial non-alignment of principal and agent will also increase the risks of gaming and manipulation.

The main risks and costs of PbR result from three trade-offs in establishing performance indicators. First, it is difficult to establish indicators that can be sufficiently influenced by the agents *and* lead to the ultimate objective *plus* stimulate autonomy and innovation. For example, incentivizing school enrolment rates (that can be influenced by the agent) may not lead to better educated pupils. At the same time, indicators at the lower end of the results chain (the steps between activities and final results), like number of antenatal visits, do not incentivize innovation for achieving reduced maternal mortality (a final result). Second, there is a trade-off between predictability of disbursement *and* incentivizing performance. Donors tend to select easy-to-reach indicators at process or

output level in order to safeguard disbursements, but these indicators hardly stimulate performance. Third, the higher the material incentive, the higher are the chances of distortion and gaming.

However, these trade-offs vary by sector, types of agents and circumstances. The first and second trade-off are particularly serious in sectors with a long results chain (many steps between activities and final results), for example in health and education, and much less in other sectors, like water or energy. Moreover, the first trade-off is probably more severe in education than in health, given the somewhat higher certainty about the relationship between outputs and outcomes in health than in education. The second dilemma is more severe if there is greater mutual dependence between principal and agent, for example in an aid relationship between a donor country and a low-income recipient country. It is usually weaker in case of an NGO and absent if the agent is a commercial actor. The third trade-off is more serious when agents are supposed to engage in multiple and complex tasks, and when the results of these tasks are not easily measurable. These two circumstances hold for health and education more than for other sectors.

It seems that PbR can have an added value if it can mobilize private resources for delivering goods and services to the poor, and especially in sectors, like energy or water, where the results chain is short and the results are tangible and can be measured relatively easily. In other sectors and situations, the added value of PbR is much less certain. If principals and agents can be expected to have the same objectives (say, improving learning outcomes), there is no need for having disbursements depend on costly and risky definitions and measurements of targets. PbR can be valuable in specific contexts, for example if there is a need to improve services for the poor or for vulnerable groups like mothers and new-borns. In most cases, implementing agents are in need of resources in order to achieve the – shared – objectives. Flexibility and innovation of agents can be fostered by providing them with core financing, or budget support. All in all, the drive to payment by results in sectors like health and education often seems to be more induced by domestic motivations to "show results" than by considerations of aid effectiveness.

1 Introduction

The aim of this report is to provide an overview of the evidence of the effectiveness of payment by results (PbR) in international cooperation. PbR has many different forms, but the basic idea is that the donor transfers aid funds *after* some earlier specified results have been achieved. The aid contract can be seen as a Principal-Agent Relationship, in which the donor is the "principal" who wants to achieve certain objectives, and the recipient is the implementing agency, or the "agent", who will be rewarded when certain results have been achieved.

The main motivation for PbR is the need for donors to show results from their aid efforts. In project aid, the money is usually provided for inputs and it is not clear whether the ultimate objectives, the desired outcomes and impact, will be achieved. With PbR, at least part of the aid money will only be disbursed when the results are achieved. This is expected to allow for greater accountability to tax payers and ultimate beneficiaries, while at the same time making the recipient more accountable for results. This drive for results can be seen as part of wider 'New Public Management' (NPM) reforms in the public sector and, in particular, in the delivery of public services (DfID, 2014; Pereira & Villota, 2012). In the aid sector, these reforms were reflected in the "managing for results" principle of the 2005 Paris Declaration. The adoption of this principle stimulated PbR, while more critical stances among domestic constituencies about aid budgets also played an important role (Janus, 2014; Pereira & Villota, 2012).

PbR can also be seen as implementing 'ex post' conditionality, in response to the failure of 'ex ante' conditionality where recipients promised to carry out certain policies in return for aid. In practice, the promised reforms were often not implemented or only implemented cosmetically (Collier et al., 1997, Killick et al., 1998; Dijkstra, 2002). Another motivation for introducing payment by results is that it would lead to more ownership of the recipient and to more flexibility and innovation in how to achieve the aid objectives (DfID, 2014).

Around 2000, budget support was seen as the answer to the failure of 'ex ante' conditionality and as a way to foster recipient ownership. In this aid modality, the resources are freely spendable but recipient countries must meet certain eligibility criteria and respect certain "underlying principles": 'ex post' conditionality. In addition, this aid is accompanied by a policy dialogue on results to be achieved. In practice, however, eligibility criteria were not always met and donors began to use the policy dialogue for

ensuring that "underlying principles", like fighting corruption and holding free and fair elections, became respected. The lack of success of this return to 'ex ante' conditionality led to suspensions of budget support and to increased disillusionment with this aid modality (Hayman, 2011; Dijkstra et al., 2012; Swedlund, 2013; Molenaers et al., 2010). When many bilateral donors discontinued budget support around 2010, payment by results came to be seen for some as an attractive alternative. The British Department for International Development (DfID),[1] was one of the pioneers of payment by results, along with the World Bank. Other donors active in RBA/RBF include the Inter-American Development Bank (IADB), the Asian Development Bank (ADB), The German development bank KfW (Kreditanstalt für Wiederaufbau), the Swedish International Development Agency (Sida), Norway, and the Dutch NGO Cordaid.

This report reviews the evidence on PbR. It is not a systematic review as defined by Cochrane or the Campbell collaboration, but I attempted to find and select sources in a systematic way as discussed in the methodology section below. Although the number and quality of studies and evaluations on PbR has increased over time, recent systematic reviews in health, the sector for which most evaluations are available, still conclude that the evidence base is small (Diaconu et al., 2021; Singh et al., 2021). Some studies just present whether the targeted indicators have been achieved and whether the money has been disbursed, and do not include independent verification of these indicators. Other studies apply a simple before-after design without using a rigorous counterfactual. In presenting the results on effectiveness of payment by results, I will indicate the quality of the studies and give priority to those studies with the most rigorous evaluation methods. But I will also discuss less rigorous and qualitative studies, as they provide other information on the value of payment by results, for example on *how* it works or doesn't work, or on other objectives of PbR like fostering innovation.

This report first outlines the different forms of PbR and then discusses the theoretically expected advantages and risks. This leads to an extensive theoretical framework that is then applied in the chapters with results. Chapter 4 contains a description of the methodology for searching sources and a first classification of the sources found. The evidence on the results is presented in three chapters: one on the heath sector (5), one on education (6), and one on other sectors and bigger and hybrid schemes in payment by results (7). Chapter 8 concludes.

[1] Now called Foreign, Commonwealth and Development Office (FCDO).

2 Types of payment by results

There are different types of payment by results in aid, and the definitions used are not always the same. Most authors agree on the differences between Results-Based Aid (RBA) and Results-Based Finance (RBF) (Grittner, 2013; Musgrove, 2011; Pearson, 2011). In RBA, the donor is the funding source and the recipient is the national government in the recipient country. Examples include the Global Alliance for Vaccines and Immunisation (GAVI)[2] where countries receive an amount of money for additional children vaccinated, and results-based aid in education where countries are rewarded when specific output or outcomes in education are achieved. The Millennium Challenge Corporation (MCC) of the US can also be seen as RBA, as countries must meet strict eligibility criteria for receiving this aid. As referred to above, the eligibility criteria for budget support were not very strictly maintained. But some donors, most notably the EU, link part of the resources in budget support to the achievement of certain results.These so-called "variable tranches" in budget support can be considered a form of payment by results.

In RBF, the principals can be donors but also national or sub-national governments or NGOs. [3] Implementing agents can be sub-national governments, private sector organisations or (public or private) service providers, but also individuals. According to some authors, RBF not only includes schemes with incentives on the supply side (providers of services), but also on the demand side (Grittner, 2013; Helland & Mæstad, 2015; Pearson, 2011). On the supply side, incentives can be provided to the implementing agency or to individual workers in these agencies, for instance health or education workers.

Providing incentives on the demand side may include Conditional Cash Transfer programmes (CCT) where individuals or household are paid provided they use certain services, and voucher schemes. Voucher schemes are used mainly in health. For example, targeted women receive a voucher for a facility-based delivery. The facilities are reimbursed once they can show they have delivered the service. However, following Musgrove (2011), this report does not include voucher schemes since vouchers are given 'ex ante' and do not imply any incentive, which is the key feature of PbR. And although CCTs do imply an incentive for behavioural change, we do not include them in this report either, since there is already abundant information available about their effectiveness.

[2] Now called Gavi, the vaccine alliance.

[3] But, in order to be included in this report, some aid money must be involved.

DfID (2014) brings up another modality, namely Development Impact Bonds (DIB). In DIB, a private investor provides financing to an intermediary. This intermediary pays service providers that target a specific population. Once a validating agency has confirmed that the pre-agreed results in the target population are achieved, the donor (or other principal) reimburses the investor. In theory, this may lead to a market for social investment. But given that DIBs can take many different forms, that there are many more different actors involved than with RBA or RBF, and that there is not much evidence yet, DIBs are not included in this report.

There are also hybrid schemes combining elements of RBA and RBF, like the Global Fund to Fight Aids, Tuberculosis and Malaria (GFATM), Output Based Aid (OBA)[4] stimulated by the Global Partnership for Output-Based Aid (GPOBA), and the Health Result Innovation Trust Fund (HRITF) (Grittner, 2013). Unlike GAVI, GFATM can be considered a hybrid scheme as the agents can be government, private sector organisations or NGOs. In GFATM, good performance in the first 2 years of a scheme is rewarded by continued grants in years 3–5 (Pearson et al., 2010).

GPOBA is the first of a series of multi-donor trust funds coordinated by the World Bank. OBA was formally introduced in 2003, and in the same year GPOBA was set up. It stimulates output-based aid for delivering basic services to the poor in developing countries. These include access to water and sanitation, energy, health care, education, communication services, and transport. Aid money takes the form of a subsidy covering the difference between the full cost and the price that poor users can afford. One feature of OBA is that the incentive is expected to be additional to private or public funding and that it leverages such extra funding. Service providers, often private sector agents, are rewarded after delivering the services to the poor. The aid contract can be with a national government, or with lower level government, NGOs or the private sector (Mumssen et al., 2010; Musgrove, 2011). This makes OBA a hybrid between RBF and RBA.

Mumssen et al. (2010) report that there were 32 OBA projects in 2003, mostly in Latin America and the Caribbean, and that this number increased to 131 in 2010 (of which 34 closed, 78 ongoing and 19 in design stage). In 2019, GPOBA was renamed GPRBA (Global Partnership for Results Based Approaches). It is a donor-funded pilot programme

[4] "Output" in OBA refers to both output and outcomes (Musgrove, 2011:2).

administered by the World Bank aiming to expand OBA in countries funded by the World Bank window for low income countries, IDA (International Development Association). The GPRBA/IDA portfolio of OBA projects currently contains 58 projects (closed and ongoing).[5]

HRITF is another World Bank Trust Fund that finances design, implementation and evaluation of pilots of RBF in the health sector of developing countries. Agents often include sub-national governments and health facilities and/or health workers but sometimes also national governments (Helland & Mæstad, 2015). The schemes have been extensively evaluated and many of these evaluations are covered in the systematic reviews to be discussed in chapter 5 of this report. In 2015, the World Bank established a similar Trust Fund for education, REACH (Results in Education for All Children). It focuses on IDA countries and is co-financed by the Norwegian Agency for Development Cooperation (NORAD) and USAID (Hill et al., 2015). Like HRITF, a key aim of REACH is to increase the evidence base on RBF/RBA in education. By 2019, 33 projects in 23 countries were financed (Lee and Medina, 2019:13). The Dutch NGO Cordaid has financed RBF in health since 2001, and has started some RBF in education. Their schemes often operate in fragile and violence or conflict affected areas (Lee and Medina, 2019:53).

Helland and Mæstad (2015) also discuss PbR as an element of Norwegian bilateral cooperation aimed at reducing CO_2 Emissions from Deforestation and Forest Degradation (REDD+). Studies on PbR (often RBA) in the context of REDD+, and also the literature on Payment for Environmental services (PES, both RBA and RBF), is not discussed in this review. This is decided in order to keep this review manageable, and also because these are very specific forms of RBA and RBF with a huge amount of challenges (Angelsen, 2014).

[5] Who We Are | The Global Partnership for Results-Based Approaches (GPRBA), accessed 20 September 2022.

Table 1. Some of the bigger RBA and hybrid schemes covered in this report

Name	Donors/funders	Years
MCC	US	2004–
Variable tranches in budget support	EU	1999–
GAVI	Official donors and private sector organisations	2000–
GFATM	Official donors, private sector organisations and NGOs	2002–
OBA/GPOBA	In 2010: World Bank, Australia, EU, IFC Netherlands, Sweden, UK	2003–2019
OBA/GPRBA (for IDA countries only)	In 2022: World Bank, Australia, IFC, Netherlands, Sweden, UK	2019–
HRITF	World Bank, Norway, UK	2007
REACH, for IDA countries only	World Bank, Norway, USAID	2015
Program for Results of the World Bank	World Bank	

Source: The author.

This review incorporates studies on RBF, on RBA, on studies that combine the two and on hybrid schemes (Table 1). As explained above, it will not consider DIBs, REDD +, PES, CCTs and voucher schemes. Yet, to the extent that some RBF modalities focusing on the supply side also include incentives for stimulating demand (like outreach activities or reduction of user fees) they are included in this review.

3 Theory: expected advantages and risks of payment by results

The two main features of a principal-agent relationship are that the incentives between principal and agent may not be fully aligned, and that the principal cannot observe the agent's efforts (Hazeu, 2000). Payment by results can solve these issues: by paying for certain pre-agreed results, the agent will have an incentive to act in line with the objectives of the principal; if these pre-agreed results reflect the objectives of the principal, the principal does not need to observe the agent's efforts.

The alignment of objectives between principal and agent is therefore a first expected advantage of payment by results in aid. However, this assumes that there is no agreement on priorities before the aid contract. It can be expected that the extent of 'ex ante' alignment will vary among different principal-agent relationships. Agents from public sector and NGOs are more likely to share objectives with the donor/principal, while this is less likely for agents from the private sector. 'Ex ante' alignment might also be weaker for individual health workers or teachers. To the extent that objectives between principal and agent are fully aligned, payment by results is not necessary for achieving results and may only increase risks and costs.

A second expected advantage is that there will be a stronger focus on results and on performance in aid. The implementing agency will dedicate more efforts to achieving these results. Third, this focus on concrete results will increase transparency and accountability of aid, both to taxpayers and to ultimate beneficiaries. Fourth, the focus on results is expected to lead to flexibility and discretion for the agent in how it can achieve the results (for example, with which inputs and how many), and this is expected to lead to more innovation, and also to more efficiency and cost effectiveness (DFID, 2014). However, this fourth advantage only holds in comparison with project aid and not in comparison with budget support. Budget support involves the same discretion for the agent and, provided that donor and recipient have the same objectives, it potentially brings the same effectiveness and efficiency advantages as compared to project aid.

The literature on payment by results has revealed that some of these benefits, like more results and more innovation will not come forward automatically, and that there are also risks and costs involved (Clist & Verschoor, 2014; DFID, 2014; Dom et al., 2021; Paul, 2015).

First, the level in the results chain at which the measure is defined matters. In order to foster flexibility and innovation, it should be defined at outcome or impact level. However, this implies less control over the result for the agent, as many other factors may influence the result. Measures at outcome or impact level tend to have a high "signal to noise ratio": they are weaker indicators for the efforts of the agent. This means a higher risk of non-payment in spite of extensive efforts (a risk for both principal and agent), as well as a higher risk of payment for limited efforts (risk for the principal). In practice, these risk considerations often lead to measures at output or process (or even input) level, at the cost of the advantage of fostering innovation.

