Department for International Development

Semi-systematic review to understand Payments-by-Results mechanisms in developing countries

DFID COVER NOTE

UKaid from the British people

What: The enclosed **semi-systematic review** attempts to:

- identify all available high-quality quantitative research of international development programmes using payment by results;
- identify trends and generalizable lessons on how, whether and under what circumstances PbR programmes were successful in international development contexts.

Who: this research was undertaken by an independent team at the University of East Anglia (UEA) led by Professor Maren Duvendack. The findings of this independent research are the researchers own, and do not necessarily reflect the views of DFID. In this cover note, the commissioning DFID team highlight key findings, limitations and recommendations.

Why: DFID's 2014 PbR Strategy[i] and 2015 Evaluation Framework for PbR[ii] commit DFID to build the evidence base of what works, and build capability to deliver PbR programmes well.

How: the UEA team sifted several thousand peer-reviewed evaluations to identify the most robust research which assessed whether PbR programmes in international development had statistically significant effects. The team identified 100 research reports which met their quality criteria, and analysed these in more detail. Twenty-three of the associated programmes had been funded by DFID, and 77 were funded by other donors and agencies. The researchers considered quantitative approaches only and screened them against their quality criteria.

Findings:

- The published, peer-reviewed evaluation evidence on PbR mechanisms in sectors beyond health and education remains scant.
- A majority of the qualifying research showed that the PbR programmes they examined had statistically significant, positive effect on their targeted result.
- Research using weaker methods with higher 'risk of bias' did not tend to report more positive findings than research using the most rigorous methods. The majority of studies in both groups found statistically significant, positive effects.

Limitations:

- Beyond health and education, the team were unable to draw robust, generalizable findings on where, how and in what circumstances PbR might be more or less successful. The high number of qualifying studies, and the large diversity of programme interventions, uses of PbR, and contexts made comparison and inference difficult.
- In addition, this review does *not consider complementary qualitative research* or other available information which could shed additional light on **why** and **how** positive or negative results were achieved in different circumstances.

Primary value: this report has rigorously sifted several thousand publications to create a database of dozens of academically robust, quantitative research papers which examine a broad range of PbR programmes in international development contexts. The report and in particular annexes will be a valuable reference resource for future researchers and programme designers, allowing highest-quality, most-relevant research to be identified quickly and with relative ease.

DFID response to the recommendations:

In line with the terms of reference, the independent researchers have put forward recommendations to DFID in this report. Unfortunately the limitations discovered during the course of the research, and highlighted above, mean that these recommendations are in part based on the subjective professional judgement of the research team rather than on robust findings established from the literature. As a result DFID accepts some but not all of the recommendations put forward, but is grateful to Professor Duvendack and her team for the spirit and integrity of the suggestions put forward.

Recommendations which we would particularly like to highlight include:

- The importance, in successful payment by results programmes, of paying close attention to the strength of the underlying **data** used to measure and verify results. The associated importance of factoring the cost of high quality data into programme design early on (see below). That **independent evaluators** can add value on this from an early stage in the design of PbR interventions.

- To identify the effect of a PbR mechanism on development outcomes, there is a need for programmes to have **credible counterfactuals**, ideally comparing the chosen PbR approach to alternative payment models for delivering the same results. More programme research which considers this will help the international development community address the question of whether PbR or alternative contracting models offer greatest impact in different circumstances.

- The need for the international development research community to investigate further the way PbR impacts on the **incentives of our suppliers**.

- The need for additional care and caution to continue to be applied to all forms of programming intervention in **fragile and conflict-affected states,** particularly to PbR which is an especially challenging form of contracting.

- There is a need to better record the actual costs of **measurement and verification** of results. Initial estimates by programme teams and implementing partners are likely to underestimate the cost, time and complexity involved.

- We would respectfully consider that readers are cautious when considering the assertion put forward in this report that **output-level indicators** are flawed, and note that this was not a finding robustly substantiated from the evidence which was assessed in this semi-systematic review, but rather a hypothesis put forward by the research team. We do, however, respect the independence of the researchers and their views.

How we will use these findings:

DFID will incorporate these insights into our institutional PbR learning, training, guidance and support to programme design teams, as part of DFID's institutional learning strategy.

We thank the researchers for their efforts and in particular the constructive approach of Professor Duvendack during consultations with DFID. The inferences and conclusions reached are those of the UEA researchers alone.

---

[i] https://www.gov.uk/government/publications/dfids-strategy-for-payment-by-results-sharpening-incentives-to-perform.

[ii] https://www.gov.uk/government/publications/dfids-evaluation-framework-for-payment-by-results

FULL REPORT: Review Of Payment By Results In DFID: Establishing The Evidence Base

Semi-systematic review to understand Payments-by-Results mechanisms in developing countries

Dr Maren Duvendack, July 2017, University of East Anglia[1]

Abstract

We conducted a semi-systematic review on Payments-by-Results (PbR) in international development with a focus on synthesising the health evidence. The objective of this review was to draw out lessons learnt about what works for the implementation and impact of PbR mechanisms in developing country government and non-DFID donor/NGO-led programmes.

The main research questions guiding this work were:

- What PbR instruments work best in different circumstances?
- When and how do PbR incentives work in practice?
- What is the value for money of different types of PbR instruments?

A number of evidence gap maps have been included to complement the semi-systematic review and to highlight some of the research gaps in the PbR literature - this has been an important part of many systematic reviews and will be useful to guide future priorities.

We identified a highly heterogeneous evidence base with 71 studies covering a range of health sub-themes, geographical regions and methodological quality. We found that governments are the dominant principal in 63% of the health evidence with health facility and individuals being the dominant agent. In terms of PbR measures, 60% of the evidence reported output level measures with high levels of heterogeneity (e.g. 26 different PbR measures across 71 studies). We assessed success and failure of PbR measures and observed that positive and significant effects dominate; however, publication bias may play a role to explain these observations. We have learnt that PbR mechanisms are not well understood outside of health and education.

[1] Acknowledgements: Many thanks to Sonja Marzi and Asta Hansen for research assistance.

Table of Contents

1. Introduction

This report synthesizes the health evidence drawing on a systematic review methodology. A systematic review is a variation of a literature review but more rigorous and transparent following clear protocols. Systematic reviews originated in the medial sciences in the 1970s to understand the effectiveness of health care interventions but they have now been adopted by a wide range of academic disciplines to understand what works. According to Mallet et al (2012) "systematic reviews involve identifying, synthesising and assessing all available evidence, quantitative and/or qualitative, in order to generate a robust, empirically derived answer to a focused research question" (p. 445-446).

The purpose of this review is to allow us to generate lessons learned about what works for the implementation and impact of Payments-by-Results (PbR) mechanisms in developing country government and non-DFID donor/NGO-led programmes.

The main research questions guiding this work are:

- What PbR instruments work best in different circumstances?
- When and how do PbR incentives work in practice?
- What is the value for money of different types of PbR instruments?

2. The method: Semi-systematic review process

As indicated in the title, this review is semi-systematic which means that we follow the systematic review process where possible but due to limited resources, we adjust it in terms of breadth and depth of search and screening processes, quality assessment and synthesis. To complement the semi-systematic review approach we include a number of evidence gap maps to highlight some of the research gaps in the PbR health literature – this has been an important part of many systematic reviews (Gough et al, 2013).

The report is structured as follows: We first outline the search methodology, this is followed by providing details on screening, search results and data extraction. The next step involves an assessment of validity and quality of the evidence before presenting the evidence gap maps – these steps all relate to the health and non-health evidence, however, the synthesis documented in this report focuses on the health evidence. Finally, an in depth synthesis which is guided by the Measure-Agent-Principal (MAP) framework developed by Clist (2016) is presented where we assess success and failure of PbR measures.

Search methodology

Search strategy – External evidence

This section outlines our search strategy allowing us to identify the relevant literature related to programmes using PbR. The search process is the same for the health and non-health evidence. The search process was initiated by a snowballing approach that focused on 5 existing systematic and non-systematic reviews (Table 1), the results of the snowballing exercise were complemented by extensive electronic searches of academic and institutional repositories as well as Google Scholar.

Table 1: Snowballing of systematic and non-systematic reviews

Authors	**Title**	**Published By**
Mason, Fullwood, Singh, and Battye (2015)	Payment by Results Learning from the Literature	ICF International
Perrin (2013)	Evaluation of Payment by Results (PBR): Current Approaches, Future Needs	Department for International Development (DFID)
Webster (2016)	Payment by Results: Lessons from the Literature	Russell Webster
Eijkenaar (2012)	Pay for Performance in Health Care An International Overview of Initiatives	Medical Care Research and Review
Oxman and Fretheim (2009)	Can paying for results help to achieve the Millenium Development Goals? Overview of the effectiveness of results-based financing	Journal of Evidence-Based Medicine

'Snowballing' through the bibliographies of 5 well-known systematic and unsystematic reviews allowed us to create an initial database of relevant key evidence.

Table 2: List of key terms

"big results now"	"payment by results"
"cash on delivery aid"	"paying for results"
"financial incentives"	"payment to results"
"incentive contracts"	"performance-based contracting"
"output based aid"	"performance-based incentive"
"outcome-based commissioning"	"performance-based aid"
"outcome-based contracting"	"performance-based financing"
"outcome-based payment"	"performance incentives"
"payment for performance"	"performance related pay"
"pay for performance"	"program-for-results"
"pay for quality"	"results-based financing"
"pay for success"	"results based aid"
"payment by outcome"	

This key evidence was then reviewed to derive a list of key terms (Table 2) that best describe PbR and related financing schemes. The list of key terms then informed the search strings we employed in the electronic searches of a range of academic and institutional repositories (Table 3).