Second, the quality of the performance measure matters. The indicator should be measurable at sufficiently short time intervals, and it should correspond as much as possible with the ultimate objective of the donor, also after it has become the target. The latter is difficult. Goodhart's law, already formulated in 1975 when commenting on monetary policy in the UK, is likely to apply: "When a measure becomes the target, it ceases to be a good measure" (cited in Eldridge & Palmer, 2009:164). Suppose that the ultimate objective is to have a well-qualified workforce, to be achieved through high quality education of as many children as possible. The indicator can then be, for example, high completion rates. High completion rates are expected to result in a large number of better qualified workers. However, in order to increase completion rates, government or schools can choose either to improve quality of education, or to reduce repetition rates. It is obviously easier to do the latter, but this will reduce the quality of education and will not bring about the ultimate objective. 'Ex ante', completion rates seemed to be a good proxy for a well-qualified workforce, but 'ex post', after it became the target, it no longer is (Clist & Verschoor, 2014:11).

There are several other possible distortions or unintended consequences of rewarding performance based on one or more specific measures. A focus on pre-defined results may lead to lower efforts dedicated to other valuable goals. Most notably, a focus on quantity tends to lead to lower quality, but neglects of other valuable goals are also possible. A focus on specific targets may also lead to gaming and manipulation: agents will make sure the targets are met, if necessary by manipulating the numbers (cheating) or by artificially inflating them through non-valuable actions. In addition, the drive to achieve the pre-defined results may lead to "cherry-picking": agents will focus their efforts on easiest to reach regions or target groups, with negative effects on equity. Lastly, there may be "hidden costs of control".

Agents may perceive this control of performance as distrust and may lower their efforts in consequence. The literature also reveals cases in which this attempt to increase extrinsic motivation crowds out the intrinsic motivation, resulting in lower overall goal achievement. This holds, in particular, when incentives are targeted to individuals.

Some PbR schemes explicitly target poorer segments of the population. This means the reward will only be provided if this target population has been shown to benefit. The success for improving equity depends, among other things, on the effectiveness of the targeting mechanism. In schemes where results are rewarded without explicit targeting, equity may be affected negatively because the agent may engage in "cherry picking" in order to achieve the targets more easily.

In order for PbR to work, there are also issues related to the agent and to the principal (Clist, 2016). The agent can be more or less risk averse, and this will influence the extent to which she discounts the expected payments by the principal. The extent to which the agent perceives the measures as acceptable and fair also matters, as well as the extent to which agents have more intrinsic or more extrinsic motivation. With more intrinsic motivation (more alignment), PbR is less needed and implies a risk that agents will focus their efforts on the measures rather than the ultimate objectives (Clist & Verschoor, 2014:15). In addition, in order for innovation to occur, agents must have a sufficient level of autonomy, and sufficient resources and capacities.

As regards the principal, two issues are important. First, the principal must be willing to withhold aid if the targets are not met, otherwise the incentive will not work. The experiences with earlier aid conditionality (both "ex ante" and "ex post") do not augur well for this. Donors tend to continue disbursing regardless of whether conditions are met, for two reasons: the "Samaritan's dilemma" (those not meeting the conditions are usually the poorest, so it is difficult to stop funding), and institutional reasons within donor organisations, in particular the pressure to disburse. McGillivray and Pham (2017) argue that performance-based aid allocation as carried out by the MCC or the World Bank suffers from some fundamental problems. The neediest countries may be the worst performing, and donors should take into account, in particular, countries' lack of human capital and economic vulnerability (McGillivray & Pham, 2017). Second, and especially in order to stimulate innovation, the principal must have a sufficiently long time frame, of at least five years. This is usually also difficult for donors (Clist & Verschoor, 2014).

Payment by results on the basis of pre-agreed targets also implies extra costs. First, there are the costs of the tariff or bonus to be paid. Depending on the risk aversion of the implementing agency, the donor will have to pay a risk premium in order to compensate for the uncertainty involved in the payments (Clist & Verschoor, 2014:5). The costs depend on the tariff paid per agreed result: the higher the tariff, the higher the incentive but also the higher the cost and the higher the likelihood of distortion and gaming. Usually, there is information asymmetry: the agent knows better how much effort are needed to achieve a certain result (Clist and Verschoor, 2014:17). This leads to the paradox that donors prefer payment by results because they cannot observe the actions of the recipient and don't know which and how many actions are necessary to achieve the objective, but in order to make a good contract, they need to have good knowledge of these required efforts and actions (Clist, 2016:309). Second, there are the transaction costs of preparing the aid contract, but this holds for all aid contracts and they are not necessarily higher for payment by results.

Third, there are always verification costs. Preferably, and in order to reduce the chances of gaming, verification is not left to the agent but to an independent agency, but this will increase costs. Although other aid modalities usually also involve monitoring, the monitoring costs in payment by results will be higher since the stakes are higher: more is required in terms of accurateness and timeliness of measurement, and the higher stakes imply a higher likelihood of gaming and manipulation. On the other hand, the (higher) emphasis on measuring results may foster better data systems, leading to improved monitoring and evaluation of policies. If PbR targets the poor, targeting costs are also involved.

Table 2. Advantages, mechanisms, risks, costs of payment by results, and conditions related to agent and principal for advantages to come about

Advantages	Mechanisms	Risks	Costs	Conditions agent	Conditions principal
Effectiveness: results on outputs, outcomes, impact	Incentives work: efforts to achieve results increase	Measure not in line with objective Distortion, gaming Distrust, loss of intrinsic motivation	Costs of incentives: risk premium	Agent can influence result Has some extrinsic motivation, is not much risk averse, perceives measures as fair	Willing to withhold finance
Alignment of objectives P-A	Discussion with all stakeholders; good contract		Preparation costs	Involved in preparation	Willing to align
Transparency and accountability	Good data systems and/or independent verification, possible positive effect on data systems	Hidden costs of control: distrust, loss of intrinsic motivation	Costs of verification		
Innovation	Results must be defined at outcome or impact level			Agents have autonomy, skills, resources	Long-term commitment

Equity	Performance is easier for richer countries (e.g. in MCC) or vis-a-vis richer districts/ municipalities/ clients	Affects equity (cherry picking), poorest clients or countries excluded	
Targeting	Good targeting mechanism		Costs of targeting

Source: The author.

Table 2 gives an overview of expected advantages, mechanisms, costs and risks involved, and conditions related to principal and agent. The table also serves as basis for structuring the empirical evidence later in this report. Each chapter starts with discussing evidence on effectiveness, including evidence of gaming, manipulation and distortions. If available, this will be followed by evidence on alignment, transparency and accountability, innovation, equity and targeting, costs effectiveness and conditions related to the principal. Under alignment, this report discusses not only the alignment between principal and agent, but also possible effects of PbR programmes on aligning all actors: different government agencies, and donor alignment and harmonization. Under transparency and accountability, possible effects on data systems are also discussed.

4 Methodology for literature search

4.1 Search strategy

The aim was to find publications on payments by results in aid. We started searching in two important social science databases, Web of Science and Scopus, with the following search terms:

- Cash on delivery
- Output-based financ*, output-based aid, output-based loans, output-based lending
- Payment by results, payment for results, payment for performance
- Performance-based financ*, performance-based aid, performance-based loans, performance-based lending
- Program for results, program* for results
- Results based financ*, results-based aid, results-based loans, results-based lending
- Result, variable, output, outcome, performance *with* tranche

In the search strings, these terms were combined with "aid" *or* "international cooperation" *not* "aids". We searched in English language only, and in title, keyword and abstracts. We then performed a search in Google Scholar, with a slightly adjusted search string. Since the number of results for some of the terms sometimes became very large (>100), we decided to limit the search in those cases to the title only.[6] This is indicated in the list below by adding "(title)":

- Cash on delivery (title), cash on delivery aid (title), cash on delivery aid *not* aids
- Output based financ*, output-based finance, output based financing (title), output based aid (title), output based loans, output based lending
- Payment for results (title), payment for results aid *not* aids, payment by results (title), payment by results aid *not* aids, payment for performance (title), payment for performance aid *not* aids

[6] The risk of missing relevant sources was probably small, as the title of relevant studies (those that did not have to be excluded later) usually included one of the search terms.

- Program for results (title), program* for results, programme for results (title)
- Performance based financ*, performance based finance (title), performance based financing in aid *not* aids (title), performance based aid *not* aids (title), performance based loans, performance based lending
- Results based financ* (title), results based financing in aid *not* aids, results based finance (title), results based finance aid *not* aids
- Variable tranches (title), variable tranche

In addition, we searched in institutional databases, in particular from the World Bank, the Inter-American Development Bank, the Asian Development Bank, the African Development Bank (in particular IDEV, Independent Development Evaluation), and KfW. Some further references were obtained by contacting staff from IEG (World Bank), Sida Evaluation Unit, IOB (Netherlands), KfW Evaluation department, and DEVAL (Germany).

The search produced 867 results. In a next step, we started reading titles and, if necessary, abstracts, in order to apply criteria for exclusion. These include:

- Duplicates
- Sources that are not on aid projects or - programs
- Sources that are not on payment by results/output/performance
- Sources that do not focus on low or middle income countries
- Sources for which payment by results is only a very small part of content
- Project documents (World Bank)
- Drafts for comments
- Master theses
- Power point presentations
- Sources on REDD+ and PES
- Sources on voucher schemes and CCTs
- Sources for which we cannot access full text

After removing duplicates and applying the other exclusion criteria, 174 sources remained. The remaining references were articles, chapters, books and reports on aid, on payment by results (or similar) and on low or middle income countries. We then read abstracts and sometimes full sources in order to apply the inclusion criteria:

- The source should provide new empirical evidence on how payment by results in aid works (how it is implemented), and/or what the effects are on achieving one or more objectives of the aid project/program.
- In the studies providing original evidence, there should be some discussion of the methodology used to come to the evidence (in order to provide a basis for assessing the quality of the methodology).
- In addition, reviews of the evidence on how payment by results works and what its effects are on achieving the aid objectives should be included.

The second inclusion criterion meant that, among other studies, 32 short pieces on "best practices" and "lessons leant" for the application of, usually, "output-based aid" (so called "OBA-notes") were left out. All in all, 45 sources remained. During the process of reading and writing, I found some other relevant reference and I added them to the list. The total number became 48. The sources were coded according to the following criteria: academic versus grey; RBA, RBF, combination or hybrid; whether they were reviews of the literature, analyses of several cases or single case studies; sector; and academic versus grey. Grey publications are studies published or commissioned by an agency involved in the implementation of PbR. It is sometimes expected that they are less rigorous than academic publications, which are usually subject to more extensive quality review.

4.2 Quantitative results of literature search

Nine studies were or included reviews of the literature on PbR); four of them were systematic reviews,[7] meaning that they followed a strict protocol for the search process (Table 3). In addition, there were eleven studies that aimed to analyse multiple experiences or cases of PbR. These studies used secondary sources (literature) in addition to primary sources on these cases. The other studies either discussed one case or more than one, but did so on the basis of primary sources only. Most sources (29, or 60%) were academic studies and these included 2 PhD theses.

[7] One PhD thesis, listed as case study in Tables 3 and 4, includes a fifth systematic review (on cost effectiveness).

Table 3. Classification of sources by type: academic versus grey, reviews versus case studies

	Academic	Grey	Total	In %
Reviews[1]	5	4	9	19
Cases + review	2	9	11	23
Multiple cases	4	3	7	15
Single case	18	3	21	44
Total	29	19	48	100

Source: The author.

[1] Including three studies that *contain* a literature review. In order to avoid double counting, one single-case PhD study containing a limited literature review (discussed in section 5.1) is only classified under Single cases.

The vast majority of studies (60%) were about the health sector. The second category is that of analyses of PbR in multiple sectors (17%). These sources always include experiences in the health sector. The next category is education, with five studies (Table 4).

Table 4. Classification of sources by sector and type

	Reviews	Cases+ review	Multiple cases	Single case	Total	Percent
Health	6	1	5	17	29	60
Education	3		1	1	5	10
Water		1			1	2
Energy		1		1	2	4
Governance			1	2	3	6
Multiple		8			8	17
Total	9	11	7	21	48	100

Source: The author.

The large attention to the health sector is in line with a systematic review that registered the type of evidence available: almost 50% of studies was on the health sector. This review includes CCTs, voucher schemes and studies on the environment, mainly PES. When excluding vouchers and CCTs, there is an even larger dominance of studies in health, followed at a big distance by those in education (Meuth Alldredge et al., 2020) (p. 23).[8] This dominance of the health sector can be explained by the fact that there is a lot more experience with PbR in the health sector. All five systematic

[8] This study covers both developing countries and low-income contexts in developed countries.

reviews are on the health sector. Most PbR schemes in health are in the form of RBF (Table 5). The programmes in education are mostly targeted at national level, but usually involve some incentives at lower government levels or facilities as well. The studies discussing PbR in multiple sectors usually include RBF, RBA and hybrid schemes. The studies discussing hybrid schemes are about the above mentioned mechanism of Output-Based Aid.

Table 5. Classification of sources by sector and type of RBF/RBA

	RBF	RBA	RBF + RBA	Hybrid	Multiple	Total
Health	20	6	3			29
Education			4	1		5
Water				1		1
Energy	1		1			2
Governance		3				3
Multiple		1	2	3	2	8
Total	21	10	10	5	2	48
In %	44	21	21	10	4	100

Source: The author.

Although I am confident that the remaining 48 studies are sufficiently representative for being able to conclude something on PbR, the search strategy may have some limitations. First, one sector, namely agriculture and forestry (or the environment), is excluded. Second, we only searched two academic databases, which means, for example, that some studies on the health sector may have been missed as they are more likely to appear in medical science databases. Third, while checking google search, we sometimes limited the search to words in the title only in order to keep it manageable. There is a small chance that we missed relevant studies for that reason. Fourth, we may also have missed grey publications. And finally, even though most inclusion and exclusion decisions were taken by two persons (author and research assistant), there is always an element of subjectivity.

5 Payment by results in health

This chapter first discusses reviews and systematic reviews on payment by results in the health sector, and then the studies on individual cases/programmes in health. Most of the 22 case studies are on RBF, but six of them are on one particular RBA programme, the Salud Mesoamérica Initiative. For this reason, these six are discussed separately in section 5.3. Table 7 gives a schematic overview of the results of the reviews, and Table 9 of the case studies in health.

5.1 Reviews and systematic reviews of RBF/RBA

This section discusses the results of three regular reviews and five systematic reviews in the health sector, one of which is a review of cost effectiveness only. Most studies are academic, except for one review and one systematic review.

Effectiveness

Oxman and Fretheim (2009) critically review four PbR schemes for which evaluations were available. One of these is GAVI, an RBA that will be discussed in chapter 7. Evaluations of the other three schemes reported positive or mixed (one scheme) trends in outputs, but the positive trends could not be attributed to RBF due to lack of rigorous evaluations (two programmes), or the positive effects proved to be due to incentives provided to the demand side (one programme).

Grittner (2013) reviews 12 RBF programmes in 13 developing countries[9] for the German Development Institute (now called German Institute for Development and Sustainability). In four schemes, no positive effects on the targeted indicators were found. In two cases, the bonuses were too small or came with too long delays to be effective. In seven programmes there were positive trends in health care supply and health coverage (outputs), but there were no rigorous control groups so it was not clear whether the result could be attributed to the performance-based payment. Other contextual factors and increased funding may also have played a role. This lack of

[9] Afghanistan, Bangladesh, Bolivia, Burundi, Cambodia, Costa Rica, DRC, Haiti, Nicaragua, Rwanda, Senegal + Madagascar (one scheme), and Tanzania.

rigorous evaluation also holds for one out of three programmes reporting positive health outcomes, and for three schemes reporting positive health impact on, for example, reduced child malnutrition. To the extent that service delivery increased and improved it was often more due to non-monetary incentives like increased flexibility and more involvement of staff in management decisions, than to the financial incentives. There was limited evidence on unintended effects. The exceptions include studies on Rwanda and Tanzania, which observed reduced attention for preventive care and quality, and possibly for other necessary care that was not rewarded.

Eichler et al. (2013) focus in particular on what performance-based incentives can do for the health of mothers and newborns. They review nine schemes/sources, out of which eight represent RBF[10] while three of them apply both supply and demand side incentives. Seven out of nine studies report an increase in institutional deliveries, but only three of them use rigorous control groups. While programmes also incentivized improved antenatal care, only three found a positive effect and only one did so with rigorous methods. There is no information on mortality of mothers and babies (impact). Some studies report that quality of services is rewarded, but it is not clear how quality is measured or no results are reported (Eichler et al., 2013).

Ogundeji et al. (2016) write a systematic review of the effectiveness of RBF in health as well as of possible factors that influence effectiveness. The authors review 96 primary studies that cover 68 different programmes. Even though this review also covers studies and programmes from developed countries, it is interesting because it provides a statistical meta-analysis of the quantitative studies.

For 37 studies they computed an average outcome variable, and for all 96 studies they performed a logistic regression using a binary outcome variable indicating whether or not the scheme was effective. They find that 70% of all result variables measured show a positive, albeit very small, effect. The effect was smaller for variables at impact or outcome level[11] than for variables at process or intermediate outcome level. In addition, the size of the effect varied with evaluation design: studies with less rigorous designs were 24 times more likely to find a positive effect than

[10] The ninth scheme represents payment for performance as part of a social insurance reform (Egypt).

[11] Impact variables include morbidity and mortality rates, and outcome variables are defined as those requiring behavioural change.

randomized controlled trials. These two findings imply that the positive effects of PbR are probably overestimated. They also found that the larger the incentive and the lower the risk of not receiving the incentive, the higher the chance of a positive effect. The former finding was more robust than the latter (Ogundeji et al., 2016).

Duvendack (2022) reviews aid-funded experiences with PbR in health in low and middle income countries. She carried out a systematic review, including studies published between 2000 and 2020 and with quantitative designs only. For the 81 studies selected, a quality review was carried out identifying low, medium or high risk of bias. About one-third of the reviewed studies had a high risk of bias, so a weak methodological design. The studies included results for 52 output variables and 39 outcome variables. Positive results dominate negative ones, but like Ogundeji et al. (2016), Duvendack finds that measures at output level register more positive and significant effects than measures at outcome level. On the other hand, she does not find differences in outcomes between studies with high and low methodological risk. She did not find evidence for publication bias either: academic and grey publications report about the same rates of effectiveness. A minority of studies reports significant negative effects of RBF in health. About half of these negative outcomes proved to be from schemes in fragile states such as Afghanistan or the Democratic Republic of Congo (DRC). Although the overall findings are positive, the author warns that results are probably overestimated due to factors like lack of independent verification of performance measures, the dominance of output measures and poor quality of evaluation designs (Duvendack, 2022).

Diaconu et al. (2021) carry out a systematic review of paying for performance in health in low and middle income countries for the Cochrane collaboration. They only include studies with rigorous quantitative methods: randomized or non-randomized trials, controlled before-after studies, and interrupted time series.[12] All studies compare PbR with a counterfactual, be it a situation of standard care or another intervention. The review focuses on low and middle income countries. The search leads to a group of 59 studies, 58 RBF and one RBA. Not all schemes are aid-funded: 23 out of the 59 studies refer to programmes funded by national governments. The authors assess the certainty of evidence, making four categories – very low, low, moderate and high certainty – , on the basis of the methodological

[12] Interrupted time series implies that the variable of interest is measured several times before and after the intervention.

quality of studies (deducing 2 points, for example, for results that came mostly from controlled before-after studies), taking into account also whether measured effects were large, and the number of studies reporting a particular result (Diaconu et al., 2021:15).

Most of the evidence had a low level of certainty. When compared with standard care, PbR schemes may slightly improve health outcomes and service quality, but the effects on service delivery and utilization of services was mixed, with some indicators improving (share of people receiving HIV testing, outreach of family planning), but others deteriorating (decrease in proportion and children and households using bed nets) and in yet other indicators there were mixed effects (use of antenatal care). There were probably (moderate certainty evidence) increases in human resource and medicine availability and in functioning infrastructure when these were targeted. If they were not targeted, there is no good evidence. With regard to non-targeted outcomes, there were probably few or no distorting unintended effects (low certainty).

When RBF interventions were compared with other interventions, there was little to no effect on health outcomes, and the effect on indicators for access and use was mixed. If targeted in the RBF scheme, the effect on quality may have been positive; if not there is only low certainty evidence. If resource use was targeted, the evidence is mixed and if not targeted, the effects on resource use are uncertain. In this smaller group of studies (those that compare with other interventions), none reported on unintended effects.

The authors also looked at subgroups of studies in order to identify factors influencing outcomes. It appears that RBF programmes that target outputs and make an adjustment for quality have the largest chance of success. The authors conclude that payment for performance schemes have mixed effects and that there are large differences in types of schemes and in evaluations conducted. Outcomes of these schemes tend to depend on their design, on the amount of extra funding, on complementary elements such as technical assistance and on context (Diaconu et al., 2021).

Singh et al. (2021) also wrote a systematic review of payment for performance in health. It is a realist review, so they also assess which mechanisms and which contexts support positive findings. This means that they included qualitative studies, while excluding studies that only assess health outputs or outcomes. They discuss 117 studies covering programmes in 36 low and middle income countries.

Thirty-two studies included data on the use of health care services, and 19 of them found a positive and significant effect on at least one indicator; most often institutional deliveries. In the other studies there were either non-significant effects, or the study design did not allow for conclusions. Possible factors that could explain positive effects on health care use include improved availability of medicines and equipment, a reduction in user fees, more efforts of providers, and changed work procedures like more adherence to clinical care guidelines, improved interaction with patients, and more outreach activities in the community. Several of these changes increased patient trust which in turn helped increasing demand for services. Increased autonomy in facilities regarding financial management and other decisions helped fostering these supply side improvements.

Seven studies report negative effects on activities that are not incentivized, and two conclude that this is more likely to occur at lower levels of care where too few staff is available. And seven studies find that payment for performance leads to misreporting and gaming. One study reports that this misreporting frequently happened, and that it was induced by perceptions of inappropriateness of the incentives and perceived lack of time to meet the job requirements. Another study relates misreporting to insufficient verification methods and weak sanctions for this behaviour (Singh et al., 2021).

With respect to the contextual factors influencing the effectiveness of payment for performance, they distinguish between "distal factors", related to the wider health system, and "proximal factors", related to the environment of the facilities themselves (Singh et al., 2021:12). With respect to distal factors, the level of autonomy for health facilities was a key factor (for the availability to increase supplies and equipment and for motivation and productivity of staff), as well as well-functioning banks (for the timeliness of payments) and the initial level of user fees. With respect to proximal factors, PbR was more likely to succeed if initial levels of staffing, staff skills and motivation and quality of infrastructure were better.

Singh et al. (2021:12) observe that the overall outcomes of PBR are "mixed and indeterminant". In addition, they identify some further qualifications to the perceived success of PbR in health. First, an important mechanism for the success of RBF seems to be the greater availability of drugs and supplies and the improved infrastructure. However, increasing the core finance to facilities could have had the same effect. Second, there seems

to be limited attention for the cost effectiveness of the RBF schemes. Six reviewed studies report that it is not clear whether the implementation costs, and in particular costs for verification, outweigh the improvements in indicators. Third, it is not clear whether schemes are sustainable. Six studies express concerns on the sustainability of RBF, in particular if external finance plays a key role. However, this risk pertains to all externally financed programmes.

Transparency and accountability

There are four studies reporting positive effects on internal and external accountability (Singh et al., 2021). Giving incentives to district officers for the performance of facilities in their district led to improved internal accountability and better performance in Tanzania. And giving community-based organisations a role in the verification process improved community participation.

Targeting and equity

Five out of 12 programmes reviewed by Grittner (2013) explicitly targeted poor areas and households. They mostly used geographic targeting and means-tested targeting. In four of them, targeting proved successful in improving access to health care for the poor and/or reducing the poor's health spending. For the fifth there were no data. In two of the other schemes (without explicit targeting) it proved difficult to design incentives for health providers to deliver services to remote and poor parts of the population (Grittner, 2013:24). Diaconu et al. (2021) conclude that the impact of PbR on equity is mixed; like most other evidence in their review, this conclusion has a low level of certainty.

Cost effectiveness

Grittner (2013) observes that most studies do not provide sufficient financial information to assess cost effectiveness. In the four studies that do, the evidence was mixed.

Ogundeji et al. (2016) find that larger financial incentives are accompanied by larger positive effects, but this will increase the cost of the schemes. Very few studies in their review examine whether the positive outcomes could have been achieved more cost effectively in other ways.

As mentioned above, Singh et al. (2021) observe that there is limited attention for the cost effectiveness of RBF schemes. Six studies report that it is not clear whether the implementation costs, and in particular costs for verification, outweigh the improvements in indicators.

The PhD thesis by Salehi (2020) studies RBF in health in Afghanistan and includes a systematic review of studies on cost effectiveness of RBF in health. This review includes seven studies, but only two of them are on low or lower middle income countries (Tanzania and Zambia). These schemes target maternal and child health. Most of the other studies examine hospital care. Five studies conclude that the RBF programme was cost effective as compared to the status quo or to control groups. The study on Zambia was one of them, concluding that more lives were saved[13] in districts with RBF than in districts without. However, there are concerns about the study methods used. The study on Tanzania finds that RBF was not cost-effective, but could become so if scaled up to the national level. That study has institutional deliveries as outcome variable, and not lives saved. Salehi argues that most of the studies have limitations in the sense that they either do not use a cost effectiveness threshold, or apply the one suggested by the WHO, which is not considered adequate. In addition, none of the seven studies compares RBF with other interventions.

5.2 Case studies on RBF in health

Many of these case studies on RBF (12 out of 16) are covered by one or more of the above discussed reviews. Some highlights of these studies are included below, as well as results of other studies. Only one publication is grey. Many studies provide evidence on the same programme/country, and they are discussed together.

Effectiveness

Basinga et al. (2011) carried out a quantitative study on the RBF scheme in Rwanda, which gave bonuses to both facilities and health workers. They find large increases in some rewarded activities (e.g. institutional deliveries) but no increases in other. They also find that positive effects are linked to activities that were most rewarded and/or that required relatively less effort. Kalk et al. (2010) is a qualitative study on Rwanda. They give several examples of distortion and gaming, but the scale on which this happens is

[13] Measured in QALY, Quality-Adjusted Life Years.

not so clear. Their interviews reveal that health workers sometimes experience an ethical conflict when they must choose between necessary care and rewarded care.

Soeters et al. (2011) conclude in a rigorous quantitative study that the RBF scheme in DRC that was similar to the one in Rwanda had significant positive effects on two outputs and outcomes, next to insignificant effects on five others. Huillery and Seban (2017) come to more negative conclusions on the same scheme. In a randomized controlled trial they find that the use of services decreased in the facilities where services were rewarded, and that health outcomes of newborns also decreased. This happened despite increased efforts of health workers to attract patients, including by lowering user fees. Apparently, the lower fee signaled lower quality and did not increase the number of patients. It also led to less income for the facilities. They also find that the fee for service system lowered the intrinsic motivation of health workers and reduced their job satisfaction.

In Tanzania, PbR was introduced in 2011 by the government in one region. Eight studies in this review analysed this programme. It was financially supported by Norway and offered financial incentives to health facilities (both for staff and for the facility itself), district and regional managers if they increased service delivery in maternal and child health above a baseline (Binyaruka & Borghi, 2017). According to an impact evaluation cited in Binyaruka et al. (2018b:3) this RBF had significant effects on two out of eight incentivized indicators: institutional deliveries and the provision of drugs against malaria as part of antenatal care. The RBF was also associated with improvements in the availability of medicines and reductions in stock-outs. Key factors behind this were the incentives to facilities that could be used at discretion, the incentives to district managers (who had an important role in drugs supply) and the increased district supervision as a result of the districts' role in verification (Binyaruka & Borghi, 2017).

Anselmi et al. (2017) investigate the pathways for the increased institutional deliveries and provision of antimalarial drugs in a quantitative way. They find that the reduction in stock-outs is a key factor behind the former, as it reduced the probability that patients had to pay for medicines themselves. Another factor was increased kindness of staff (Anselmi et al., 2017). Cassidy et al. (2021) aim to understand the pathways through which the two positive effects in the Tanzanian RBF came about in a qualitative way. They come to similar conclusions. The incentive payments helped facilities to purchase drugs and other supplies. With fewer stockouts, patients perceived health services to be of higher quality.

Chimhutu et al. (2019) investigate the effects of the RBF in Tanzania by conducting interviews with health workers, patients and members of Health Facility Governance Committees. They find that performance in several aspects of maternal and child care improved due to the incentives, but that this cannot be generalized to all health care services. They find that attitudes of staff, service delivery, and teamwork improved, but only for maternal and childcare services, but that this was certainly not the case for other health services. The authors also observe that there are persistent barriers to access to care and to quality, most notably poor infrastructure of facilities, lack of medicines and other supplies, and lack of adequate staff. Lack of money and long distances are hampering the use of healthcare. The authors conclude that these persistent barriers plus the unintended negative consequences for other care outweigh the benefits of the improvements in maternal and childcare (Chimutu et al., 2019:11). Borghi et al. (2021) examine the sustainability of the (limited) positive effects of the scheme, and find that these positive effects reduced over time, and that drug availability also decreased. On the other hand, reporting became more integrated with routines, and financial autonomy and supervision were maintained with positive effects on all health services provided.

Bergman et al. (2021) carried out a qualitative study of the first 18 months of an RBF programme in Zambia. Sida supported the government of Zambia in financing the Reproductive, Maternal, Newborn, Child, Adolescent Health and Nutrition (RMNCAH) programme in 22 districts. Districts received a fixed and a variable tranche, and the latter was based on performance according to five indicators. The study does not report results yet, but finds that most district officers and health facility staff did not know about the results tranches nor did they know the indicators. There was also uncertainty about how the performance on the indicators would be translated into disbursements. All in all, it was unlikely that the system would provide incentives for better performance (Bergman et al., 2021).

Petrosyan et al. (2017) discuss the successful scale up of an RBF in Armenia aimed at increasing use of Primary Health Care (PHC) services. It was financed by USAID and the national government together, and later also partly by the World Bank. Bonuses were provided to health workers (80%), administrative staff (13%) and facilities (7%), if 27 indicators were met. The programme achieved an increase in use of facilities. Key success factors were the well-embeddedness in national policies, the changing of indicators on a regular basis to bring them in line with national priorities and to avoid perverse effects and gaming, and the free choice for patients, allowing for competition between facilities.