Table 3: Electronic databases

Academic Repositories
- Web of Science (WoS) - Science Direct (SD)
Institutional Repositories
- Centre for Global Development (CGD) - Deutsches Institut fuer Entwicklungspolitik (DIE) - World Bank Open Knowledge Repository (WB) - Overseas Development Institute (ODI) - Institute of Development Studies (IDS) - International Initiative for Impact Evaluation (3ie) - Research4Development (R4D) - OECD iLibrary - African Development Bank (AfDB) - Asian Development Bank (ADB) - Inter-American Development Bank (IDB)
Other Repository
- Google Scholar (GS)

Given this is a semi-systematic review involving limited resources, we adjusted the breadth and depth of the search process. An initial search in Google Scholar and Science Direct using the search terms outlined in Table 2 returned a very high number of records (many of them irrelevant when subjected to our inclusion criteria outlined in the next section); hence, we re-visited and refined the key terms to optimise the search process[2]. Details of this iterative search process are reported in Appendix 1 and Appendix 2.

Search strategy – DFID evidence

Searching academic and institutional databases led us to identify some DFID funded studies but this formal search approach did not allow us to capture all the evidence generated by DFID internally and hence the use of a range of more informal search methods was required.

One of these informal methods led us to trawl through an Excel file given to us by DFID where all 19 DFID funded PbR programmes were listed with links to their corresponding DEV tracker pages. Going through all the documents made available on DEV tracker for all the completed and on-going projects the most recent documents with information about the set-up of each programme was identified and compiled for later screening and data extraction. At the time of screening, 186 files were available on DEV tracker for 19 projects; these were all screened by type of document, time of publication and title. For each project, one document was included for later full-text screening.

In addition, DFID staff involved in the PbR projects gave us access to internal documents by using Trello (30 documents). The search process here was similar to the one used for DEV tracker.

After this initial search of documents produced by DFID, we searched the GPOBA database, as DFID is the founder of this donor partnership, and identified 139 documents (working papers, lessons learned papers, annual reviews). All 139 documents were screened full-text.

Similarly, HRITF, partly funded by DFID, shared a set of 16 studies/reports on on-going projects with us which we screened full-text as well.

Publications by DFID on the GEC were also screened following the same approach as with GPOBA and HRITF. Documents published on the relevant governmental GEC website were screened by title

[2] E.g. in Science Direct, an iterative search process was adopted where key terms were adjusted after an examination of the search results in terms of relevance. If many of the search results were still irrelevant (judged by the inclusion criteria described further below), we further optimised the key terms and restricted the search to titles and abstracts to improve the search output. Similarly, in Google Scholar, the original key terms listed in Table 2 were combined with additional search terms using the AND Boolean operator. E.g., the terms "big results now" and "Tanzania" and "program-for-results" and "World Bank" were combined because we knew that some studies we identified were specific to Tanzania and the World Bank. We also combined the terms in Table 2 with additional terms such as: i) development ii) international development iii) development aid iv) aid.

(20 documents). When the titles seemed to be of interest the full text was screened (6 documents), however none of these documents met our inclusion criteria (described in the next section) and were therefore not included. In addition, GEC staff shared 15 midline reports with us which were all screened by full text and included.

In total, we identified 406 internal DFID documents of which 199 were screened full-text, the remainder were screened only by title and abstract, 23 of 406 DFID documents met our inclusion criteria.

Screening and data extraction

After concluding the search process the documents in the database were initially screened by our research assistant with independent cross-checking by PI and Co-I[3]. The following inclusion criteria guided the screening process:

1. At least one lower/middle income country[4]
2. Quantitative empirical data, e.g. primary and/or secondary survey data drawing on a range of research designs such as randomised controlled trials (RCTs), quasi-experiments, basic surveys, etc.
3. Payment according to a pre-agreed measure

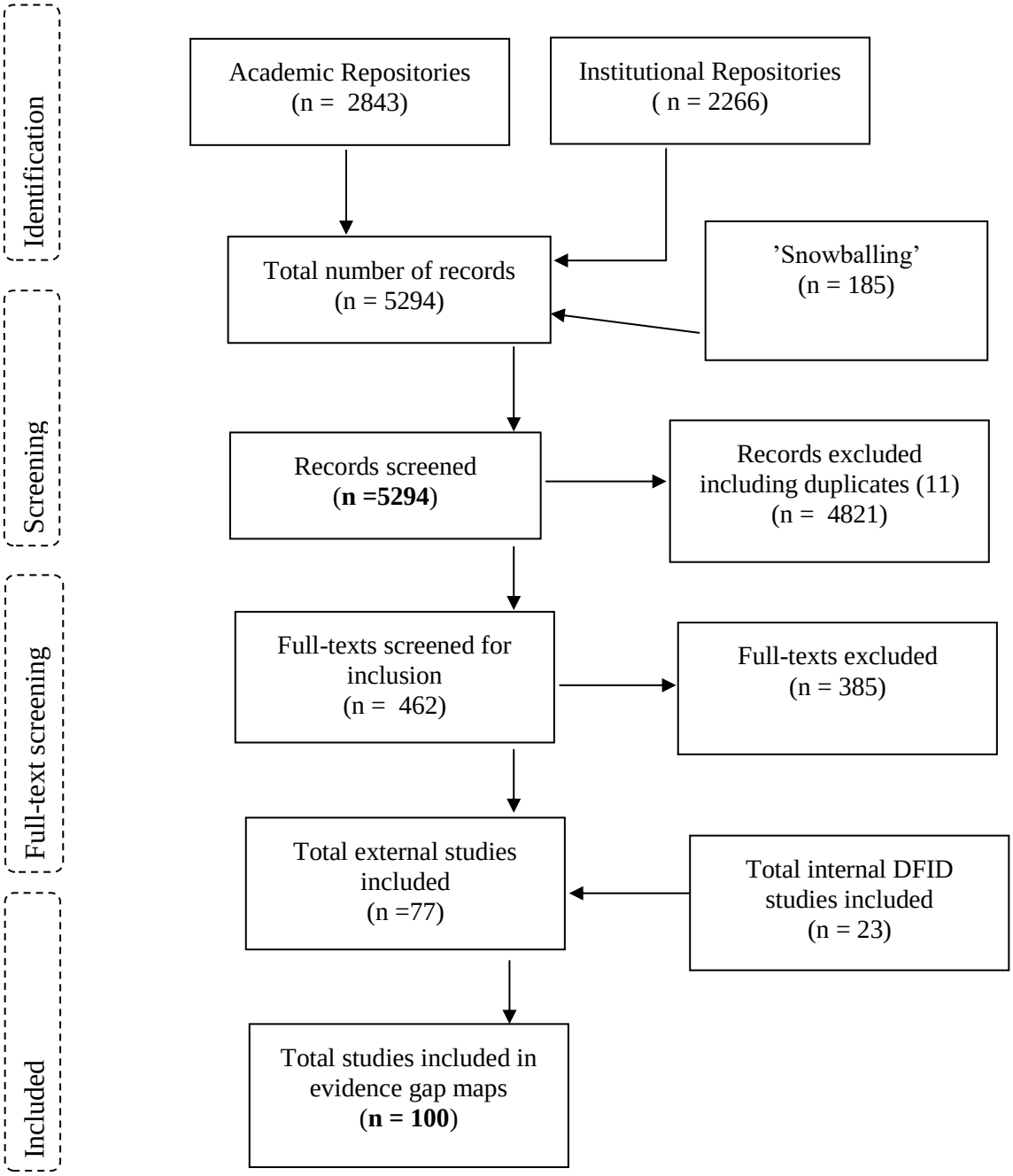
As common in the systematic review setting, we use a PRISMA diagram (Table 4) to present the different stages of the search and screening process. The initial snowballing approach that was complemented by searching academic and institutional repositories led us to identify 5294 records. These records were screened by title and abstract using the inclusion criteria listed above. We excluded 4832 records (this included 11 duplicates) as they did not meet at least one of our inclusion criteria. Of these 5294 records, 462 did not allow us to make an assessment based on screening title and abstract and hence full-text reading was required to judge their inclusion or exclusion. This is not an unusual additional step in the screening process as many studies in the social science and international development context are often not written in uniform ways and this creates challenges from a systematic review perspective (see Mallet et al, 2012 for a discussion). After reading 462 full-texts, we excluded 385 as they did not meet at least one of our inclusion criteria. This left us with 77 records to which 23 records were added that were identified after searching and screening the DFID

[3] After completing the screening process, we shared our database of included studies with DFID for consultation. See email exchange initiated on 17 January 2017. The inclusion criteria were also agreed on in the same email trail and confirmed during the workshop at the DFID offices in London on 15 March 2017.

[4] The World Bank definitions of lower/middle income countries were used: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

evidence (as outlined in the ‘search methodology’ section above). In total, 100 studies remained for further review and synthesis.

Table 4: PRISMA Diagram



These 100 studies were then read in detail to extract information for quality assessment as well as synthesis. We extracted descriptive information from each of the studies using the adapted data extraction form designed by Boaz et al (2002), see Table 5 below.

Table 5: Basic data extraction tool

Details of publication
Author Title Source (journal, conference etc.) Year
Context and population
Country Sector Donor Who gets paid? What is the measure? How is payment calculated? How much was disbursed? Any treatment effect sizes reported?
Study design
Research design category Type of secondary data used Data source
Analysis
Statistical techniques used

Source: Adapted from Boaz et al (2002).

The content from the data extraction for each study fed into the validity and quality appraisal outlined in the next section.

Validity and quality appraisal

Assessing the validity and quality of evidence has a long history in the systematic review context. Systematic reviews commonly adopt criteria for assessing each piece of evidence according to whether it is reliable within its own methodological paradigm and aims. The quality appraisal process can and should be used to validate what constitutes evidence in relation to the specific question(s) that a particular study seeks to find answers to.