In Afghanistan, the government contracted out a basic package of health services to non-state providers with support from USAID, WB, EU and other donors (Salehi, 2020). This programme, carried out between 2003 and 2010, was a success: it led to much greater health service coverage and to lower infant, child and maternal mortality rates. In 2010, the government introduced a payment by performance system with support from the World Bank. Health facilities and health workers in randomly selected facilities received a bonus for delivering additional maternal and child health services as compared to a base line, while the facilities in the control group did not. The number of services delivered was higher in the treatment facilities, but the difference was not statistically significant. Factors explaining the lack of results include the overall low level of resources in facilities, the lack of facility autonomy, the unbalanced incentive structure, with some bonuses probably too high (for deliveries) and others too low (for antenatal and postnatal care). There was also evidence of health workers and facilities not fully understanding the incentive system (Salehi, 2020).

Alignment

The study on the Zambian RBF programme shows that there was insufficient alignment to the programme of all actors involved (Bergman et al., 2021). And the same conclusion was drawn for RBF in Afghanistan (Salehi, 2020).

Transparency and accountability

Soeters et al. (2011) find that the RBF in the DRC increased transparency and reduced corruption. Anselmi et al. (2017) find that the positive effect of RBF in Tanzania on increased provision of antimalaria drugs during antenatal care is related to the increase in number of supervision visits, induced by two elements of the scheme: bonus payments to district managers and the role of these managers in verification. Cassidy et al. (2021) and Mayumana et al. (2017) confirm that increased district supervision improved internal accountability and played a role in increased availability of medical supplies, which helped to achieve more institutional deliveries and better antenatal care. However, Cassidy et al. (2021) argue that demand for services depends on citizens participating in the Community Health Fund (CHF), a community-based health insurance scheme. Community health workers and members of Health Facility Governance Committees were supposed to persuade citizens to contribute to the CHF, but none of these groups were incentivized in the Tanzanian RBF and this was an important omission.

Targeting and equity

Binyaruka and Borghi (2017) find that the increased availability and reduced stockouts of medicines in Tanzania as a result of the RBF was mostly pro-poor, i.e., better in facility catchment areas with lower income households, and also greater in rural than in urban areas. Binyaruka et al. (2018b) examine the distributional effects of the Tanzanian RBF at the individual level. They find that the scheme has a greater effect on institutional deliveries among women of middle wealth status, uninsured women, and women living in rural areas, than on wealthier and insured women and women living in urban areas. This implies a positive effect on equity. With respect to the other targeted result with a positive effect, the improved uptake of antimalarial drugs during antenatal care, there was no pro-poor effect. In another study, Binyaruka et al. (2018a) investigate the effects of the scheme on inequality between health providers. It turns out that the better-off health facilities received much more incentive payments. This inequality reduced somewhat over time, but remained high.

Cost effectiveness

Two studies explicitly assess cost effectiveness. Mariviglia (2011) examines the efficiency and cost effectiveness of a performance driven loan (PDL) of the IADB for support to the Expanded Immunization Programme in Colombia. After a first tranche in the first year (20%), the next four tranches would be given depending on performance on five indicators. Three were related to vaccination coverage, also in prioritized municipalities, and two to the government's operational capacity for immunization. The study compares the cost effectiveness of this PBL with a CCT approach and with a traditional investment loan. This cost benefit analysis reveals that the PBL was most cost effective and that CCTs are also more cost effective than a traditional loan. The analysis uses aggregate data so it is not clear to what extent all targeted municipalities have benefitted from this PDL. In addition, the author raises some doubt on the generalizability of this result to all low and middle income countries. As a higher middle income country, the systems for monitoring outcomes in Colombia were already quite good, and selective audits were sufficient to ensure reliable information (Maraviglia, 2011).

The costs of the RBF programme in Afghanistan were high, and consisted of incentive payments (70%), verification costs (23%), staff time (6%) and administration (2%) (Salehi, 2020). The programme was not cost-effective,

with much higher unit costs of services in the treatment facilities than in the control facilities. The cost of PbR per Disability Adjusted Life Years (DALY) averted was US$1.241, which was far above the estimated cost effectiveness threshold for Afghanistan of US$349 (Salehi 2020:213).

5.3 RBA in health: the Salud Mesoamerica Initiative

Six papers, five academic and one grey, examine the effects of the Salud Mesoamerica Initiative (SMI), which is results-based aid (RBA) in the health sector in seven Central American countries[14] plus the region Chiapas in Mexico. The Initiative is administered by the IADB and financed by the Bill and Melinda Gates Foundation, the Carlos Slim foundation, and the Spanish Agency for International Development Cooperation. It aims to improve health services and outcomes for mothers and children in the poorest and most underserved municipalities (El Bcheraoui et al., 2017). The SMI and national governments (and the regional government in Chiapas) each provide 50% of the funding for the project, and if the pre-agreed targets are met, the SMI rewards this with an additional 25% financing to be spent in the health sector at the discretion of national health authorities.[15] Most studies qualify this scheme as a success. Five of the eight countries participating in SMI reached the target indicators for the first operation (between mid-2012 and early 2014), and the other three countries also made significant progress in these indicators (El Bcheraoui et al., 2017).

Effectiveness

Three studies examine the effects quantitatively; one of them (Bernal et al., 2018), on El Salvador, is covered in the review by Singh et al. (2021). Bernal et al. (2018) compare outcomes in maternal and child care services, both targeted by the SMI, in three groups of municipalities: those with RBA, those with conventional input-based aid, and those with national funding for health facilities. The results are spectacular: preventive health services for maternal and child care increased by 42% in RBA municipalities and by 20% in conventional aid municipalities in comparison with national funded municipalities. Other health services in the RBA municipalities also

[14] Belize, Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua, and Panama.
[15] iniciativa Salud Mesoamérica | Iniciativa salud mesoamericana, accessed 19 September 2022.

increased. The authors find that the positive effects are not due to increased productivity of health workers, but to increases in staff and investment in additional facilities in RBA targeted municipalities. Apparently, national health authorities were motivated to perform well due to the large reward they could obtain.

Hernández et al. (2022) examine the effects of SMI on available resources for delivery care, on institutional deliveries, and on the proportion of women giving birth in a facility that was not closest to their residence.[16] They compare treatment and control areas in Guatemala, Honduras and Nicaragua with a difference-in-difference approach. They find no statistically significant difference in available resources, but institutional deliveries significantly increased and the proportion of women not using the closest facility for delivery decreased.

Duber et al. (2018) investigate another area targeted by SMI, namely the appropriate and timely provision of antibiotics for neonatal sepsis. They focus on facilities and areas participating in SMI and compare the situation before and after the first 18 months of implementation the SMI. They find that the availability of antibiotics increased, but that the percentage of newborns with sepsis that received appropriate antibiotics decreased. This points to an early success in increasing supplies, but it means that more is needed in terms of improving diagnosis and management of neonatal sepsis in the targeted municipalities.

One of the qualitative papers on SMI examines its sustainability (El Bcheraoui et al., 2018). The authors conclude that most elements of a "Dynamic Sustainability Framework" are present, in particular the evolving design that could move with national needs, the regional approach that led to prioritizing health care, the culture of learning from evidence, knowledge sharing, system improvements and evidence of scaling up. There are also some threats to sustainability: in the absence of finance, regional partnerships may weaken and the priority for health care may drop. In addition, turnover of personnel is a risk for continuity.

[16] The rationale for this indicator is that women prefer a facility close-by if quality of service is considered sufficient. So, the more women use a close-by facility, the better.

Alignment, transparency and accountability, and innovation

Two studies examine the reasons for the perceived success of SMI. They show that aligning all actors, stimulating a culture of accountability, and learning (innovation) were important factors. According to El Bcheraoui et al. (2017) factors driving the success proved to be 1) a flexible design that valued the opinion of participating countries and had them participate in setting appropriate indicators; indicators also changed over time 2) the regional approach that led to competition between participating countries to achieve the targets, 3) the inclusion of planned external evaluations after each operation which led to a culture of accountability, and 4) the fostering of an environment that stimulates experience-based learning.

Eichler et al. (2017) examine in particular the role of external monitoring as driver for success of the SMI. From the start it was agreed that there would be independent verification of the targeted indicators. Donors wanted to ensure that the bonus payments to countries were based on true numbers, and the IADB wanted to reduce fiduciary risks in the administration of the funds. In addition, external measurement would ensure that data would be comparable across countries. Given that little was known on the health situation in the poorest municipalities, donors financed the data collection, including baseline and follow-up surveys. The IADB provided technical assistance for improving the country's own data systems ("dashboards") and countries were very motivated to do so in order to be able to monitor progress themselves. All this led to more evidence-based policies on health for the poor.

Targeting and equity

The SMI aims to improve the reproductive health and the health of children of the poorest 20% of the population. The scheme targeted municipalities, so it applied geographic targeting. In El Salvador, municipalities with the highest rates of extreme poverty and of malnutrition were selected, based on household surveys, and this selection was finetuned by using the most recent census data (Bernal et al., 2018). Given the success of the SMI, it can be said that it improved health equity in the region.

6 Payment by results in education

Compared to the health sector, much fewer rigorous studies are available on RBF or RBA in education. The rigorous studies that exist and that are included in some of the reviews discussed below are often on developed countries. RBA programmes in education in developing countries are often nation-wide, so it is difficult to include a good counterfactual. On the other hand, and especially with regard to RBA, more evidence is available on more qualitative aspects of PbR, like distortion and gaming, alignment, transparency and accountability, and innovation. See Table 9 for an overview of the results of reviews in education and in other sectors, and Table 10 for the case studies in education and other sectors.

6.1 Literature reviews on RBF and RBA in education

This section discusses three reviews of PbR in education. None of them are systematic reviews. All are grey publications (of the World Bank), and one discusses experiences from all countries of the world.

Effectiveness

Hill et al. (2015) analyse 24 OBA projects in education in a qualitative way to assess factors that facilitate success. However, the report does not define "success". As a result, it is not clear how many of these projects were successful, nor what the meaning of the findings is, like "community engagement … is important for project success" (Hill et al., 2015:5). The OBA projects include incentives for facilities and sometimes also for teachers.

Lee and Medina (2019) carry out a literature review that includes 41 quantitative studies with experimental or quasi-experimental designs. They also cover 8 earlier reviews. Most of these studies and reviews are on developed countries. In order to examine pathways to success, they also use project documents (of RBF and RBA in developing countries) and reports on recently started REACH projects, sent out a survey among 46 staff of relevant donor institutions and conducted follow-up interviews with 15 of them.

Lee and Medina (2019) separately analyse schemes with incentives to teachers, to schools, and to governments, but there proved to be too little evidence on the latter, so on RBA. The evidence on the effect of incentives for teachers on learning outcomes is mixed, but the effects seem to be a bit more positive and larger in developing countries. Incentivizing teacher attendance may have a positive effect if this is well monitored, but there are mixed findings for the impact of improved attendance on learning outcomes. There is some evidence of gaming ("teaching to the test") and cheating in measuring the learning outcomes (Lee & Medina, 2019:21).

They find limited evidence of performance-based grants to schools. One problem is that this measure is often accompanied by other interventions, such as abolishment of school fees, school management training or the setting up of school committees. The reviewed studies show a large variation in outcomes. Meta-analyses of the effects of these schemes do not reveal statistically significant outcomes on completion or dropout rates, or on learning outcomes (Lee & Medina, 2019:34). There is some evidence that when performance-based school grants increase, households reduce their own education expenditure which reduces the effect of the grants.

The authors conclude that there is no rigorous evidence for concluding that RBF is better for achieving learning outcomes than other financing modalities, but that RBF, if adequately designed, is able to stimulate stakeholders for achieving positive effects in education. In developing countries, it is often necessary to target several obstacles simultaneously, and this appears to be most promising. Examples include: incentives to teachers improved learning outcomes if sufficient textbooks were available (Uganda), incentives needed to be given to both teachers and schools (Tanzania), or to school managers, teachers *and* households (Mexico) (Lee & Medina, 2019:38).

Terway et al. (2021) identify and discuss 51 RBF projects that incentivize the meso-level (districts, school managers or NGOs) in education in developing countries. Most projects are financed by the World Bank, but principals often include national government agencies (48%) or sub-national governments (24%). Agents include schools (29%), service providers (20%) and tertiary education institutions (20%). Most indicators were output or process indicators; (learning) outcome indicators represented only 17% in these projects. Evaluations of these projects report very diverse results. In addition, when studies conclude on statistically positive effects on learning outcomes or changes in management behaviour,

it is often not clear whether they are due to RBF, to other components of the programmes, or to other factors. Other studies find negative effects (Terway et al., 2021:23).

Targeting and equity

All 24 OBA projects in education examined by Hill et al. (2015) proved to target poor students, and sometimes also other specific groups like girls, orphans and vulnerable children, disabled children or adults without education. Several targeting mechanisms were used but it is not clear which one was most effective. The schemes sometimes used weighted subsidies for different target groups, in order to avoid "cherry picking".

6.2 Case studies on RBF and RBA in education

This section discusses the two remaining studies on education as listed in Table 4. The section also incorporates a discussion on RBF and RBA programmes as analysed in some of the reviews on multiple sectors (Table 4). These programmes include the Girls Education Challenge (GEC) and two DfID financed pilots in Rwanda and Ethiopia.

Effectiveness

Several studies discuss the GEC, a form of RBF. In 15 out of 37 GEC projects, in 18 countries, some of the money that was transferred to implementing NGOs was based on performance in the form of learning outcomes (Clist, 2019). Given that there was no random assignment of treatment and control groups, no rigorous evaluations are possible. Available evidence so far shows that the financial incentive of 10% stimulated the participating NGOs to focus more on learning outcomes. However, there was also a tendency among these NGOs to give more priority to the short term than to the long term, and to take fewer risks (Holden and Patch, 2017, as cited in Clist, 2019; and Bond, 2017 as cited in Lee & Medina, 2019).

Both DfID-financed pilot programmes in Rwanda and Ethiopia were contracts between DfID and the governments, but the Ethiopian government passed on the incentives to certain regions. In both countries, a first indicator was the number of students sitting key exams. In Rwanda

there were additional payments for improvements in the teacher's level of English, and in Ethiopia the exam pass rates were a second indicator (Clist, 2019:724).

The main conclusions of independent evaluations of the schemes were similar: although both countries registered increases in number of students sitting exams and other positive trends (e.g., in completion rates in Rwanda), these cannot be attributed to the RBAs (respectively Upper Quartile, 2015 and Cambridge Education, 2015 cited in Clist, 2019:724). There were several problems with the chosen indicators. The tests for monitoring the level of English in Rwanda proved to be different from year to year, not sat by the same teachers and taken in different contexts. This means that the measured improvements were not based on actual improvements. In addition, number of students sitting exams and even completion rates do not tell much about students' knowledge. In Ethiopia, pass rates were not a good indicator because in principle they remained the same every year (they are "graded on a curve"). In practice they increased during the programme, but this was due to a change in the norms defining pass rates (Clist, 2019:726). The evaluation of the Ethiopian RBA found that the effects on performance were hindered by the fact that many regional officers and school heads did not know about the programme (Cambridge Education 2015, cited in Lee and Medina, 2019:46). Yet, in some regions and schools the evaluation found that there were improvements in strategic thinking and in prioritization (Lee and Medina, 2019:55).

Dom et al. (2021) carried out an evaluation of three World Bank PfR programmes in basic education: in Mozambique, Nepal and Tanzania. Moran et al. (2020) evaluated the same programme in Tanzania, the Education Program for Results (EP4R). It was initially funded by the World Bank, DfID and Sida (2014–2017), and in a second phase (from 2017 onward) also by the Global Partnership for Education (GPE) and the Korean International Cooperation Agency (KOICA). This RBA included several input and process indicators, as well as output indicators like a more equitable teacher pupil ratio, transition rates for girls and survival rates for girls and boys. Only nine percent of potential disbursements was based on learning outcomes. Although the overall disbursement rate was quite good, some important targets were missed. An important context factor was that the government abolished school fees in 2015, leading to large enrolment increases. This made achievement of some of the targets challenging (Dom et al., 2021).

In Nepal, the PfR was related to the School Sector Development Plan (SSDP) adopted in 2016. It included 84 individual DLIs, and most of them are process and output indicators, with learning outcomes only representing 0.1 percent of the total funding. Six out of nine donors supporting the SSDP based their disbursements on the DLIs. Many DLIs were met late or not at all, in part due to the simultaneous decentralization process which replaced 75 districts with 753 local governments as primary responsible agents for education. In addition, the number and complexity of DLIs seemed too high.

In Mozambique, RBF for education started with the World Bank's Public Financial Management for Results program, 2014–2018, which had four process DLIs for the education sector: budgeting, timeliness of school grants, number of schools covered by district supervision, and number of functioning school councils. Many other donors supported the pooled fund for Education, and two of them, GPE and KfW, based their funding on, respectively, four output indicators and one input indicator (school construction). The WB DLIs were achieved by the end of the programme, also due to extensive technical assistance, but it is not clear whether the broader goal of improving sectoral accountability was achieved. One-third of the KfW construction indicators was missed and the programme did not contribute to solving underlying bottlenecks in school construction. Most of the process indicators used by GPE were met.

There is some evidence for all three countries that the incentives worked. In Tanzania and Mozambique, some of the incentives were passed to lower government levels and in Tanzania also to schools. Since these funds represent a larger part of their budgets they have a greater incentive effect (Moran et al., 2020). The local incentives seemed to have a positive effect on, for example, survival and transition rates (Dom et al., 2021). In Mozambique, there was a statistically significant effect of the scheme on a more equal distribution of PTRs in lower primary schools (Dom et al., 2021). In Tanzania, local governments were also able to achieve a more equitable teacher deployment, but after the school fee abolishment, there were not enough teachers.

With respect to several indicators in all three countries, successes were only apparent and underlying goals were not achieved. In Mozambique, the Ministry of Finance did not always pass on the funding to the education sector, thus reducing the incentive effect. For other DLIs, the incentive was probably not needed as they were already a government priority (Dom et al., 2021). Furthermore, in all three countries there was

some evidence that objectives that were not targeted in DLIs received less focus, both in reporting and in efforts to achieve them (Dom et al., 2021). All in all, effectiveness of the three schemes can be considered limited.

Alignment

Moran et al. (2020) are positive on the contribution of EP4R in Tanzania to alignment with government priorities and government systems and to donor coordination and improving the sector dialogue. But they also note that donors had different DLIs and sometimes different time frames. There is also some evidence of improved coordination within governments, for example between central and local levels in Tanzania, and between different ministries in Mozambique and Tanzania (Dom et al., 2021; Moran et al., 2020). On the other hand, the voices of civil society and NGOs were sometimes less heard in policy dialogues in the three countries due to a narrow focus on DLIs (Dom et al., 2021).

Transparency and accountability

In the two RBA schemes in Rwanda and Ethiopia, standard administrative data were used, which was a cheap measure of verification. However, Clist (2019:727) reports that this did not lead to better data management systems; to the contrary, the evaluations suggest that the quality of data deteriorated due to the high stakes involved.

In Tanzania, Nepal and Mozambique, there is evidence as well that data collection agencies did not always maintain transparency and rigor in measurement, in order to avoid signaling bad performance (Dom et al., 2021). On the other hand, the use of domestic agencies for data measurement and verification strengthened capacities in Mozambique, and the inclusion of learning outcomes among the DLIs in Nepal and Tanzania led to better and more regular data collection on learning outcomes (Dom et al., 2021). Moran et al. (2020) speak about a contribution of the Tanzanian EP4R to establishing "world class data management systems" on learning outcomes, survival and retention rates, in which non-incentivized indicators can also be monitored.

Innovation

The evaluations of the RBAs in Rwanda and Ethiopia find that these RBAs did not foster much innovation. This was probably also due to the short time horizon of the programmes (Clist 2019; Lee and Medina, 2019:68).

Comparing the schemes of Tanzania, Nepal and Mozambique, the Tanzanian scheme was most focused on outcomes and gave most autonomy to the government. However, the evaluations conclude that this autonomy did not stimulate policy learning. This is considered due to the fact that there was no explicit theory of change (Moran et al., 2020), and to the fact that there was a "missing middle": no intermediate outcomes or outputs were incentivized (Dom et al., 2019). The Nepalese scheme with its large number of DLIs was at the other end of the spectrum, and severely limited autonomy. Donors micro-managed reforms by including DLIs for all different steps in the result chain, thus preventing domestic learning (Dom et al., 2021). In all countries, autonomy at local level was insufficient to achieve the targets, for example the fact that local governments were not able to hire and fire teachers made it difficult to achieve more equitable pupil teacher ratios in Tanzania and Nepal. In general, the schemes did not foster a culture of learning. For example, data were only collected on the DLIs and not for possibly answering questions on why certain DLIs were not achieved (Dom et al., 2021). In addition, the focus on DLIs and the accompanying disbursements sometimes reduced the policy dialogue on how to solve bottlenecks or led to less attention for the policy dialogue on other indicators. In addition, when there were many DLIs, reporting on them was a heavy burden for the planning system and this reduced possibilities for learning and adaptation.

Targeting and equity

In both Tanzania and Mozambique, the programmes contributed to a more equitable pupil teacher ratio. On the other hand, in all three countries some DLIs may have induced "cherry picking" and may have led to less equity in resource allocation (Dom et al., 2021).

Principal

In the GEC, the threat of the principal withholding the performance bonus was real, and this implied a large risk for the participating service providers (Clist, 2019; Lee & Medina, 2019).

In the Rwanda RBA, some funds were withheld but the evaluation reports that it is not clear how the unspent funds were used (Upper Quartile 2015:44 cited in Clist, 2019:729). As mentioned, the relatively short time frame of these pilot projects limited innovation.

Although non-payment occurred in all three countries evaluated by Dom et al. (2021), this often led to roll-overs, restructuring of programmes, changes in design, and more flexible interpretations of the targets. Apparently, donors wanted to maintain the flow of money. This reduced the credibility of the incentives, and the flexibility itself implied high transaction costs (Dom et al., 2021).

Lee and Medina (2019) carried out a survey among 46 respondents working on education, in five agencies that were all engaged in PbR (ADB, Sida, DfID, GPE and World Bank). Sixty percent of respondents indicated that they tended to change the indicators when they were not met, because it is politically difficult to not disburse the funds. Fifty-four percent of respondents had experience with reducing or ending the funding, but funds were then often re-allocated to a part of the same project that was not results-based.

7 Hybrid schemes and PbR in other sectors

This chapter discusses the experiences of MCC, GAVI, GFATM, OBA, RBA/RBF in energy, the PfR of the World Bank, RBA for decentralization, and the variable tranche in budget support. As most of these schemes are country-wide (RBA or hybrid), rigorous evaluations are scarce. The eight reviews that cover multiple sectors (Table 8) usually discuss one or more of these programmes, next to programmes in health or education. The evidence presented here is often based on these reviews. In addition, some single case studies on governance and energy are used.

7.1 Effectiveness

MCC was announced by President Bush in 2002 and started to be implemented in 2004. It implied roughly a doubling of aid from the US. Countries had to comply with strict eligibility criteria, in three areas, "ruling justly" (6 indicators), "encouraging economic freedom" (6 indicators), and "investing in people" (5 indicators). The performance of countries is assessed relative to other countries. Eligible countries can get access to a large amount of resources; they can propose projects and MCC decides (Pearson et al., 2010).

Öhler et al. (2012) examine quantitatively the effect of the eligibility criteria on performance, especially on the indicator "control of corruption". This is the only one that has an absolute threshold: a country must score above the median in this indicator in the World Governance Indicators. They find that countries that had a chance to access the MCC given their performance on other indicators, improved their score in fighting corruption in the first four years after the announcement of the initiative. But after 2006, performance did not increase anymore. The authors suggest that uncertainty about whether aid would be received reduced the incentive for reforms over time.

GAVI's Immunization Services Support (ISS) was the first GAVI programme, and aimed to increase the number of children vaccinated against DTP (Diphtheria, tetanus and polio) in recipient countries. It is an RBA scheme, so a contract between GAVI and national governments. In the first two years countries received a fixed amount based on an estimate of what is needed to expand coverage with a certain number of children above the baseline. From the third year onward, countries received $20 for

each child vaccinated above the baseline. The figures on the number of immunized children must be audited (Eichler, & Glassman, 2008). Early evaluations report that ISS funding correlated with higher DPT coverage, but only in countries with baselines of between 65 and 80% (CEPA 2010, cited in Helland & Maestad, 2015:9). In addition, there was no evidence of reduced immunization with other vaccines (Lim et al., 2008, cited in Helland & Maestad, 2015).

However, a more recent evaluation concludes that although vaccination rates increased in recipient countries, this cannot be attributed to GAVI (Dykstra et al., 2015, cited in Clist, 2019:726). There is a lot of evidence that GAVI's auditing system did not work well (Lim et al., 2008, cited in Pearson et al., 2010). Another study also found that the increases were often a result of over-reporting of vaccination rates, when these numbers were compared with those of the Demographic and Health Surveys (Sandefur & Glassman, 2015).

The Global Fund to fight Aids, Malaria and Tuberculosis (GFATM) is a hybrid PbR mechanism. It has contracts with national governments or with NGOs. The performance criteria are country-specific. Payments are based on self-reported progress, on verification of results and approved expenditures by the Local Fund Agent, and on "contextual information and mitigating circumstances that may affect performance" (Eichler and Glassman, 2008:10). The decision rules are therefore flexible, and this makes rigorous evaluation difficult. The Fund uses a combination of output, outcome and impact indicators. An evaluation found that the focus on quantitative indicators had led to less attention for quality in more than half of the recipient countries (Pearson et al., 2010:36).

The experiences with OBA have been documented and analysed extensively in order to promote this aid modality (Brook & Smith, 2001; GPOBA, 2016; Mumssen et al., 2010). But as Pearson et al. (2010:35) also observe, most evaluations or studies on OBA assess results against the pre-defined targets and there are hardly comparisons with alternative modes of financing. Rigorous impact evaluations are scarce (Mumssen et al., 2010:107). But there are some positive indications of relative effectiveness. For example, Mumssen et al. (2010) compared the assessments in (World Bank) Implementation Completion Reports (ICR) of 37 OBA projects in water, energy and health closed in 2007, with the ICRs of 13 traditional (input-based financing) projects in the same three sectors. It turns out that 85% of the OBA projects achieved or overachieved their objectives, against about 57% of traditional projects.

With respect to OBA projects in the water sector, a review of 12 projects concludes that eight of them achieved 95% of their predefined targets in terms of household connections, and five of them exceeded these targets by at least 15%. If the targets were not achieved, this was often due to an inadequate regulatory environment (GPOBA, 2016). But from this it is not possible to assess effectiveness of this aid modality in comparison with traditional input financing. The chapters on other sectors in these books on OBA, for example on energy, telecommunications, and roads (Brook & Smith, Mumssen et al., 2010) are similarly descriptive on the experiences in those sectors and list some challenges or lessons learnt, but they do not provide an assessment of this aid modality.

Helland & Mæstad (2015) analyse some Norway-funded RBA and RBF schemes in the energy sector. In Uganda, subsidies are provided to private companies for setting up renewable energy plants. But this can hardly be called results based payments as donors give the subsidies during the first five years and not when, after 20 years, energy is delivered to customers. In general, the authors conclude that there is limited evidence yet on the effectiveness of this modality.

Another example in the energy sector is the Beyond the Grid Fund Zambia (BGFZ) funded by Sida and started in 2017. It aims to make off-grid energy connections available to households in rural and peri-urban areas.[17] It works closely together with the Zambian government and provides finance to four contracted companies on the basis of connections made. As of September 2019, 145,000 connections were made and the programme was on track for connecting 306,000 households, institutions and businesses (1.6 million people) before the planned date in 2021.[18] Based on this success, Sweden initiated the Beyond the Grid Fund for Africa (BGFA) in 2019, with the aim to expand the programme to other countries.[19]

60_Decibels (2021) carried out a (mostly) telephone survey among 626 customers of the four contracted companies in Zambia. The report does not provide any information on how the programme works, for example on how companies were selected, how areas or customers were targeted, or on the kind of PbR contract (the size, the how and the when of the payment by connection).[20] This means the report is not an

[17] Beyond the Grid Fund for Africa | REEEP, accessed 23 November 2022.
[18] Beyond the Grid Fund for Zambia | UNFCCC, accessed 23 November 2022.
[19] Beyond the Grid Fund for Africa | REEEP, accessed 23 November 2022.
[20] The consulted websites do not provide this information either.

evaluation of the BGFZ, but it does shed light on customer satisfaction with the programme. This appears to be quite good. Eighty-three percent of respondents indicates that the quality of life "very much improved". Yet, many customers experienced a problem with using the connections.[21]

The Clean Cooking Alliance (CCA) and Modern Energy Cooking Services (MECS) describe 12 RBF programmes funded with public money that promote clean cooking. The programmes all finance private companies on the basis of delivered products or services. The report does not evaluate the modality, but does list some challenges and provides recommendations (CCA & MECS, 2022). All in all, there is no evidence yet of the effectiveness of these forms of RBF.

The IEG assessment of the early experiences with the World Bank's Program for Results (PfR) instrument cannot assess effectiveness yet either. However, it does conclude that outputs dominate outcomes among the DLIs. In addition, the DLIs do not always connect well to the longer-term objectives of the projects.

A very specific case is the use of payment by results (RBA) for governance, in particular in decentralisation. In Ghana, four donors (Canada, Denmark, France, and KfW) plus the Government of Ghana established the District Development Facility (DDF) in 2008. Next to a capacity building grant (12%), there are basic grants for districts that satisfy minimum conditions (20%), and performance grants for districts meeting more ambitious performance conditions (68%) (Janus, 2014). Meeting minimum and performance conditions only influences the allocation across districts, not the donor disbursements to the government. The money can be freely spent by the districts. Criteria for disbursement include a large number of input and compliance indicators, like publishing annual statements of accounts and holding meetings. Although the ultimate goal is to improve and increase services, donors motivated the choice for process indicators by arguing that district authorities should be able to influence them. In addition, they pointed to the complexity and longer time horizons involved in measuring quality of services (Janus, 2014:21).

Independent consultants visit all districts to verify the indicators, but donors were aware that it would be difficult to discover gaming (Janus, 2014:26). Over the years, districts achieved striking improvements in performance on

[21] 36% of respondents have experienced challenges in using the product or service, and 63% of these challenges had not been resolved yet at the time of the survey.

the incentivized indicators: the percentage of districts satisfying the minimum conditions increased from 36% in 2008 to 96% in 2013, and the average performance score increased from 45% to 82% in the same period (Janus, 2014:28).