Criteria for judging validity used in the systematic review literature are often adapted from the Cochrane Handbook (Higgins and Green, 2011) and EppiCentre (Gough, 2007; EppiCentre, 2010). The Cochrane Collaboration suggests that the key components of bias (and therefore in assessment of validity) in any study are:

A. selection bias (systematic differences between baseline characteristics of the 2 groups);

B. performance bias (systematic difference between care or support provided to the 2 groups);

C. attrition bias (systematic differences between the 2 arms in withdrawals from the study);

D. detection bias (systematic differences between groups in how outcomes are determined); and

E. reporting bias (systematic differences between reported and unreported findings).

An important aspect of quality appraisal is the consideration of whether a particular study's methods are suitable in relation to the relevant research question(s) (e.g. qualitative evidence may not be relevant to a cross-comparison of regression analyses for an effectiveness review question on economic impacts). Thus, the EPPI-Centre advises appraising each piece of evidence according to their Weight of Evidence (WoE) scale where it formulates the risk of bias as being composed of the following:

F. trustworthiness of results (methodological quality, as discussed by Cochrane, including transparency, accuracy, accessibility and specificity of the methods);

G. appropriateness of the use of that study design to address the review question (methodological relevance, including purposivity);

H. appropriateness of focus for answering the review question (topic relevance, including relevant answers and legal and ethical propriety); and

I. overall weight of evidence (a summary of the above).

These criteria are often used to guide specific validity and quality assessments of methodologically and conceptually diverse studies. Given the dominance of quantitative evidence in the systematic review context, many quantitative quality appraisal tools, or also known as risk of bias tools, have been developed. Many of them focus on assessing the validity of experimental designs (e.g. Coalition for Evidence-Based Policy, 2010 for social experiments, Higgins and Green, 2011 for medical experiments) but increasingly tools for assessing quasi-experimental evidence have been developed (for an in depth list see Valentine and Cooper, 2008).

Many of these tools involve checklists that can be lengthy and subjective; therefore, Duvendack et al (2011, 2012) developed a quality appraisal tool that distils the risk of bias components outlined above into an objective scoring scheme. We adopt their scoring scheme to assess the quality of the 100 studies we included in this review. Their scheme categorises each study by scoring their reported research design and analytical method (see Table 6); these scores are then combined into an index. A cut-off point of 2 is applied, e.g. a study with a score of 2 and above is considered to have high threats to validity – this would correspond to a study based on a basic survey reporting descriptive statistics. Studies with scores of less than 2 have lower threats to validity, e.g. an RCT using econometric techniques would fall into this category (for details see Tables 6 & 7). We acknowledge that this tool is not without limitations (alternative quality assessment or risk of bias tools exist for quantitative

studies, see Waddington et al, 2012 for a discussion and overview) but perfectly feasible given our resource constraints and sufficient given our explicit focus on a semi-systematic review.

Table 6: Potential risk of bias in quantitative studies

	Statistical methods of analysis		
Research design	DID, PSM, IV, RDD	Multivariate (or bivariate with covariate means tests)	Tabulation
RCT	Low	Low	Low–Medium
Natural experiment	Low	Low	Low–Medium
Pipeline	Low–Medium	Medium-high	High
Panel	Low–Medium	N/A	High
Cross section	Low–Medium	High	High

Source: Duvendack et al. (2012).

Applying this tool to the 100 studies we included, we find the following:

Table 7: Potential risk of bias in PbR studies

	Statistical methods of analysis		
Research design	DID, PSM, IV, RDD	Multivariate (or bivariate with covariate means tests)	Tabulation
RCT	18	21	3
Panel	4	1	0
Cross section: b/a or w/wo	20	8	12
Basic survey	1	3	9

Source: Adapted from Duvendack et al. (2012).

Legend for Table 7:

Low score	43	High score	32
Medium score	25		

There were no natural experiments or pipeline designs present in our sample and we have therefore removed these research designs from Table 7. Also, some categories in Table 6 indicate low-medium, but based on the actual scores we calculated for each of our studies we made a clear decision on either low or medium rather than sit on the fence between 2 categories.

Table 7 indicates that in our sample of 100 studies, 43 have a low score indicating a low risk of bias[5], 25 have a medium score indicting a medium risk of bias[6] and 32 studies have a high score meaning

[5] Low risk of bias studies in our sample include RCTs or panel data designs in combination with using advanced econometric techniques for analysis.
[6] Medium risk of bias studies draw on cross-sectional research designs using a mixture of advanced econometric techniques or multivariate statistics as analytical tools.

they have a high risk of bias[7]. Overall, we have 42 RCTs in our sample of which 23 RCTs are in the health sector and 19 RCTs are in the education sector (dominated by Girls' Education Challenge (GEC) RCTs – 15 in total – which are discussed in more depth in report 2).

High risk of bias does not mean that studies do not contribute in significant ways either substantively or methodologically, only that they have shortcomings in how they deal with threats to internal and external validity and thus we should treat their findings with caution. In the systematic review literature there is a debate on whether to include studies with low methodological quality, i.e. high risk of bias, in the synthesis. We feel that lessons can be learnt from high, medium as well as low risk of bias studies. Therefore, we include all studies irrespective of their methodological quality in the evidence gap maps and in the in-depth synthesis but then provide further sub-group analysis by risk of bias levels to investigate whether any patterns can be detected particularly in terms of success and failure of PbR mechanisms (discussed in section 4, details in Appendix 6 and Appendix 7).

Summary of this section

Due to resource constraints, we adapt standard systematic review procedures to review the evidence base on PbR mechanisms. An iterative search process was followed that was initially driven by a snowballing procedure focusing on 5 existing systematic and non-systematic reviews. The database generated as a result was complemented by extensive electronic searches of academic and institutional repositories as well as Google Scholar. This process did not allow us to capture all of the internally generated DFID evidence which required the adoption of a more informal search approach. We refined the key search terms to optimise the search results of the external evidence base. We identified 5294 PbR related documents that were screened by title and abstract as well as by full-text where required. Applying 3 inclusion criteria, the evidence base was reduced to 100 studies for further assessment and synthesis. A quality scoring scheme was employed to assess the methodological quality of the 100 studies in our sample, we find that a third of all studies suffer from high risk of bias.

3. Evidence gap maps

The evidence gap maps as well as the in-depth synthesis are guided by the Measure-Agent-Principal (MAP) framework described by Clist (2016). The principal refers to the donor or funding body who commits to pay for aid on a pre-agreed measure, the agent is the party being paid to deliver results by the principal and the measure is the pre-agreed measure that PbR contracts are based on.

[7] High risk of bias studies in our sample adopt cross-section designs or basic surveys in combination with less rigorous analytical techniques such as multivariate statistics or basic descriptive statistics.

Principal

As principals, donors of aid, we identified the following from the 100 studies we included:

- The World Bank,
- NGOs,
- governments (this includes DFID),
- development financing organisations (e.g. IADB) and
- public-private partnerships (e.g. GAVI).

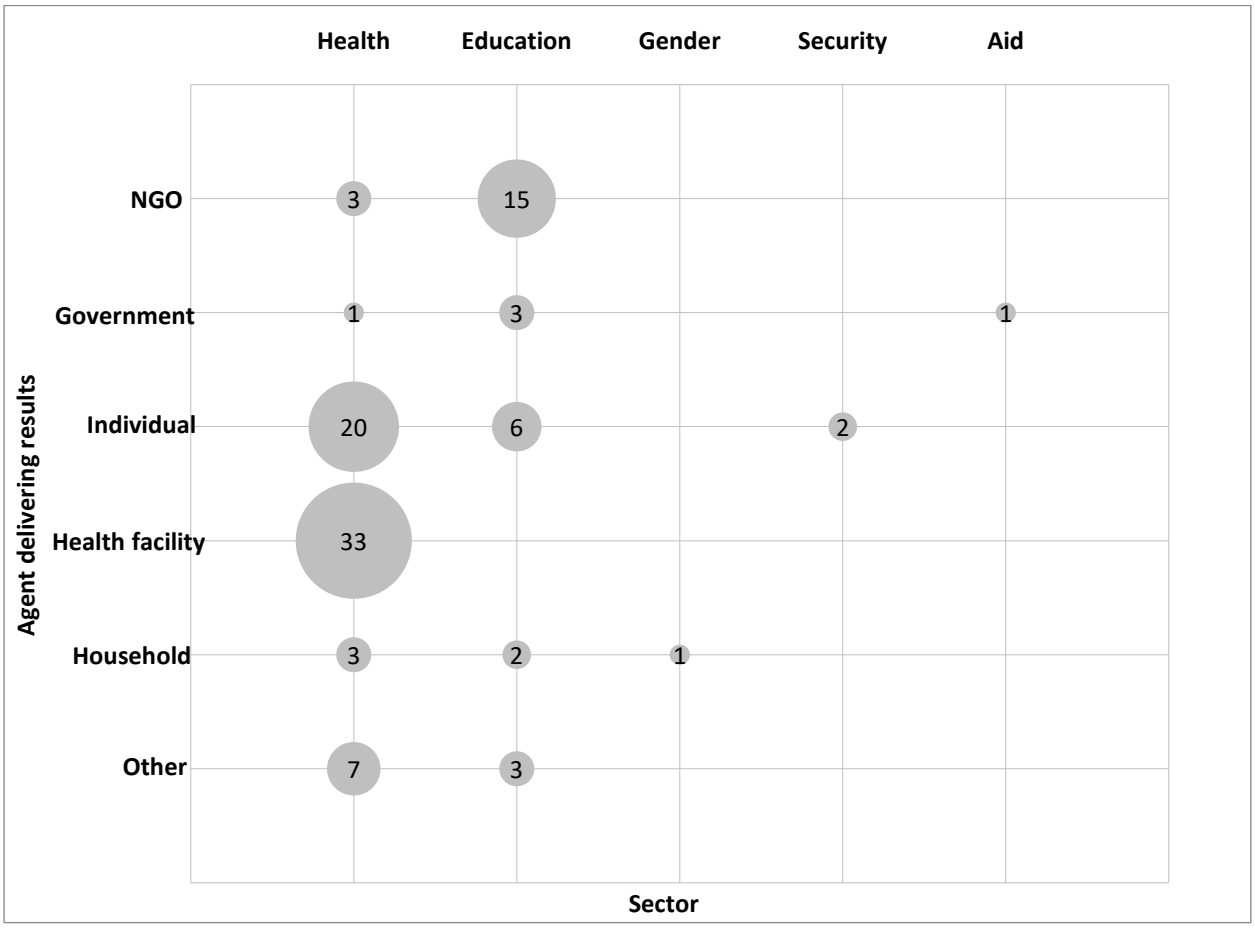
72% of the 100 included studies list governments as the main principal, the remaining 28% were a mix of NGOs (8%), the World Bank (5%), GAVI (2%), development financing organisations (1%) and 12% did not provide any data on the principal (see Appendix 3 for details of these studies). Looking at only the health evidence (accounting for 71 of the 100 studies we included), 63% of studies list governments as the main principal, followed by NGOs (10%) and the World Bank (7%), 15% of health studies did not provide any information on the principal (see Appendix 3 for details) – this lack of reporting is a known challenge in the systematic review context, hence the development of the CONSORT statement[8] to improve the quality of reports of studies (with a focus on RCTs).