Sabbi and Stroh (2020) analyse the effects of the DDF several years later. First, they compare scores on the DDF indicators with scores in the District League Table (DLT).[22] This is an index measuring quality and quantity of services available in the districts in five sectors: water and sanitation, health, education, security and governance. There proves to be a correlation of only 0.11 across 216 districts. They examine three districts in more detail. They all have high scores in the DDF but two of them score very low in the DLT. Second, and looking for explanations by conducting interviews, they show that local officers passively accepted the external (donor) definitions of capacity building as represented in the indicators. In one of the three examined districts, interviews revealed that nobody was interested in meetings, but they were held nevertheless and minutes were written in order to comply with the indicator. In another, minutes were written while no meeting was held at all. The authors conclude that both donors and recipients had an interest in "playing the numbers game" (Sabbi & Stroh:2). Both donors and local governments gained legitimacy if they were able to show improvements. On both sides, this meant a reduction in genuine commitment to capacity building in the districts. Public officials stated that "capacity problems will always be with us" (Sabbi & Stroh:9). Although this account might seem disturbing for the value of payment by results, it can be argued that the record for other aid modalities that focus on capacity building is not much better (Andrews et al., 2012).

Some of the literature also deals with (sector) budget support, and in particular with the variable tranches of budget support as applied, for example, by the European Union (EU). EU Budget support disbursements include a fixed tranche that can be disbursed if more general, (sector) policy and governance criteria are met, and a variable tranche. For the variable tranche, countries must achieve some specific targets. The disbursement of the variable tranche depends on the extent of meeting each of the pre-defined – weighted – targets. The EU uses data systems of the recipient for measuring the indicators, and often provides technical assistance for improving data collection and data systems. There

[22] This DLT is an initiative of the National Development Planning Commission in Ghana and UNICEF and is published since 2014, so could not be used at the start of DDF. 2020 District League Table | UNICEF Ghana, accessed 27 September 2022.

is no independent verification. Eichler and Glassman (2008) already report (for the health sector) that these official data are not always accurate. In addition, countries do not always meet the targets. They explain this from the fact that the financial incentive provided by the variable tranche is relatively small, and even smaller for individual indicators. Another problem is that the funds are disbursed to the Ministry of Finance, while the sector ministry is responsible for meeting the targets.

Some more recent evaluations and studies of budget support also assess the variable tranches. Disbursement on variable tranches (by EU and other donors) usually does not exceed 20% of the total budget support envelope (Lawson et al., 2014). It leads to delays in disbursements and decreases in the predictability of aid; this led to increased domestic borrowing in several recipient countries. Furthermore, the indicators for the variable tranches may dominate the policy dialogue, resulting in less attention for other important objectives or for more strategic issues (Dijkstra, 2018). Finally, there seems to be no evidence that compliance with the variable tranches is better than for the fixed tranche (ITAD, 2014; Lawson et al., 2016; Ronsholt, 2014).

In Rwanda, even though most indicators and targets for the variable tranches were based on national plans, the government often did not allocate sufficient resources to measuring the indicators and to achieving them, resulting in delays and in non-disbursements. Also in this case, the money flowed to the Ministry of Finance while line ministries were responsible for meeting the targets (Dijkstra et al., 2020).

7.2 Evidence on other effects

Alignment

Forms of RBA and RBF that pay private companies after they have delivered products or services to underserved and usually poor customers (OBA, and programmes in energy) appear to succeed in aligning the interests of the principal with those of the agent, the private company. However, rigorous evaluations of these programmes are not available.

IEG (2016) found that the objective of fostering harmonization and alignment with other donors did not materialize in the first 27 experiences with "Program for Results" (PfR) of the World Bank. Financing from other donors for the same disbursement-linked indicators (DLIs) was rare.

Janus (2014:24) observes that although four donors worked together in the DDF in Ghana, this RBA is less aligned and harmonised than sector budget support, as some donors have added specific disbursement triggers and require separate assessments of the DDF. In addition, the World Bank has a separate scheme, the Urban Development Grant, with additional indicators that must be verified.

In budget support, the application of variable tranches – not only by the EU, but by other donors as well – reduces donor harmonization because different donors often apply different disbursement indicators (Dijkstra, 2018).

Transparency and accountability

GAVI did not lead to improved data systems (Clist, 2019). The monitoring for GFATM contributed to extensive data collection but not always to better and more useful information. In addition, it contributed to fragmented information systems and uncoordinated surveys (Pearson et al., 2010:39).

OBA projects require external verification for the pre-agreed results, but this monitoring is only used for the payment decisions. There is no evidence of broader monitoring of quality of service delivery, outputs and outcomes or of civil society involvement in this monitoring (Mumssen et al., 2010:132). The same conclusion is drawn in GPOBA (2016:38) on the water sector.

Innovation

The way service providers (agents) are selected may influence the extent of innovation and efficiency. Out of the 79 OBA projects in which some selection of providers was involved, 57% used competitive bidding, 29% worked with an incumbent provider, and 20% selected a number of certified providers and had them compete in quality.[23] The 45 projects that used competitive bidding were in transport, telecommunication, water, off-grid energy, and health (in the latter case with NGOs as providers). The providers compete on the lowest subsidy required or on the largest number of beneficiaries that can be served. There is evidence that this competitive bidding increased efficiency. In one project (telecommunications), no

[23] A few projects used more than one selection method.

subsidy was needed at all. On the other hand, competitive bidding requires capacity and takes time. In addition, there is a risk that providers underbid, which may lead to financial problems later. Independently of the selection method, some providers were also able to innovate and increase operational efficiency, but this depended on the regulatory environment, and in particular on procurement rules (Mumssen et al., 2010:120–124).

In the early experiences with the Payment for Results instrument of the World Bank, IEG (2016) observes a tendency to select easy to reach indicators in order to ensure predictability of funding, even including routine actions, rather than providing incentives for improving performance. This means that innovation is probably not stimulated.

Targeting and equity

On the one hand, Low Income Countries Under Stress (LICUS) were found to benefit less from GAVI-ISS funding. On the other, the paid bonus of $20 had a greater incentive effect in poorer countries and in countries with lower initial immunization coverage (Pearson et al., 2010:42–43; Oxman & Fretheim, 2009). For the GFATM there are no equity concerns, in the sense that poorer countries also benefit from the Fund, and that there doesn't seem to be a difference in coverage between better-off and worse-off groups within countries (Pearson et al., 2010:42).

All GPOBA water projects were targeted to poor households, and most used geographic targeting. Sometimes there were unexpected obstacles on the demand side for making the OBA water projects succeed in improving equity. In rural areas, households had cheaper alternatives than connections to piped water, and in peri-urban and informal settlements connections were sometimes impossible due to lacking formal titles to land. In another case (Uganda), the technical availability of water was too limited and this was not sufficiently taken on board in project design (GPOBA, 2016). There is also evidence that OBA has been able to target the poor. Komives et al. (2005) conclude that targeted OBA funding for new connections for the poor in water and electricity has more pro-poor effects than subsidies on tariffs for lower quantities (cited in Mumssen et al., 2010:101–102).

Several RBA and RBF schemes in the energy sector aim to bring electricity or clean cooking methods to those who do not have them yet, so they are targeted to the poor. The extent to which they succeed in this is not fully clear. In one case (Liberia, see Helland & Mæstad, 2015) this was hampered by a high electricity tariff.

Cost effectiveness

There is no evidence that OBA projects involve higher costs than traditional projects, and some little evidence to the contrary. According to the ICRs of 37 World Bank projects in 2007, the OBA projects scored better on staying within planned budgets than regular projects (Mumssen et al., 2010:108–109). In addition, five ICRs of OBA projects published an economic internal rate of return, and they range from 31 to 126%, which is much higher than the average for World Bank projects of 10 to 12% (Mumssen et al., 2010:35). On the other hand, the GPOBA study on the water sector concludes that there is no evidence of OBA being more efficient than traditional aid (GPOBA 2016:32).

In its evaluation of the 27 early experiences with the World Bank's Programmes for Results (in many different sectors), IEG (2016) concludes that the preparation costs for these programmes are the same as for investment loans, but that implementation costs are significantly higher. But what matters is of course whether the benefits of these programmes outweigh the extra costs, and this could not be established yet in this early evaluation.

In the DDF in Ghana, the assessment costs represent about 2–3% of the annual disbursements, which is not high. But there are also large reporting costs for the districts (Janus, 2014:26).

Principal

The fact that the variable tranches in budget support led to delays in disbursements and lower aid predictability means that principals managed to withdraw aid in case of non-performance. On the other hand, the evaluation of EU budget support to Rwanda found that the EU sometimes allowed that the disbursements foregone due to non-compliance were allocated to the sector in the form of project aid. This reduced the incentive effect greatly. Had the targets been met, the funds would have been disbursed to the Ministry of Finance, while the line ministry now benefits from missing them (Dijkstra et al., 2020).

8 Conclusion

This review assessed available literature on payment by results (PbR), and looked, in particular, at RBF, RBA, combined and hybrid schemes. It reviewed the experiences in the sectors health (mostly RBF) and education (often combinations of RBA and RBF), and then examined other and bigger schemes, and other sectors such as energy and governance. Following the theoretical framework developed in chapter 4, I analysed expected benefits of PbR as well as possible risks and costs. I examined effectiveness (at input, output, outcome, and/or impact level, plus quality), unintended effects like gaming, manipulation and distortions, extent of alignment, effects on transparency and accountability, on innovation, on equity, and PbR's cost effectiveness. In addition, I assessed whether the behaviour of the principal fostered expected benefits.

This concluding chapter first comments on the quantity, scope and quality of the reviewed studies, then summarizes the evidence on effectiveness and on other expected benefits and risk and costs, and finally gives an overall conclusion.

8.1 Quantity, scope and quality of included studies

An extensive search in two academic databases, Web of Science and Scopus, in Google Scholar, and in websites of relevant organisations led to 867 results. After applying exclusion and inclusion criteria, 48 studies on PbR remained. A decision was made to exclude DIBs, CCTs, and studies on the environment like REDD + or PES. Most studies proved to be on the health sector. Other sectors covered include education, energy, water and governance.

The largest number of rigorous evaluations of effectiveness also proved to be on the health sector. PbR in health is mostly RBF, in which the principal is a donor, a national or sub-national government, and agents are sub-national governments, facilities, NGOs, or workers. This facilitates high-standard evaluations of effectiveness, like Randomized Controlled Trials (RCTs), where implementing agents can be assigned randomly to treatment and control groups. In sectors like education or governance, PbR often implies RBA, in other words, a contract between a donor (principal) and the government of a recipient country (agent). At this

national level it is difficult to include a rigorous counterfactual for evaluating PbR – unless regional variation in national application of PbR can be exploited, like in (sector) budget support (Elbers et al., 2009). In sectors like telecommunications or energy, the incentive is meant to foster products or services for the poor, and agents are usually from the private sector. This also complicates assessing effectiveness against a counterfactual, since the private sector usually does not receive aid or a subsidy.

Looking at the schematic evidence presented in Tables 7–10 in the appendix, it can be concluded that most studies report on at least one measure of effectiveness, but that there is much less evidence on quality and all other possible benefits, costs and risks (the columns on the right side of "quality"). This means that even with good quality evidence on effectiveness (in a limited sense, so regarding output, outcome or impact), the value added of PbR is not certain. This will be further elaborated upon below.

8.2 Effectiveness

The percentages and assessments in Tables 7–10 seem to point to a slightly positive balance on effectiveness of PbR. However, the quality of the evidence is often weak or very weak (red, purple and green colours), and the dominance of positive assessments among the case studies is largely due to just two programmes in health (in respectively, Tanzania and Central America) that are the subject of 14 single-case studies. Qualifications to some positive assessments in the reviews are also necessary.

Three of the systematic reviews assess the evidence on effectiveness of PbR in the health sector on the basis of quantitative studies. Although the results seem quite positive, with 70% and 85% success rates and the third with an average positive effect, the authors of these reviews assess the overall evidence as, respectively, weak, very weak, and with low certainty. The quantitative studies included in these reviews often suffer from one or more limitations: they are not always rigorous (RCTs), use data that have not been verified independently, they do not examine whether gaming or manipulation of numbers has occurred, they do not assess the effects on outputs and outcomes that have not been incentivized, they do not assess whether "cherry picking" has occurred with negative effects on equity, and they do not assess cost effectiveness. Furthermore, one study

finds more positive effects with less rigorous methods and with indicators at process and output level (Ogundeji et al., 2016), and another does not find this difference in method but notes that *most* studies use indicators at process or output level (Duvendack, 2020). The fourth systematic review on health is a realist review and includes many qualitative studies. This study concludes that the evidence on effectiveness of RbF is "mixed and undetermined" (Singh et al., 2021).

When looking at the sources that review experiences in other or multiple sectors (Table 8), the evidence on effectiveness is either non-existent or weak, and the percentages are much lower than as reported in the systematic reviews. It must be noted that the schemes discussed in some of these studies are to a large extent overlapping.

When taking into account that the case studies in the health sector only cover eight programmes, the results of all 28 case studies are mixed, with positive and negative results balancing out (Tables 9 and 10). In addition, the positive effect reported in (most) sources on the much-studied Tanzanian programme is based on one quantitative evaluation showing that two out of eight output indicators have a positive and significant effect, implying that this does *not* hold for the other six output indicators. The other positive and much-studied case in health is the Salud Mesoamerica Initiative (SMI). This programme seems to have been a success in leading to greater prioritization of health services for the poor, while avoiding risks like distortion and gaming.

Tables 7–10 reveal that, in general, there is little evidence on unintended effects: gaming, manipulation, distortions, negative effects on other areas of service, or uncertainty about whether meeting the targeted indicator contributes to achieving the ultimate objective. Most quantitative studies do not examine this, and there is clearly a need for more qualitative and mixed methods evaluations. The realist systematic review (Singh et al., 2021) lists the highest number of studies in which unintended effects are reported: 12 out of 36. Most qualitative case studies (at least, those examining effectiveness of PbR) report at least one type of unintended effects. On the whole, the incidence of unintended effects is probably underreported. This again cautions against too positive conclusions on PbR's effectiveness.

8.3 Evidence on other effects

Alignment

PbR is supposed to align objectives of the agent with those of the principal. In some cases and sectors, this seems to have worked. PbR schemes that reward agents only if they achieve results for poor or vulnerable populations, seem capable of changing the priorities of these agents. This holds for agents in the private sector for which servicing the poor is not profitable, as demonstrated by projects in energy or in many of the OBA projects. But it also holds for national governments or other agents that may, in theory, be willing to prioritize poor and vulnerable groups, but that are in practice constrained by a lack of knowledge, awareness or capacities, like the Central American governments with regard to health outcomes of the poorest population. On the other hand, in other cases of PbR, incentives proved much less necessary as objectives and interests of principals and agents were already aligned.

A condition for incentives to work is that all agents "are aligned", i.e., are aware of how the PbR scheme works and preferrably have been involved in its design. This condition proved not always met. Another issue is whether PbR leads to alignment of all principals/donors. The early PfR evaluation and also many other schemes (education, governance, the variable tranche in budget support), reveal that this alignment is negatively affected by PbR since donors tend to establish different performance indicators.

Transparency and accountability

The results on improving transparency and accountability, and on improving data systems, are mixed. In some cases, like the SMI, data systems improved, but in other cases like that of GAVI and of education programmes in Rwanda and Ethiopia, they deteriorated. Similarly, transparency and accountability seem to have improved in the Tanzanian health scheme and in the SMI due to specific design features, but there were mixed effects in OBA.

Innovation

There is very little information on whether PbR stimulated innovation in the health sector. The exception are two studies on the SMI that conclude that improved data, regular evaluations and healthy regional competition stimulated evidence-based policymaking. Some of the studies on OBA cautiously conclude that innovation and efficiency increases have occurred, the latter due to competitive bidding among private providers. However, studies on all sectors observe that many targeted indicators in PbR are on input, process, or output level, which means that the programmes cannot allow for much flexibility or innovation. Some studies on education and governance conclude explicitly that innovation did not occur, either because of the dominance of indicators at the lower level of the results chain (closer to the inputs and further away from final results), or due to a too short time frame of donors.

Targeting and equity

The PbR projects and programmes that explicitly target poor or vulnerable groups seem to be largely successful in improving equity. This includes OBA but also other schemes, such as SMI and BGFZ. There is much less evidence on equity effects of schemes that do not explicitly target the poor. But in so far as studies do report equity effects, the results are mostly mixed or negative.

Cost effectiveness

There is very little information on cost effectiveness of PbR. This means that even if studies show that PbR is effective as compared to other ways of financing, we do not know whether this still holds when the costs are taken into account. To the extent there is information on cost effectiveness of RBF and RBA, the results are mixed, with positive and negative outcomes more or less in balance. The OBA projects appear to be relatively efficient, but on their effectiveness there are no good comparisons with other aid modalities.

Principal

There is not much information on the extent to which principals are willing to withhold aid or resources when targets are not met. This appears to be less of a problem in RBF and in multilateral RBA like the GEC, GAVI or GFATM. But in RBA where the principal is a bilateral donor or the EU, and the recipient country is the agent, several cases are reported in which principals have problems in withholding the resources.

8.4 Conclusion

Overviewing the presented evidence, we can conclude that PbR appears to have been effective in some cases and circumstances, but that effectiveness, let alone cost effectiveness (or value added) of PbR in general is by no means certain. The risks and costs of PbR as identified in the theory are real. However, these risks and costs vary by type of PbR and by sector. This means that the effectiveness of PbR not only depends on design and context, as concluded by many other reviews (Diaconu et al., 2021; Duvendack, 2020; Ogundeji et al., 2016), but also on type and sector.

In terms of design, key factors appear to include the involvement of all stakeholders, flexibility in setting indicators from year to year, a perception of fairness of indicators and targets, and independent verification of targeted indicators. The amount of extra funding and the size of the incentive also matter, and accompanying technical assistance often increased effectiveness. However, the list of desired design features is long and it is not certain whether the benefits of PbR outweigh the costs of meeting them.

In terms of context, it can be concluded that PbR is more challenging in fragile contexts, and in general in situations in which not enough staff, resources, and infrastructure are available. But this of course holds for all interventions. Another important context factor, especially for RBF and with agents at sub-national level, is that agents must have some level of autonomy. And finally, a condition for PbR to work is that the principal is willing to withhold the money in case of non-performance. This seems more difficult in RBA than in RBF, and in particular for bilateral donors.

PbR seems to be most appropriate if the objectives and interests of principal and agent are not fully aligned at the outset. PbR has been used to induce (commercial) agents to deliver goods and services to poor and vulnerable groups. Schemes that target the poor were often successful in

improving equity. However, we often do not know whether these programmes have unintended effects (for example, on quality or on other service areas) or whether they are cost effective in comparison with other interventions. The initial non-alignment of principal and agent will also increase the risks of gaming and manipulation.

The main risks and costs of PbR result from three trade-offs in establishing performance indicators. First, it is difficult to establish indicators that can be sufficiently influenced by the agents *and* lead to the ultimate objective *plus* stimulate autonomy and innovation. The experiences in health and, particularly, in education show that incentivizing processes or even outputs like school completion rates may not lead to better educated pupils. At the same time, indicators at the lower end of the results chain, like school enrolment rates or number of antenatal visits, do not incentivize innovation for achieving the ultimate learning or health outcomes. Second, there is a trade-off between predictability of disbursement *and* incentivizing performance. Donors tend to select easy-to-reach indicators at process or output level in order to safeguard disbursements, but these indicators hardly stimulate performance. Third, the higher the material incentive, the higher the chances are of distortion and gaming (Table 6).

Table 6. Trade-offs in choosing performance indicators for PbR

1	Can be influenced by agent	<>	Lead to ultimate objective Stimulate innovation
2	Ensure predictable disbursement	<>	Incentivize performance
3	Provide sufficient material incentive	<>	Avoid gaming or distortion

However, these trade-offs vary by sector, types of agents and circumstances. The first and second trade-off are particularly serious in sectors with a long results chain, so if there are many steps between inputs and final results, for example in health and education. If the results chain is shorter, for example in water and energy, these dilemmas hardly play a role. Moreover, the first trade-off is probably more severe in education than in health, given the somewhat higher certainty about the relationship between outputs and outcomes in health than in education. For example, there is probably a stronger relationship between institutional deliveries and (reduced) maternal mortality, than between school enrollment and literacy. The second dilemma is more severe if there is greater mutual dependence between principal and agent, for example in an aid

relationship between a donor country and a low-income recipient country, or between higher and lower government levels. This dependency is absent if the agent is a commercial actor and also weaker in case of an NGO, provided the latter has multiple sources of funding. The third trade-off is more serious when agents are supposed to engage in multiple and complex tasks, and when the quality and results of these tasks are not easily measurable. These two circumstances hold, in particular, for the health sector, where incentizing one health service (e.g. institutional deliveries) may lead to the neglect of other activities, and where the quality of the incentivized service is difficult to measure. But also in education the tasks are complex and the quality is not easily measurable.

It seems that PbR can have an added value if it can mobilize private resources for delivering goods and services to the poor, and especially in sectors, like energy or water, where the results chain is short and the results are tangible and can be measured relatively easily. In other sectors and situations, the added value of PbR is much less certain. If donors and governments can be expected to have the same objectives (say, improving learning outcomes), there is no need for having disbursements depend on costly and risky definitions and measurements of targets. The same holds for donors or national governments vis-à-vis sub-national governments, health facilities or schools. PbR can be valuable in specific contexts, for example if there is a need to improve services for the poor or for vulnerable groups like mothers and new-borns. In most cases, however, implementing agents are in need of resources in order to achieve the – shared – objectives. Flexibility and innovation of agents can be fostered by providing them with core financing, or budget support. Depending on circumstances this core financing (of sub-national government, facilities) or budget support (of countries) can be accompanied by regulation (e.g. minimum standards of care) or by a policy dialogue, and with technical assistance. All in all, the drive to payment by results in sectors like health and education often seems to be more induced by domestic motivations to "show results" than by considerations of aid effectiveness.

References

60_Decibels. 2021. Beyond the Grid Fund Zambia: Energy Service Subscription Verification & Customer Insights: 60_Decibels.

Andrews, M., Pritchett, L., & Woolcock, M. (2012). *Escaping capability traps through problem-driven iterative adaptation (PDIA).* (Working Paper No. 299). Washington: Center for Global Development.

Angelsen, A. 2014. "The Economics of REDD+." In *Handbook of Forest Resource Economics*, 290–306: Taylor and Francis. doi:10.4324/9780203105290-26.

Anselmi, L., Binyaruka, P., & Borghi, J. (2017). Understanding causal pathways within health systems policy evaluation through mediation analysis: An application to payment for performance (P4P) in Tanzania. *Implementation Science, 12*(1), 1–18.

Basinga, P., Gertler, P. J., Binagwaho, A., Soucat, A. L., Sturdy, J., & Vermeersch, C. M. (2011). Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: An impact evaluation. *The Lancet, 377*(9775), 1421–1428.

Bergman, R., Forsberg, B. C., & Sundewall, J. (2021). Results-based financing for health: A case study of knowledge and perceptions among stakeholders in a donor-funded program in Zambia. *Global Health Science and Practice, 9*(4), 936–947. https://doi.org/10.9745/GHSP-D-20-00463

Bernal, P., Martinez, S., & Celhay, P. (2018). Is results-based aid more effective than conventional aid?: Evidence from the health sector in El Salvador. IDB Working Paper Series No. 589.

Binyaruka, P., & Borghi, J. (2017). Improving quality of care through payment for performance: Examining effects on the availability and stock-out of essential medical commodities in Tanzania. *Tropical Medicine & International Health, 22*(1), 92–102.

Binyaruka, P., Robberstad, B., Torsvik, G., & Borghi, J. (2018a). Does payment for performance increase performance inequalities across health providers? A case study of Tanzania. *Health Policy and Planning, 33*(9), 1026–1036.

Binyaruka, P., Robberstad, B., Torsvik, G., & Borghi, J. (2018b). Who benefits from increased service utilisation? examining the distributional effects of payment for performance in Tanzania. *International Journal for Equity in Health, 17*(1), 1–16.

Borghi, J., Binyaruka, P., Mayumana, I., Lange, S., Somville, V., & Maestad, O. (2021). Long-term effects of payment for performance on maternal and child health outcomes: Evidence from Tanzania. *BMJ Global Health, 6*(12), e006409.

Brook, P. J., & Smith, S., eds. (2001). *Contracting for public services: Output-based aid and its applications*. The World Bank.

Cassidy, R., Tomoaia-Cotisel, A., Semwanga, A. R., Binyaruka, P., Chalabi, Z., Blanchet, K., Singh, N. S., Maiba, J., & Borghi, J. (2021). Understanding the maternal and child health system response to payment for performance in Tanzania using a causal loop diagram approach. *Social Science & Medicine, 285*, 114277.

CCA and MECS (2022). Clean Cooking RBFs: Key Design Principles.

Chimhutu, V., Tjomsland, M., & Mrisho, M. (2019). Experiences of care in the context of payment for performance (P4P) in Tanzania. *Globalization and Health, 15*(1), 1–13.

Clist, P. (2016). Payment by results in development aid: All that glitters is not gold. *World Bank Research Observer, 31*(2), 290–313. https://doi.org/10.1093/wbro/lkw005

Clist, P. (2019). Payment by results in international development: Evidence from the first decade. *Development Policy Review, 37*, 719–734. https://doi.org/10.1111/dpr.12405

Clist, P., & Verschoor, A. (2014). The conceptual basis of payment by results. Norwich: University of East Anglia, School of International Development.

Collier, Paul, Patrick Guillaumont, Sylviane Guillaumont, and Jan Willem Gunning. 1997. "Redesigning Conditionality." *World Development* 25 (9): 1399–1407.

DFID. (2014). DFID's evaluation framework for payment by results. London: DFID. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/436051/Evaluation-Framework-Payment-by-Results3.pdf

Diaconu, K., Falconer, J., Verbel, A., Fretheim, A., & Witter, S. (2021). Paying for performance to improve the delivery of health interventions in low- and middle-income countries. *Cochrane Database of Systematic Reviews, 2021*(5). https://doi.org/10.1002/14651858.CD007899.pub3

Dijkstra, Geske. 2002. "The Effectiveness of Policy Conditionality: Eight Country Experiences." *Development and Change* 33 (2) (April): 307–334.

Dijkstra, G. (2018). *Budget support, poverty and corruption: A review of the evidence.* (EBA Report No. 2018-04). Stockholm: EBA, Expert Group for Aid Studies, Sweden. www.eba.se

Dijkstra, Geske, Antonie de Kemp, and Denise Bergkamp. 2012. *Budget Support: Conditional Results; Review of an Instrument, 2000–2011*. The Hague: Ministry of Foreign Affairs, Department for International Research and Policy Evaluation (IOB).

Dijkstra, Geske, Caldeyro, Martin, Käppler, Ruth M., & Kasprowicz, Leszek. (2020). *Evaluation of EU budget support to Rwanda (2011–2018), final report, volume 1.* (Evaluation carried out on behalf of the European Commission). Galway, Ireland: GDSI Limited. https://ec.europa.eu/international-partnerships/evaluation-eu-budget-support-rwanda-2011-2018_en

Dom, C., Fraser, A., Holden, J., & Patch, J. (2021). *Results-based financing (RBF) in the education sector: Country-level analysis; final synthesis report.* Submitted to the REACH Porgram at the World Bank by Mokoro Ltd.

Duber, H. C., Hartford, E. A., Schaefer, A. M., Johanns, C. K., Colombara, D. V., Iriarte, E., Palmisano, E. B., Rios-Zertuche, D., Zuniga-Brenes, P., Hernández-Prado, B., & Mokdad, A. H. (2018). Appropriate and timely antibiotic administration for neonatal sepsis in Mesoamérica. *BMJ Global Health, 3*(3). https://doi.org/10.1136/bmjgh-2017-000650

Duvendack, M. (2022). Payment-by-results for health interventions in low- and middle-income countries: A critical review. *Development Policy Review, 40*(1). https://doi.org/10.1111/dpr.12538

Eichler, R., & Glassman, A. (2008). Health systems strengthening via performance-based aid: Creating incentives to perform and to measure results. Brookings Global Economy and Development, Global Health Financing Initiative Working Paper 3.

Eichler, R., Agarwal, K., Askew, I., Iriarte, E., Morgan, L., & Watson, J. (2013). Performance-based incentives to improve health status of mothers and newborns: What does the evidence show? *Journal of Health Population and Nutrition, 31*(4 Suppl 2), S36–S47.

Eichler, R., Regalia, F., Gigli, S., Tapia Conyer, R., Kress, D., Wong, E., Mujica, R., Betancourt Cravioto, M., Ciria Matilla, M. C., Álvarez, M. C., Iriarte, E., Neldon, J., Ríos-Zertuche, D., & Zúñiga Brenes, P. (2017). External measurement as a catalyst for change in a regional results-based aid initiative – the Salud Mesoamerica experience. Inter-American Development Bank (IDB), Technical Note No. 1312.

El Bcheraoui, C., Palmisano, E. B., Dansereau, E., Schaefer, A., Woldeab, A., Moradi-Lakeh, M., Salvatierra, B., Hernandez-Prado, B., & Mokdad, A. H. (2017). Healthy competition drives success in results-based aid: Lessons from the Salud Mesoamérica Initiative. *PLoS ONE, 12*(10). https://doi.org/10.1371/journal.pone.0187107

El Bcheraoui, C., Kamath, A. M., Dansereau, E., Palmisano, E. B., Schaefer, A., Hernandez, B., & Mokdad, A. H. (2018). Results-based aid with lasting effects: Sustainability in the Salud Mesoamerica Initiative. *Globalization and Health, 14*, 97. https://doi.org/10.1186/s12992-018-0418-x ER

Elbers, Chris, Jan Willem Gunning, and Kobus de Hoop. 2009. "Assessing Sector-Wide Programs with Statistical Impact Evaluation: A Methodological Proposal." *World Development* 37 (2): 513–520.

Eldridge, Cynthia and Natasha Palmer. 2009. "Performance-Based Payment: Some Reflections on the Discourse, Evidence and Unanswered Questions." *Health Policy and Planning* 24: 160–166. doi:10.1093/heapol/czp002.

GPOBA (Global Partnership on Output-Based Aid). (2016). *Water sector experience of output-based aid*. The World Bank.

Grittner, A. M. (2013). *Results-based financing: Evidence from performance-based financing in the health sector*. German Development Institute (DIE) Discussion Paper 6/2013.