Agent

The first evidence gap map (Map 1 below) describes all 100 studies by sectors and by agent, i.e. 'Who gets paid?' across all risk of bias levels to get an understanding of the evidence base we are dealing with.

[8] http://www.consort-statement.org/

Map 1: All studies, all risk of bias levels, by sector and by agent



Notes: Map 1 lists 67 rather than 71 health studies because 9 health studies do not provide any information on 'Who gets paid?' - the agent - while 5 studies report 2 different agents (these appear twice in the map), hence 67 studies. The section 'Other' includes very study specific agents, e.g. villages, cooperatives, central medical stores, schools or school principals, local civic organisations. Appendix 4 has details on each of the studies included in each bubble in the map.

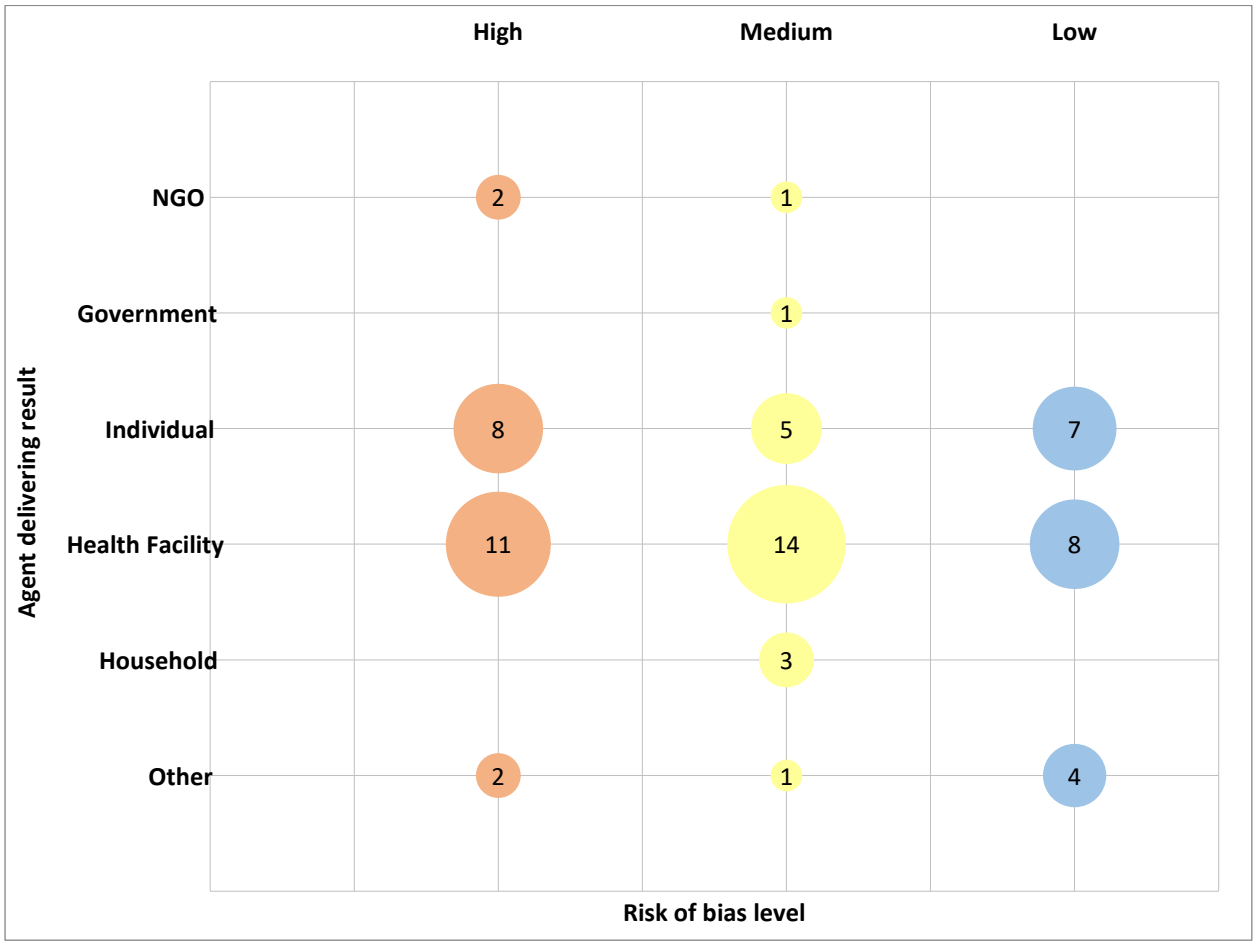
It is clear from Map 1 that health and education are the dominant sectors. The y-axis lists the agent, i.e. 'Who gets paid?', we can see that in more than half of all studies health facility and individual are listed as agents. In total, we find 71 studies for health (71%)[9]. **Given these initial results, we agreed with DFID to focus the synthesis on the health evidence rather than contrast external versus DFID internal evidence.**

Informed by the findings of Map 1 and given concerns related to methodological quality of the included studies (as outlined in Table 7), we created Map 2 describing the health evidence by risk of

[9] Within the sample of 71 health studies, we identified 6 studies as internal DFID documents, the remaining 65 studies were identified through the external search process.

bias levels (high, medium and low) and by agent, i.e. 'Who gets paid?' (NGO, government, individual, health facility, household and other).

Map 2: Health studies by risk of bias levels and by agent



Notes: As in the case of Map 1, 67 rather than 71 health studies are captured because 9 health studies do not provide any information on 'Who gets paid?' - the agent - while 5 studies report 2 different agents (these appear twice in the map), hence 67 studies. The section 'Other' includes very study specific agents, e.g. villages, cooperatives, central medical stores, schools or school principals, local civic organisations. Appendix 4 has details on each of the studies included in each bubble in the map.

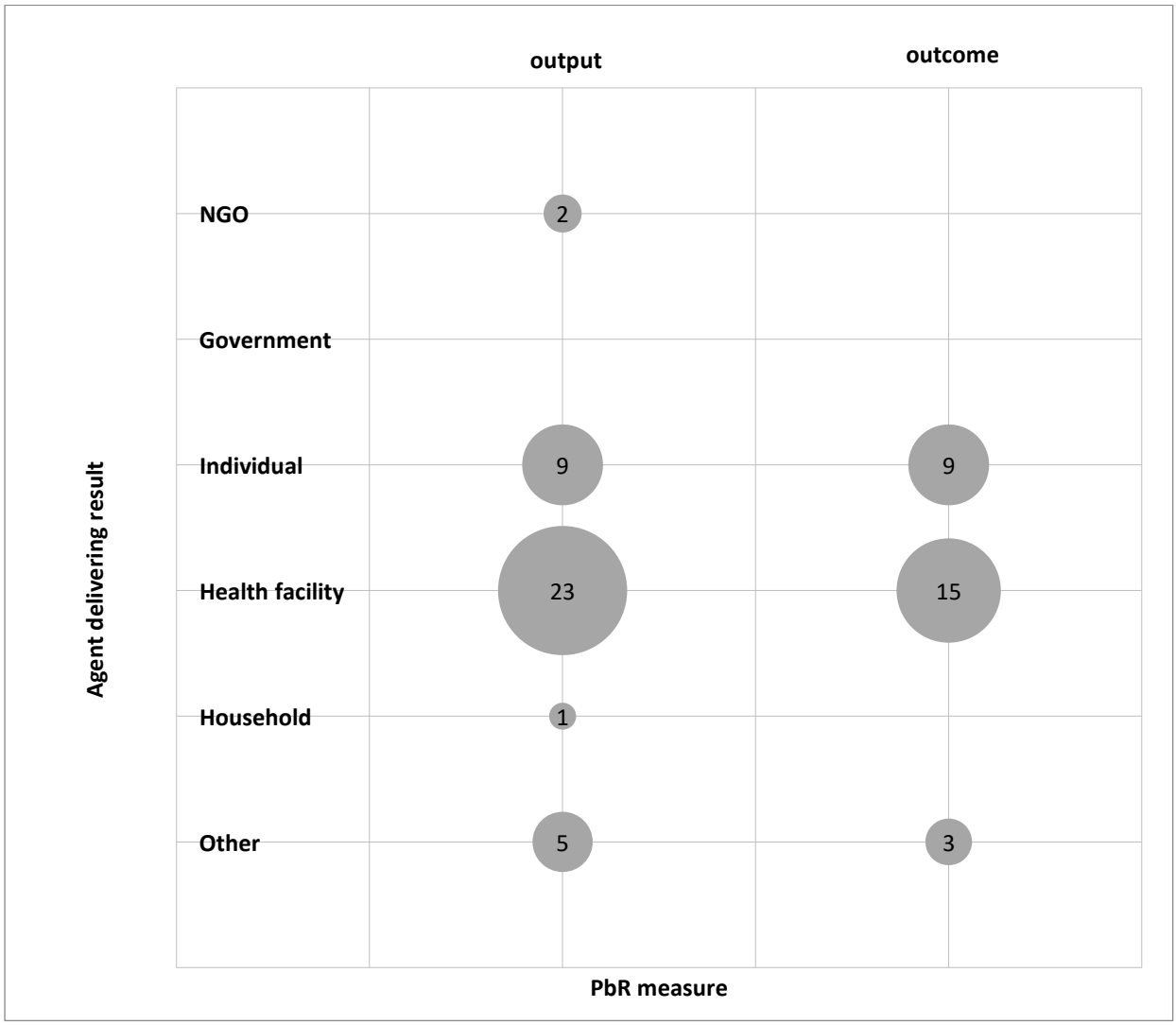
Of the studies captured in Map 2, a third (34%) are classified as suffering from high risk of bias, of the 53 studies listing individual and health facility as an agent 36% (19 studies) are classified as high risk of bias.