Hayman, Rachel. 2011. "Budget Support and Democracy: A Twist in the Conditionality Tale." *Third World Quarterly* 32 (4) (June): 673–688.

Hazeu, C. A. (2000). *Institutionele economie: Een optiek op organisatie- en sturingsvraagstukken*. Coutinho.

Helland, J., & Mæstad, O. (2015). *Experiences with results-based payments in Norwegian development aid*. Norad Evaluation department, Report 4/2015.

Hernandez, B., Harris, K. P., Johanns, C. K., Palmisano, E. B., Cogen, R., Thom, M. G., Linebarger, E., El Bcheraoui, C., Kamath, A. M., Camarda, J., Rios-Zertuche, D., Zuniga-Brenes, M. P., Bernal-Lara, P., Colombara, D., Schaefer, A., Salvatierra, B., Mateus, J. C., Casas, I., Flores, G., Mokdad, A. H. (2022). Impact of the Salud Mesoamerica Initiative on delivery care choices in Guatemala, Honduras, and Nicaragua. *BMC Pregnancy and Childbirth, 22*(1), 5. https://doi.org/10.1186/s12884-021-04279-2 ER.

Hill, T., Fredriksen, B. J., Isenman, P., Rosenberg, A., & Jayaram, S. (2015). Paying for performance: An analysis of output-based aid in education. Results for Development Institute, commissioned by GPOBA.

Huillery, E., & Seban, J. (2017). Money for nothing? The effect of financial incentives on efforts and performances in the health sector.

IEG (Independent Evaluation Group). (2016). *Program-for-results: An early-stage assessment of the process and effects of a new lending instrument*. The World Bank, IEG.

ITAD. (2014). *Independent evaluation of budget support in Mozambique, final report, volume 1*. ADE, ITAD and COWI.

Janus, H. (2014). Real innovation or second-best solution? first experiences from result-based aid for fiscal decentralisation in ghana and Tanzania. German Development Institute Discussion paper.

Kalk, A., Paul, F. A., & Grabosch, E. (2010). 'Paying for performance' in Rwanda: Does it pay off? *Tropical Medicine and International Health, 15*(2), 182–190. https://doi.org/10.1111/j.1365-3156.2009.02430.x

Killick, Tony, Ramani Gunatilaka, and Ana Marr. 1998. *Aid and the Political Economy of Policy Change*. Routledge.

Lawson, A. e. a. (2016). *Evaluation of budget support to Sierra Leone 2002–2015, final report volume 1*. (Report commissioned by DfID, UK and managed by Evaluation Unit of DG International Cooperation and Development, EuropeAid). Rotterdam: Ecorys and Fiscus.

Lawson, A., & with Gonzalo Contreras, Gonzalo Alvarez de Toledo, and Virginie Morillon. (2014). *Synthesis of budget support evaluations: Analysis of the findings, conclusions and recommendations of seven country evaluations of budget support, volume 1*. (Report commissioned by the European Commission, DG for Development and Cooperation, EuropeAid). Oxford UK: FISCUS and ADE.

Lee, J. D., & Medina, O. (2019). *Results-based financing in education: Learning from what works*. The World Bank, REACH. http://hdl.handle.net/10986/31250

Maraviglia, A. R. (2011). *Effectiveness of performance driven aid: The case of Colombia's expanded immunization program*. PhD thesis. George Washington University.

McGillivray, M., & Pham, T. K. C. (2017). Reforming performance-based aid allocation practice. *World Development, 90*, 1–5. https://doi.org/10.1016/j.worlddev.2015.05.006

Mayumana, Iddy, Jo Borghi, Laura Anselmi, Masuma Mamdani, and Siri Lange (2017). "Effects of Payment for Performance on Accountability Mechanisms: Evidence from Pwani, Tanzania." *Social Science & Medicine* 179: 61–73.

Meuth Alldredge, Josh, Emma De Roy, Elangtlhoko Mokgano, Peter Mwandri, Tulika Narayan, Martin Prowse, Jyotsna Puri, William Rafferty, Anu Rangarajan, and Faraz Usmani (2020). *Evidence Review on Results-Based Payments: Evidence Gap Map and Intervention Heat Map. IEU Learning Paper*. Songdu, South Korea: Independent Evaluaiton Unit, Green Climate Fund.

Molenaers, Nadia, Linas Cepinskas, and Bert Jacobs. 2010. *Budget Support and Policy/Political Dialogue: Donors Practices in Handling (Political) Crises*. Antwerp: University of Antwerp/IOB.

Moran, G., Connal, C., Kirama, S., & Leung, Y. (2020). *Evaluation of the Sida-supported education program for results (EPforR) 2014–2021, Tanzania*. Nordic Morning. www.sida.se/publications

Mumssen, Y., Johannes, L., & Kumar, G. (2010). *Output-based aid: Lessons learned and best practices*. The World Bank.

Musgrove, P. (2011). Financial and other rewards for good performance or results: A guided tour of concepts and terms and a short glossary. The World Bank.

Ogundeji, Y. K., Bland, J. M., & Sheldon, T. A. (2016). The effectiveness of payment for performance in health care: A meta-analysis and exploration of variation in outcomes. *Health Policy, 120*(10), 1141–1150. https://doi.org/10.1016/j.healthpol.2016.09.002

Öhler, H., Nunnenkamp, P., & Dreher, A. (2012). Does conditionality work? A test for an innovative US aid scheme. *European Economic Review, 56*(1), 138–153. https://doi.org/10.1016/j.euroecorev.2011.05.003

Oxman, A. D., & Fretheim, A. (2009). Can paying for results help to achieve the millennium development goals? A critical review of selected evaluations of results-based financing. *Journal of Evidence-Based Medicine, 2*(3), 184–195. https://doi.org/10.1111/j.1756-5391.2009.01024.x

Park, S., & Kwak, J. S. (2017). Understanding results-based conditionality in development cooperation: A comparative case analysis. *Journal of International and Area Studies, 24*(1), 125–146.

Paul, E. (2015). Performance-based aid: Why it will probably not meet its promises. *Development Policy Review, 33*(3), 313–323. https://doi.org/10.1111/dpr.12115 ER.

Pearson, M. (2011). Results based aid and results based financing: What are they? have they delivered results. London: HLSP Institute.

Pearson, M., Johnson, M., & Ellison, R. (2010). *Review of major results based aid (RBA) and results based financing (RBF) schemes.* London: Department for International Development (DFID), UKAid.

Pereira, J., & Villota, C. (2012). *Hitting the target? evaluating the effectiveness of results-based approaches to aid.* Brussels: EURODAD. https://www.oecd.org/dac/peer-reviews/Hitting-the-target.pdf

Petrosyan, V., Melkomian, D. M., Zoidze, A., & Shroff, Z. C. (2017). National scale-up of results-based financing in primary health care: The case of Armenia. *Health Systems and Reform, 3*(2), 117–128. https://doi.org/10.1080/23288604.2017.1291394

Ronsholt, E. (2014). *Review of budget support evaluations.* (Evaluation Study No. 1). Copenhagen: Ministry of Foreign Affairs of Denmark, DANIDA.

Sabbi, M., & Stroh, A. (2020). The "numbers game": Strategic reactions to results-based development assistance in Ghana. *Studies in Comparative International Development, 55*(1), 77–98. https://doi.org/10.1007/s12116-019-09296-z ER.

Salehi, A. S. (2020). *Political economy analysis and economic evaluation of results-based financing in Afghanistan.* PhD Thesis. London School of Hygiene & Tropical Medicine.

Sandefur, J., & Glassman, A. (2015). The political economy of bad data: Evidence from African survey and administrative statistics. *Journal of Development Studies, 51*(2), 116–132. https://doi.org/10.1080/00220388.2014.968138

Singh, N., Kovacs, R. J., Cassidy, R., Kristensen, S. R., Borghi, J., & Brown, G. W. (2021). A realist review to assess for whom, under what conditions and how pay for performance programmes work in low- and middle-income countries . *Social Science & Medicine, 270*. https://doi.org/https://doi.org/10.1016/j.socsimed.2020.113624

Soeters, R., Peerenboom, P. B., Mushagalusa, P., & Kimanuka, C. (2011). Performance-based financing experiment improved health care in the Democratic Republic of Congo. *Health Affairs, 30*(8), 1518–1527. https://doi.org/10.1377/hlthaff.2009.0019

Swedlund, Haley J. 2013. "From Donorship to Ownership? Budget Support and Donor Influence in Rwanda and Tanzania." *Public Administration and Development* 33 (5): 357.

Terway, A., Burnett, N. R., & Frotté, M. D. (2021). Results-based financing in education for sub-national government and school : A conceptual framework and practical recommendations. The World Bank. https://openknowledge.worldbank.org/handle/10986/37028

Appendix: Tables 7–10

Table 7. Overview of results of the reviews in health

Type	Number [1]	Author	Year	Syste-matic	Region	Sector	Output	Outcome	Impact	Quality	Un-intended effects[2]	Data[3]	Innova-tion	Targe-ting	Equity	Costs	Prin-cipal
A	4	Oxman	2009		S	Health	75% (4)				yes (2)	pos (1)			50% (1)		
G	12	Grittner	2013		S	Health	55% (9)	33% (3)	67% (3)	n.e.	yes (2)			80% (5)	neg (1)	mixed (4)	
A	9	Eichler	2013		S	Health	75% (8)		n.e.	n.e.		n.e.					
A	*37*	Ogundeji	2016	Yes	G	Health		70%[4]									n.e.
A	*81*	Duvendack	2020	Yes	S	Health		85%[5]									
G	*59*	Diaconu	2021	Yes	S	Health	pos	pos		pos					mixed		
A	36, *117*	Singh	2021	Yes	S	Health	59% (32)			71% (7)	yes (12)	pos (4)				mixed (6)	
A	2	Salehi	2020	Yes	S	Health										mixed (2)	

Legend:
Pos: positive; neg: negative; mixed: both positive and negative results
Under Type: A= academic, G= Grey
Under Region: G= Global, S = Global South
Under output, outcome, impact: percentages indicate share of projects with positive effects; projects with mixed effects get a score of 0.5
Number in brackets indicates number of studies or programmes for which this is reported

“n.e.”: no evidence as reported by author(s) of studies. “n.e.”: my assessment.
Text and numbers in red: very weak evidence, in purple: weak evidence.
1 Number means number of programmes or number of studies (in italics).
2 Unintended effects include distortion, gaming, manipulation, neglect of other areas of service, etc.
3 Data includes transparency, accountability, improved data systems.
4 More effect with less rigorous methods and at process or output level
5 No difference in quality of method, most studies only assess process or output indicators, often there is no independent verification.

Table 8. Overview of results of reviews in other sectors

Type	Number[1]	Author	Year	Syst	Region	Sector	Output	Outcome	Impact	Quality	Unintended effects[2]	Data[3]	Innova-tion	Targe-ting	Equity	Costs	Prin-cipal
G	24	Hill	2016		S	education		n.e.						pos			
G	41	Lee	2019		G	education	50%	50%			some						
G	51	Terway	2021		S	education		50%									
G	12	GPOBA	2016		S	water		n.e.				mixed	neutral	pos	mixed	pos (2)	
G		Brook	2002		S	multiple		n.e.					pos			pos	
G	112	Mumssen	2010		S	multiple		pos[4]				mixed	pos	pos		pos	
G	27	IEG	2016		S	multiple		n.e.					neg			pos	
G		Pearson	2010		S	multiple		n.e.									
G	6	Eichler[5]	2008		S	multiple	57% (4)										
G	4	Helland[5]	2015		S	multiple	25% (2)				some				neg (1)	neg	neg (2)
A	9	Clist[5]	2019		S	multiple	58% (6)			neg (1)	mixed (2)	neg (3)	neg (2)		pos (1)		
A	4	Park[5]	2017		S	multiple	25% (2)								50% (1)		neg
G	15	CCA	2022		S	energy		n.e.						pos			

Legend and notes 1, 2 and 3: see Table 7.

pos: positive, neg: negative, mixed: both positive and negative results, neutral: no effect.

[4] This source assesses effectiveness on the basis of Implementation Completion Report (ICR) ratings.

[5] The programmes discussed in these reviews to a large extent overlap with programmes discussed in other reviews.

Table 9. Overview of case studies in health

In # reviews	Type	Author	Year	Country/ region	Methods	Inputs	Output	Outcome	Quality	Un-intended effects	Align-ment	Data	Inno-vation	Tar-geting	Equity	Costs
in 4	A	Basinga	2011	Rwanda	quant		mixed		pos							
in 2	A	Kalk	2010	Rwanda	qual					yes						
in 3	A	Soeters	2011	DRC	quant		mixed		pos			pos				pos
in 2	A	Huillery	2017	DRC	quant		neg	neg		no						neg
in 1	A	Binyaruka	2018b	Tanzania	quant	pos	mixed								mixed	
in 2	A	Binyaruka	2017	Tanzania	quant	pos									pos	
in 2	A	Anselmi	2017	Tanzania	quant	pos			pos			pos				
in 1	A	Cassidy	2021	Tanzania	mixed	pos						mixed				
in 1	A	Chimhutu	2019	Tanzania	qual		pos		pos	yes		pos				
	A	Borghi	2021	Tanzania	mixed							pos				
in 1	A	Mayumana	2017	Tanzania	mixed							mixed				
in 1	A	Binyaruka	2018a	Tanzania	quant										neg	
	A	Bergman	2021	Zambia	qual			neg			neg					
	A	Petrosyan	2017	Armenia	qual		pos			yes	pos					
	A	Maraviglia	2011	Colombia	quant					no					n.e.	pos
	A	Salehi	2020	Afghani-stan	quant		neutral		pos							neg
in 1	G	Bernal	2018	El Salvador	quant	pos	pos							pos		pos
	A	Hernandez	2022	Nic, Hond, Guat	quant	neutral	pos							pos		

In # reviews	Type	Author	Year	Country/ region	Methods	Inputs	Output	Outcome	Quality	Un-intended effects	Align-ment	Data	Inno-vation	Tar-geting	Equity	Costs
	A	Duber	2018	Belize, Gua, Hon, Nica, Mexico,	quant	pos	neutral									
	A	El Bcheraoui	2017	CA plus Chiapas	qual		pos					pos				
	G	Eichler	2017	CA plus Chiapas	qual		pos					pos	pos			
	A	El Bcheraoui	2018	CA plus Chiapas	qual	pos								pos		

Legend: See Table 7. In green: weak evidence as assessed by authors of studies.

Table 10. Overview of case studies in other sectors

in # reviews	Type	Author	Year	Region/ country	Sector	Methods	Inputs	Output	Outcome	Quality	Unintended effects	Alignment	Data	Innovation	Targeting	Equity	Costs
	G	Moran	2020	Tanzania	education	qual			mixed				pos			mixed	
	G	Dom	2021	Tanzania, Moz.,Nepal	education	qual	neg (1)	some	n.e.		yes	mixed	mixed	neg		mixed	
	A	Öhler	2012	S	multiple	quant			pos								
	G	Janus	2014	Ghana, Tanzania	Decentralization	qual			some			mixed					pos
	A	Sabbi	2020	Ghana	Decentralization	mixed			neg		yes		neg				neg
	G	60-decibels	2022	Zambia	energy	qual		n.e.		mixed		pos			pos		

Legend: see Tables 7, 8 and 9. In purple: weak evidence according to my assessment