We should note that 24% of the 71 health studies have already been synthesised by other systematic and unsystematic reviews that touch on PbR themes (e.g. Blacklock et al, 2016; Eichler et al, 2013; Eldridge and Palmer, 2009; Oxman and Fretheim, 2009; Witter et al, 2012; Glassmann, Todd and Gaarder, 2007; Lindsay, Beith and Eichler, 2011).

Measure

The most important element in the MAP framework is the measure as this forms the basis for the PbR contract, therefore, Map 3 describes the health evidence by PbR measure (categorised into output and outcome measures) and by agent. We define outputs as tangible goods and services that are delivered by the project, e.g. how much money is spent, how many children are vaccinated, how many textbooks are distributed, etc. – the implementing agency has direct control over these outputs. Outcomes build on outputs, they are realised once beneficiaries have used the projects outputs, they relate to the project's intended medium-term goals. Measuring outcomes is important when trying to answer cause-and-effect questions allowing an assessment of the overall difference a project made while outputs will only be able to answer normative questions to assess project activities and whether they achieved short-term targets. Outputs are delivered while outcomes come from expected behaviour changes among project beneficiaries (Gertler, 2011).

Map 3: Health studies by PbR measure and by agent, all risk of bias levels



Notes: As in the case of Map 1, 67 rather than 71 health studies are captured because 9 health studies do not provide any information on 'Who gets paid?' - the agent - while 5 studies report 2 different agents (these appear

twice in the map), hence 67 studies. 15 studies do not report any information on the pre-agreed measure but of the 56 studies reporting pre-agreed measures 32 report 2 or even 3 measures. The section 'Other' includes very study specific agents, e.g. villages, cooperatives, central medical stores, schools or school principals, local civic organisations. Appendix 4 has details on each of the studies included in each bubble in the map.

For the purpose of Map 3, with the definition of outputs and outcomes in mind, we grouped the 26 different PbR measures that we found in our sample of included studies into output and outcome measures[10] to grapple with the high level of heterogeneity we identified. It is clear from the map that the majority of PbR measures are output related (60% of all health studies across all risk of bias levels).

Summary of this section

From the evidence gap maps outlined above it is clear that the majority of the evidence (96 of 100 studies) focuses on exploring PbR mechanisms in health and education, we know little about PbR in other sectors. We also have relatively limited knowledge about outcome level PbR measures across the 71 included health studies as 60% of these studies focus on output level measures. Furthermore, around a third of all studies suffer from high risk of bias.

4. Health synthesis

In this section, we provide a more in-depth discussion of the health evidence which is also guided by the MAP framework. Initially the aim was to employ meta-analysis to synthesize the health evidence, however, meta-analysis is only possible for studies that can be meaningfully compared, i.e. they are comparable on a conceptual level with similar constructs and relationships and similar study designs as well as statistical approaches. In reality, studies often do not share context, intervention, or outcome. Hence, we often encounter the so-called "apples and oranges" problem, which renders meta-analysis meaningless (Lipsey and Wilson, 2001).

Furthermore, methodologically flawed studies of low quality should be treated with caution, recall that a third of the health evidence is classified as suffering from high risk of bias, which implies that there are concerns in relation to how well they dealt with potential threats to internal and external validity. Finally, in the context of the PbR health evidence, we also find high levels of heterogeneity

[10] Applying the definition of outputs and outcomes outlined above and considering the broad results chain of each of the health sub-categories we identified, we categorise the PbR measures as follows: **Outputs** are number of vaccine given, condoms sold, number of people attending health services, number and type of health services delivered, number of new patients, percentage of discharges, number of screening and referring/caring malnutrition cases, length of hospital stay, number and type of STI treatment, number of women receiving PNC/ANC, number of referral of pregnant women to health facility, number of prescriptions given, number of TB treatment/referral provided. **Outcomes** are negative test for disease, giving birth at a facility, birth with skilled attendant, health insurance coverage, contraceptive prevalence rate, cases of malnutrition, investing in health and education expenditure, clinical performance vignettes, provision of family planning services.

(outlined in subsequent tables below) which in light of the arguments presented above suggests that meta-analysis is not sensible. Therefore, we have chosen to use a narrative synthesis approach that involves textual as well as descriptive analyses of the quantitative evidence. Narrative synthesis is a commonly applied method in the social science when meta-analysis is not possible. It brings together the results of a diverse set of studies teasing out similarities and differences while also investigating their outputs and outcomes. The downside of this approach is its lack of being able to provide an effect size but it can still look at effectiveness by addressing the methodological quality of the included studies. It is also common for narrative synthesis to provide a systematic and consistent record of the data extracted from the studies in the review and making this data visually available in the form of tables (Boaz et al., 2002).

A large part of our synthesis will focus on studies listing individual and health facility as agents as these dominate (75%, or 53 studies), we may deviate from this focus where necessary or sensible, some of our analysis will also provide a break-down by risk of bias levels to explore whether methodological quality plays a role in the success or failure of PbR mechanisms (details in Appendix 7).

We begin the synthesis with a description of our sample to highlight the high levels of heterogeneity that prevent us from employing meta-analysis before applying the MAP framework.

Description of the sample of included studies - Heterogeneity

Our sample of health studies is highly heterogeneous in terms of health sub-themes, geographical regions and methodological quality. Table 8 shows that almost 50% of the health evidence is dominated by maternal and child health programmes, this is followed by HIV/STI programmes and nutrition. A third of the health evidence cannot be categorised easily as it is covering a diverse set of health sub-themes.

Table 8: Health sub-themes, across all RoB, individual and health facility only

Health subtheme	**% of studies**
Maternal and child health	47.4
HIV/STI	13.6
Nutrition	5.1
Health general /Other	33.9
Total	**100**

Note: Appendix 5 has details on each study included in this table.

In terms of geographical regions, the diversity of the sample continues to play a role (Table 9), close to 53% of the sample of studies focus on Sub-Saharan Africa, followed by Latin America (13.8%), South and East Asia (both 9.8%) and Southeast Asia (7.8%). Almost 6% of studies focus on Middle Eastern countries. Overall, 21 different countries are captured in our sample of 71 studies. Given PbR

mechanisms are highly context-specific it is challenging to generate meaningful learning when faced with such a highly heterogeneous evidence base.

Table 9: Geographical region, across all RoB, individual and health facility only

Region	**% of studies**
Sub-Sahara Africa	52.9
Latin America	13.8
South Asia	9.8
East Asia	9.8
Southeast Asia	7.8
Middle East	5.9
Total	**100**

Note: Appendix 5 has details on each study included in this table.

We already discussed methodological quality for all studies in the section 'validity and quality appraisal', the quality assessment of the health evidence mirrors the findings presented earlier. Table 10 shows that more than a third of the health evidence suffers from high risk of bias (35.3% - Appendix 5 has details on each study). High risk of bias refers to studies that, for example, adopted basic surveys which do not attempt to construct counterfactual scenarios, there are some cross-section research designs using case-control comparisons but the analytical methods they employed have shortcomings, e.g. multivariate analysis or simple tabulations which do not attempt to address selection bias.

Risk of bias	**% of studies**
High	35.3
Medium	35.3
Low	29.4
Total	**100**

Table 10: Risk of bias, individual and health facility only

Note: Appendix 5 has details on each study included in this table.

The medium risk of bias evidence (35.3% of studies) is dominated by cross-section designs that employed sophisticated analytical methods such as difference-in-differences, instrumental variables or propensity score matching, while the low risk of bias studies (almost 30% of studies) are dominated by RCTs (23 studies) that draw on sophisticated analytical methods (see Tables 6 & 7 for details).

Findings

We now use the MAP framework to further synthesise the health evidence and build on the findings described in the section 'evidence gap maps'. The discussion below considers all studies across low, medium and high risk of bias. We conducted sub-group analysis by risk of bias level to explore whether any patterns can be detected in relation to changes in methodological quality especially in

relation to success and failure of PbR mechanisms (reported in Appendix 7) but our findings hold irrespective of risk of bias levels.

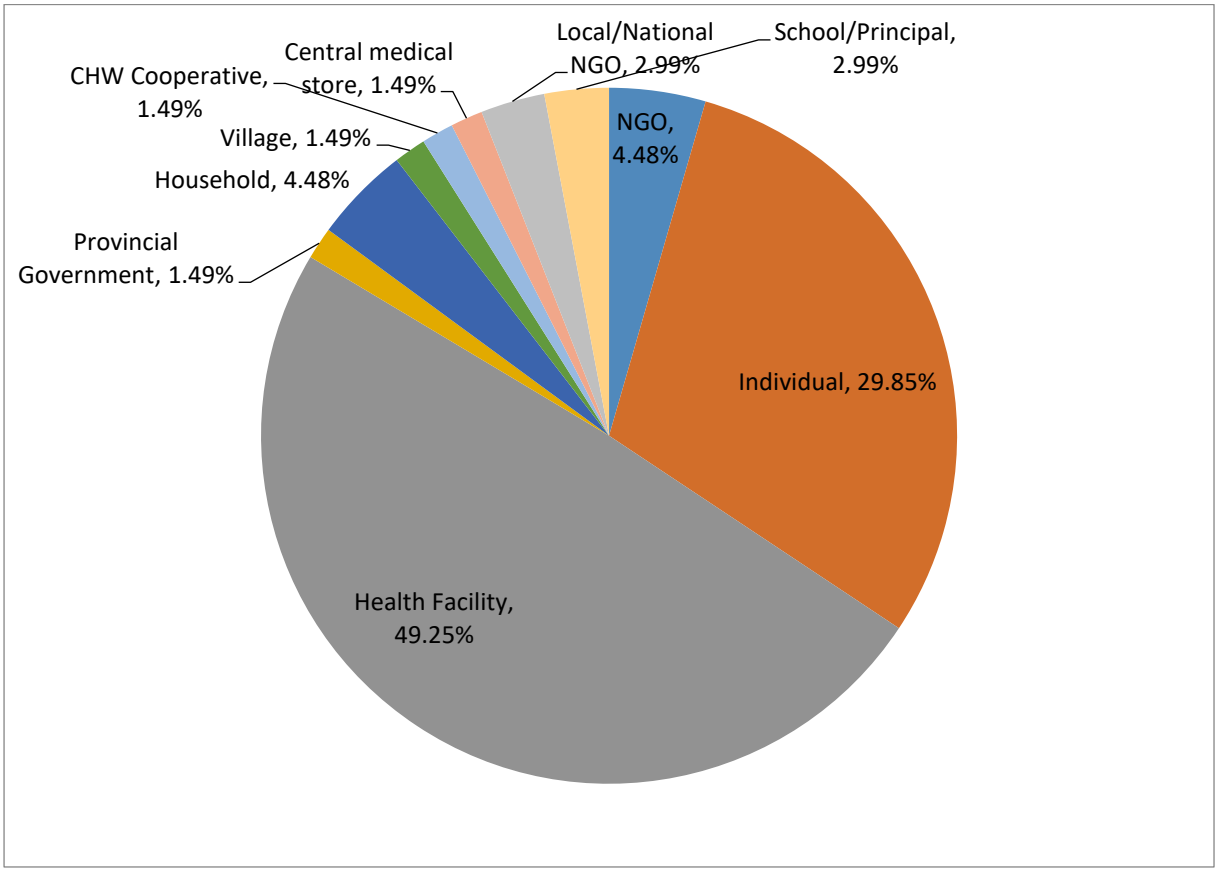
Principal: Who pays?

As mentioned above, 63% of the studies in our health sample report governments as the main principal, followed by NGOs (10%) and the World Bank (7%), 15% do not provide any information on the principal (see Appendix 3 for details). These findings may not be surprising, as governments have historically dominated aid delivery.

Agent: Who gets paid?

The picture is more mixed when examining the domain of the agent. The health evidence is dominated by 2 agents: Health facilities (almost 50% of studies) and individuals (almost 30% of studies), see Figure 1. The remaining 20% of studies present a diverse set of agents such as households (5%), international NGOs (5%), local/national NGOs (3%), schools (3%) and others ranging from provincial governments to villages and central medical stores.

Figure 1: Agent: 'Who gets paid' across all risk of bias levels



Note: Health studies only (chart based on 62 studies rather than 71 as not all studies reported information on 'Who gets paid?', all risk of bias levels. We generated the same chart for low and medium risk of bias studies but this did not substantially alter the findings.

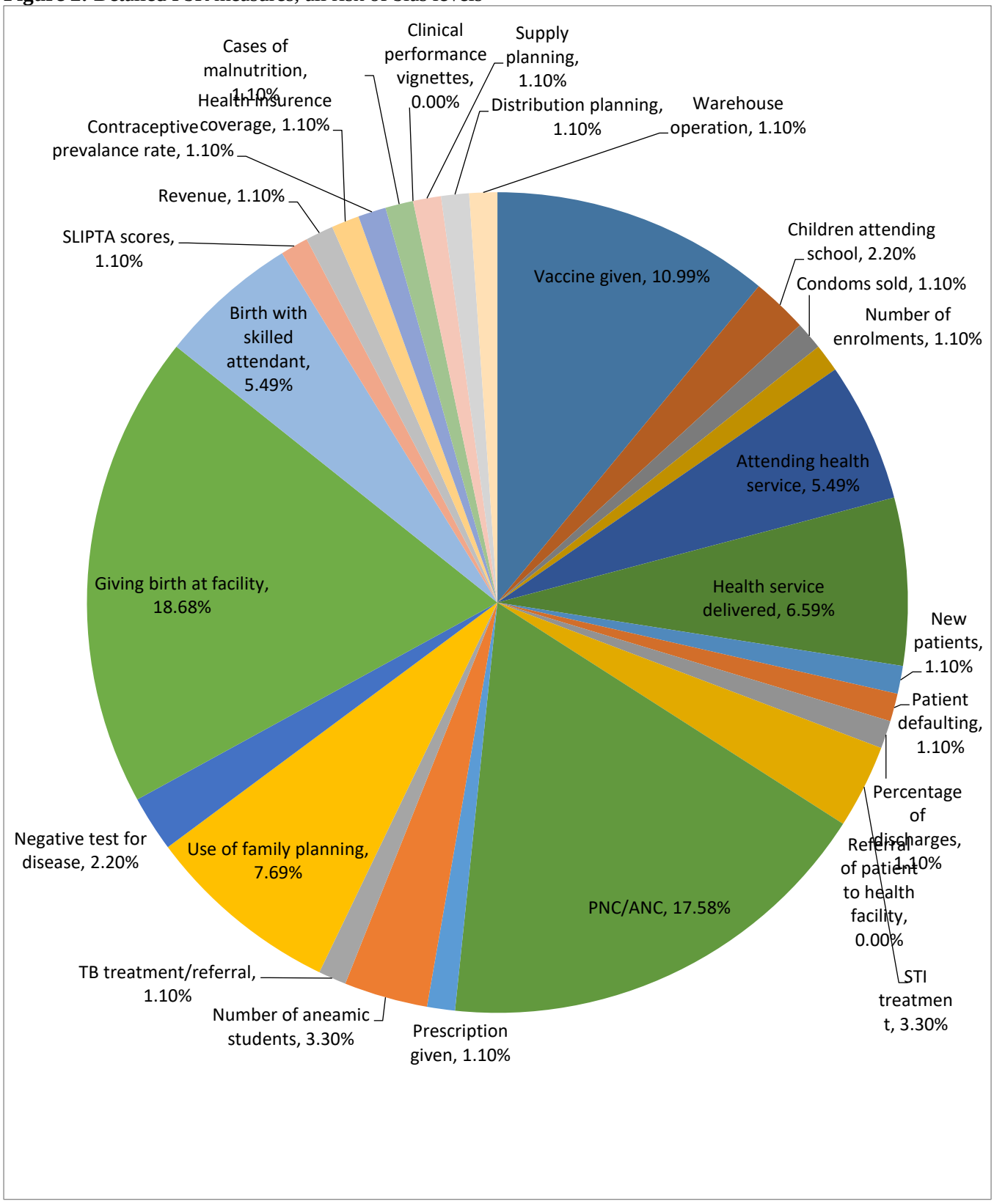
As a result of these findings, our synthesis focuses largely on the studies reporting health facilities and individuals as agents as this captures 80% of the health evidence.

Measure: What is the measure?

The quality and appropriateness of the pre-agreed measure is central to PbR contracts. Much has been written about poor performance measures and the challenges to find a measure that is perfectly aligned with a range of desired activities and incentivising those (Clist and Verschoor, 2014). Clist (2016) further argues that a decision has to be made on 'where "on the results chain" a measure is targeted' (p.13), e.g. a measure could be targeted at outcome or at output levels. Targeting outcomes may be preferred as this is a longer-term measure trying to capture behaviour change (see our earlier definition in section 3) as a result of an intervention but success may not be as easy to assess as measuring outcome levels requires high quality data collected over longer timeframes and this is may not always be feasible or affordable.

As discussed above, we find that 60% of the PbR measures target output levels while the remaining 40% target outcomes, see Map 3. Methodological quality does not alter the key message. We find a wide range of PbR measures across our sample of health measures, 26 in total (see Figure 2).

Figure 2: Detailed PbR measures, all risk of bias levels



Note: Health studies only (chart based on 47 studies rather than 71 as not all studies reported information on 'What is the measure?', but some studies reported multiple measures, all risk of bias levels.

Figure 2 further supports the evidence presented above in terms of high levels of heterogeneity within our sample. To summarise, we identified 71 health studies covering 21 countries, capturing 26 different PbR measures, 9 types of agents plus multiple health sub-themes and considerable variation in methodological quality (a third of studies suffer from high risk of bias). This generates challenges in terms of identifying generalisable findings from this diverse evidence base. In the next section, we discuss the success and failure of these diverse PbR measures to better guide the choice of PbR measures when devising PbR contracts.

What can we learn about success and failure of PbR measures?

Map 3 has clearly indicated that the majority (60%) of the PbR measures identified in the health literature are targeted at output levels. Following Eldridge and Palmer (2009), we now explore what these output but also outcome measures allow us to conclude with regard to success or failure of PbR contracts. We define success and failure in naïve terms by counting[11] the signs and levels of significance for both output and outcome level measures where positive and significant measures would suggest success. Depending on the main mechanism of each programme, success should ideally be defined at the level (i.e. input, output, outcome or impact) appropriate to the particular Theory of Change embodied in the PbR programme. This has often not been possible due to a lack of reporting of Theory of Change and or measures at the appropriate level. To help our explorations we revisited all 71 health studies irrespective of their risk of bias level to extract statistical information on each of the pre-agreed PbR measures reported in each of the studies. We extracted data on whether the results reported for the pre-agreed PbR measure were positive or negative and whether they were statistically significant or not. Table 11 presents the results for all 26 PbR measures across output and outcome levels identified in the health evidence irrespective of their methodological quality.

Table 11: Success and failure of all PbR measures (output and outcome), across all risk of bias levels

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	30	4	34
Not significant	24	3	27
Total	54	7	61

Notes: Health studies only (47 studies rather than 71 as not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures), all risk of bias levels. Appendix 6 has details on each study included in this table.

We can see that the majority of PbR measures report positive results that are statistically significant; this would suggest a success of PbR projects. ~~However, it pays to be careful to arrive at such hasty conclusions. If we breakdown the information presented in Table 11 by output and outcome level~~

[11] We adopt a 'vote counting' procedure, which was first discussed by Light and Smith (1971), where studies are sorted into categories: studies yielding significant or not significant, positive or negative results.

~~measures (see Tables 12 and 13, with more details in Appendix 6) we can see that output measures are more likely to present positive and significant effects than outcome measures.~~

Table 12: Success and failure of all PbR measures (output only), across all risk of bias levels

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	24	4	28
Not significant	23	4	27
Total	47	8	55

Notes: Health studies only (47 studies rather than 71 as not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures), all risk of bias levels. The figures of Tables 12 and 13 do not add up to the ones reported in Table 11 as some studies report both output and outcome measures, if this is the case the study is only counted once in Table 11 but appears in both Tables 12 and 13. Appendix 6 has details on each study included in this table.

The findings presented in Tables 11-13 are further supported by a breakdown by risk of bias levels (see Appendix 7 for details where we provide a sub-group analysis of success and failure of PbR by risk of bias levels). We initially hypothesised that high risk of bias studies may potentially be more likely to report positive and significant effects but this is not the case here as evidenced in Appendix 7. However, many high risk of bias studies do not report levels of significance as they only present basic descriptive statistics or graphics to demonstrate PbR effects.

Table 13: Success and failure of all PbR measures (outcome only), across all risk of bias levels

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	16	1	17
Not significant	14	3	17
Total	30	4	34

Notes: Health studies only (47 studies rather than 71 as not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures), all risk of bias levels. The figures of Tables 12 and 13 do not add up to the ones reported in Table 11 as some studies report both output and outcome measures, if this is the case the study is only counted once in Table 11 but appears in both Tables 12 and 13. Appendix 6 has details on each study included in this table.

We caution to reach any firm conclusions on success or failure of PbR measures based on the findings presented in Tables 11-13. To get more clarity on our findings, we examined the underlying data sources of a random sample of 26 studies. We find that in the case of 7 studies Demographic and Health Survey (DHS) data were used[12], 4 studies collected cross-sectional district level household surveys, in 11 cases national and/or programme health information systems formed the basis for data analysis and in 4 cases World Bank or other government data were used. Using the quality appraisal tool outlined in Table 6, we argue that the data sources we identified suffer from biases, e.g. due to

[12] Schoumaker (2014) raises concerns on some of the birth and fertility metrics found in DHS data, further discussed below.

their cross-sectional nature and lack of constructing counterfactual scenarios as well as due to concerns in relation to metrics used (in the case of DHS data).

Moreover, we should also note that publication bias might play a role in explaining the dominance of positive and significant effects presented in Tables 11-13. Publication bias is a well-recognised and serious issue in the context of systematic reviews. It is argued that studies reporting statistically significant findings are more likely to be published in peer-reviewed journals than studies reporting statistically non-significant findings (Borenstein et al, 2009). In the context of meta-analysis one can test for publication bias but in the context of a narrative synthesis this is more difficult and we can only speculate. In our sample of health studies 59% are published in peer-reviewed journals and of those 90% present positive estimates with 64% of them reporting positive as well as significant findings. This may suggest that publication bias does indeed play a role and thus we should be cautious about what to conclude about success or failure of PbR mechanisms[13].

Understanding failures

To understand success or failure of PbR measures better, we examine some of the success and failures in more depth. Let us start with some of the failures, Tables 12 and 13 list negative PbR estimates on output and outcome levels that are worth investigating. 3 of the studies reporting negative output level PbR effects are suffering from high risk of bias suggesting low methodological quality (Appendix 6 provides details on each of the studies reporting positive as well as negative effects, while Appendix 7 presents a sub-group analysis by risk of bias level as well as details of each study). A range of PbR measures are used such as type of health services delivered, number of referral of patients, and number of anaemic students treated. The principal was government in most cases and the agents largely individuals or health facility. From looking at principal, agent and the type of measure it is not obvious how to make sense of the negative effects we found.

The picture is similar for the outcome level measures reporting negative effects. Though there is a notable difference in terms of PbR measures used, e.g. negative test for disease and giving birth at facility; and the methodological quality which was either low or medium risk of bias. To explain the negative effects we examine the geographical locations of each of the studies and find that almost half of the negative outcomes are reported in fragile states such as Afghanistan, DR Congo, Zimbabwe and Pakistan. This may suggest issues related to data quality and/or measurement errors (as discussed above) but we cannot be certain given the available evidence base.

[13] Publication bias is widespread in published academic research, in particular in the social and medical sciences. Rosenthal (1979) first used the term 'file drawer problem' to draw attention to biases in published research.

Understanding successes

We conclude this section by investigating some of the successes on output and outcome levels in more detail, we focus only on the positive and statistically significant effects.

For 58% of the studies reporting positive and statistically significant output level effects the principal is the government. The agent is mainly health facility (54% of studies) followed by individual, e.g. such as a health worker and/or a patient (21%) – these findings hold when taking out the high risk of bias studies (details in Appendix 6 and Appendix 7). A range of PbR output measures are reported, 11 across 24 studies, such as number of vaccinations given, condoms sold, number of individuals attending health services, number of new patients, percent of patient discharges, type of health service delivered, number of patients for STI treatment, number of women receiving ante- and post-natal care. The dominant measures are number of vaccinations given (reported in 7 studies) and type of health service delivered (reported in 4 studies). One could argue that it may not be surprising that these measures dominate the positive and statistically significant effects as they are easily measurable and changes in these measures can be achieved in short time horizons. However, it is worth asking, especially in relation to our earlier definitions of outputs and outcomes, how meaningful is it to count the number of vaccinations given, e.g. how do we know the right sort of people were vaccinated in terms of correct age and timeliness, do they get the full course of the vaccination, are some people re-vaccinated, etc. (specialist medical literature has engaged with the usefulness of these metrics but it is beyond the scope of this report to review this literature). Furthermore, what do we learn from having an inventory of the type of health services delivered? This may not say much about quality, appropriateness or long-term effects of such health measures. In addition, the recipients of these measures, health facilities and individuals in this case, may not be encouraged to embark on activities that affect or change long-term health behaviour. To gain a better understanding of these issues, it may be worth paying closer attention to the qualitative evidence base and encourage the pursuit of more holistic evaluations that encompass a range of methodological and analytical approaches that would allow us to better unpack the drivers underpinning PbR mechanisms. We do not observe any variations across geographical regions but we should note that 7 of the 11 studies (64%) reporting on vaccinations given and type of health services delivered suffer from high risk of bias.

As for the positive and statistically significant outcome level measures, we find that in 50% of all cases the government is listed as the principal, followed by the World Bank (19%). The dominant agents are again health facility (in 69% of all cases) and individuals (25%) – as above, these findings hold when taking out the high risk of bias studies (Appendix 7 reports details). 7 different PbR outcome measures are reported across 16 studies, e.g. giving birth at facility, provision of family planning services, health insurance coverage, contraceptive prevalence rate and cases of malnutrition. The dominant measures are: giving birth at facility (in 75% of studies), birth with skilled attendant

(25%) and family planning (19%) – some studies report multiple measures hence the total percentages do not add up to 100. No geographical trends can be detected across the positive and significant outcome measures. An interesting finding here is the type of data used to verify the 3 dominant outcomes, in close to 30% of all cases DHS data are used to verify the birth and family planning type outcomes. DHS data are of good quality in most cases but Schoumaker (2014) raises quality concerns for some of the historical birth and fertility data reported in some DHS datasets (we briefly raised this issue in footnote 12 above). Furthermore, DHS data, like so many surveys, is subject to conventional biases such as respondent and authority biases and often no ex-post enumeration verification or quality assessment is conducted in the context of DHS. Clist (2016) raises the issue of being able to draw on unbiased and unincentivised data sources to accurately assess the value of PbR mechanisms – one way to achieve this could be to increasingly involve independent evaluators in the data collection process allowing them to pay attention to data quality as well as to appropriate methods and types of analysis.

Summary of this section

This section describes our sample of health studies, demonstrating its highly heterogeneous nature in terms of health sub-themes, geographical regions and methodological quality. E.g. 71 studies cover 21 countries, 5 types of principals, 9 types of agents, 26 different PbR measures with a third of studies suffering from high risk of bias. Applying the MAP framework, we find that the majority of health studies list governments as the main principal, health facility and individuals as the main agents but without providing any clear trends on dominant PbR measures. We find, however, that 60% of our PbR measures are at the output level.

To reach any conclusions on success or failure of PbR measures we extracted statistical information on all PbR measures (sign and level of significance) irrespective of their risk of bias level. The majority of PbR measures report positive results that are statistically significant suggesting a success of PbR projects (see Table 11). If we break this information down by output and outcome level measures, Table 12 suggests that output and outcome measures present positive and significant effects in quite similar proportions (as indicated by the figures presented in Tables 12 and 13 which count the sign and level of significance of PbR measures separately by outputs and outcomes, with significant positive results found in 44-47% studies of output and outcome measures respectively). We provide examples to better understand success and failure of selected PbR measures and caution the reader not to place too much faith in output level measures (as argued by Clist, 2016). We also point towards a potential lack of unbiased data sources, e.g. see the case of DHS data, that do not allow us to draw any clear conclusions on the success or failure of PbR mechanisms. We note that publication bias may explain why positive and significant effects dominate casting further doubts on the reliability of the findings.

5. Discussion points

In this section, we explore what we can learn from the PbR health evidence in terms of advancing PbR mechanisms. The issues raised in this section draw inferences beyond the evidence base included in this report and relate them to the PbR literature more broadly.

- From evidence gap Map 3 it is clear that 60% of PbR measures reported in the health evidence are output related (across all risk of bias levels). This is not surprising as outputs are easier to measure than outcomes and successes (i.e. positive effects using our definition above) can be realised within shorter time horizons. This observation is linked to the short-term nature of PbR contracts favouring measures that can be verified relatively quickly and more easily at low cost (e.g. see Clist, 2016). However, outputs may not always be a reliable measure of success or failure; thus, programmes should strive to think more carefully about the PbR measures they adopt and how they relate to the mechanisms and goals set out in their particular Theory of Change.

- To understand the failures (using our naïve definition presented in the previous section) of some of the PbR mechanisms, we examined the geographical locations of each of the studies and find that almost half of the failures are reported in fragile states. This may suggest issues related to data quality and/or measurement errors (we have already made these points above) casting doubts on the reliability of the PbR measures used in such fragile environments. Output level measures may be easier to measure and verify in fragile states but this may not always be useful from a PbR perspective as the ultimate goal of sustainable development programmes should be to create long-lasting outcome level effects that are driven by behavioural changes.

- Furthermore, the evidence does not provide any details on the costs of measuring outputs versus outcomes and we do not yet fully understand what sort of PbR measures work best in fragile states. Future studies should provide more information on the costs of collecting and verifying PbR measures so that the trade-offs between the different PbR mechanisms can be better understood, especially in fragile and conflicted-affected environments.

- Report 2 argues that evidence on which agent characteristics are the most important is thin with the exception of health worker motivation. In the case of the health evidence included in this report, it is clear that individual and health facility are the dominant agents but a clear pattern on motivation is not emerging due to the heterogeneous nature of the evidence base. We identified 2 studies dedicated to exploring health worker motivation: Huillery & Sebanz, 2014 and Robyn et al., 2014. Huillery & Sebanz (2014) conduct an experiment in DRC finding that financial incentives led to better efforts from health workers. They also find a shift from intrinsic to extrinsic motivation but caution that the capacity of the health care provider needs to be taken into account if PbR measures are to succeed. Robyn et al (2014) investigate the case of a health

insurance scheme in Burkina Faso where quality of health care declined due to health worker dissatisfaction with the payment method, e.g. levels of capitation payments were insufficient, payments were infrequent and there was no mechanisms in place to reimburse service fees. Some of these findings link to insights gained from the HRITF project discussed in report 2, i.e. lack of motivation affects the quality of the health care provided. Motivational and behavioural aspects of health care delivery within a PbR framework might be worth exploring further.

- Furthermore, around a third of all studies suffer from high risk of bias, this implies that ensuring high methodological quality should be a priority when commissioning future studies on exploring the success of PbR mechanisms – these recommendations align with similar ones made by other full systematic reviews (e.g. see Duvendack et al, 2011; Oya et al, 2017).

- Finally, we wish to draw attention to encouraging more thinking about plausible counterfactual scenarios. Is 'no aid' a plausible counterfactual or is it in fact 'other alternative payment mechanisms'? To truly understand the effects of PbR mechanisms, one has to engage in-depth with the range of counterfactual scenarios that may be on offer. In the impact evaluation literature, counterfactual scenarios are the backbone for rigorously assessing the impact of interventions and so it is surprising to find that only 50% of the studies we included in this review engage with constructing counterfactual scenarios.

6. Conclusion and recommendations

We report the findings of a semi-systematic review to gain a better understanding of what PbR instruments work best in different circumstances, when and how PbR incentives work in practice and what is the value for money for the different types of PbR instruments.

We identified 100 PbR studies of which 71 are focusing on health. Across the health evidence, we find a range of health sub-themes, a wide array of agents and PbR measures (26 different measures with a focus on output measures), 21 different countries and conclude that a third of studies suffer from high risk of bias, i.e. they have low methodological quality.

We assess success and failure of PbR measures and observe that positive and significant effects dominate; however, these successes can be misleading as argued by Clist (2016). Publication bias may also play a role to explain our observations.

Disaggregating our findings for each of the research questions, we conclude and recommend the following:

- What PbR instruments work best in different circumstances?

The evidence we reviewed indicates that most measures (60%) are focused on output level measures (most of them positive) capturing short-term effects. We point towards Clist (2016) and suggest that this could amount to fool's gold and that outcome level measures capturing longer time horizons may be a better choice – assuming the ultimate goal is to seek behaviour changes, which could lead to more sustainable effects of development programmes. One could of course argue that measuring outcomes is not realistic especially in fragile and conflicted affected areas. The evidence base we reviewed indicates that output measures were preferred in fragile environments; or where outcome measures were used within fragile and conflict affected areas these were negative in almost 50% of the cases (see discussion on p.26 and p.29). As argued above, this could be due to measurement errors, low data quality and/or a lack of sufficient research capacity.

- When and how do PbR incentives work in practice?

It is not obvious from the evidence we reviewed that they do or do not work. We also do not know how sustainable PbR measures are as the evidence points towards a preference for short-term measures that focus on the initial stages of the results chain (input to output levels). To unpack how PbR incentives work in practice we should draw more on qualitative in-depth research, the quantitative evidence we reviewed is not sufficient to help us understand this.

- What is the value for money of different types of PbR instruments?

We cannot derive any firm conclusions regarding value for money as the evidence does not report any detail on how payments were calculated or how much was dispersed. When commissioning future PbR studies donors should explicitly request engagement with value for money measures. The report discussing objective 2 provides further details on value for money of PbR mechanisms.

In a nutshell: What have we learnt from this semi-systematic review?

PbR mechanisms are not well understood outside of health and education. In terms of the MAP framework we have learnt the following for each of its components from the studies which we have considered:

Principal: Governments are the dominant principal in 72% of all studies across all sectors and in 63% of the health evidence.

Agent: The dominant agents across the health evidence are health facility and individuals while it is NGOs in the case of the education sector.

Measure: Close to two-thirds of all measures are output level measures (this finding holds for all studies but also for the health only evidence) with high levels of heterogeneity (26 different PbR

measures are reported across health alone). Many of these measures report positive and significant effects.

Final thoughts

To better understand PbR mechanisms, it may be worth examining the existing qualitative evidence in more depth as this would potentially allow us to unpack the black box that conceals PbR mechanisms. We should also be looking at sectors outside of health and beyond developing countries to be able to identify good PbR measures. Furthermore, we need more detailed reporting and investigation on particularly the type of PbR measures that were adopted in different contexts including fragile and conflict affected areas. Finally, we need to be able to draw on unbiased data sources collected by independent evaluators who pay attention to selecting appropriate and rigorous types of methods and analysis depending on context.

7. Bibliography

Blacklock, C., M acPepple, E., Kunutsor, S. & Witter, S., 2016. Paying for Performance to Improve the Delivery and Uptake of Family Planning in Low and Middle Income Countries: A Systematic Review. *Studies in Family Planning*, 47 (4): 309-324.

Boaz, A., Ashby, D. & Young, K., 2002. Systematic Reviews: What Have They Got to Offer Evidence Based Policy and Practice? ESRC UK Centre for Evidence Based Policy and Practice: Working Paper 2. *Available at: https://www.kcl.ac.uk/sspp/departments/politicaleconomy/research/cep/pubs/papers/assets/wp2.pdf.*

Borenstein, M., Hedges, L.V., Higgins J.P.T. & Rothstein, H.R., 2009. *Introduction to Meta-Analysis*, Wiley, Chichester.

Clist, P., 2016. Payments by Results in Development Aid: All that Glitters is not Gold. *The World Bank Research Observer*: 1-24.

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

Coalition for Evidence-Based Policy, 2010. Checklist For Reviewing a Randomized Controlled Trial of a Social Program or Project, To Assess Whether It Produced Valid Evidence. *Available at: http://coalition4evidence.org/wordpress/wp-content/uploads/Checklist-For-Reviewing-a-RCT-Jan10.pdf.*

Duvendack, M. et al., 2011. What is the Evidence of the Impact of Microfinance on the Well-being of Poor People? *EPPICentre Social Science Research Unit Institute of Education University of London*.

Duvendack, M., Hombrados, J., Palmer-Jones, R. & Waddington, H., 2012. Assessing 'What Works' in International Development: Meta-Analysis for Sophisticated Dummies. *Journal of Development Effectiveness*, 4 (3):456-471.

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 Popul Nutrition*, 31 (4 Suppl 2): S36-S47.

Eijkenaar, 2012. Pay for Performance in Health Care An International Overview of Initiatives. *Medical Care Research and Review*, 69 (3): 251-276.

Eldridge, C. & Palmer, N., 2009. Performance-based Payment: Some Reflections on the Discourse, Evidence and Unanswered Questions. *Health Policy Plan*, 24 (3): 160-166.

EppiCentre, 2010. Quality and relevance appraisal. *Available at: http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=177.*

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B. & Vermeersch, C. M. J., 2011. *Impact Evaluation in Practice*. Washington D.C.: World Bank.

Glassman, A., Todd, J. & Gaarder, M., 2007. Performance-Based Incentives for Health: Conditional Cash Transfer Programs in Latin America and the Caribbean. CGD Working Paper No 120. *Available at: https://www.cgdev.org/publication/performance-based-incentives-health-conditional-cash-transfer-programs-latin-america-and.*

Gough, D. 2007. Weight of Evidence: A Framework for the Appraisal of the Quality and Relevance of Evidence. In Furlong, J. & Oancea, A., eds. *Applied and Practice-based Research. Special Edition of Research Papers in Education*, 22 (2): 213-228.

Gough, D., Oliver, S. & Thomas, J., 2013. Learning from Research: Systematic Review for Informing Policy Decisions. A Quick Guide. A Paper for the Alliance for Useful Evidence. London: Nesta.

Higgins, J.P.T., & Green S., 2011. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. *Available at: www.cochrane.handbook.org.*

Light, R.J. & Smith, P.V., 1971. Accumulating Evidence: Procedures for Resolving Contradictions among Different Research Studies. *Harvard Educational Review*, 41:429-471.

Lindsay, M., Beith, A. & Eichler, R., 2011. Performance-Based Incentives for Maternal Health: Taking Stock of Current Programs and Future Potentials. USAID Health Systems 20/20 Report. *Available at: http://abtassociates.com/AbtAssociates/files/96/96b67395-90b9-4bfe-bbcf-b808fdf0a2f3.pdf.*

Lipsey, M. W. & Wilson, D. B., 2001. *Practical Meta-Analysis*. Applied Social Research Methods, Sage Publications.

Mallet, R., Hagen-Zanker, J., Slater, R. & Duvendack, M., 2012. The Benefits and Challenges of Using Systematic Reviews in International Development Research. *Journal of Development Effectiveness*, 4 (3): 445-455.

Mason, P., Fullwood, Y., Singh, K. & Battye, F., 2015. Payment by Results Learning from the Literature. ICF International. *Available at: https://www.nao.org.uk/wp-content/uploads/2015/06/Payment-by-Results-Learning-from-the-Literature.pdf.*

Oxman, A. D. & Fretheim, A., 2009. Can Paying for Results Help to Achieve the Millennium Development Goals? Overview of the Effectiveness of Results-based Financing. *Journal of Evidence Based Medicine*, 2 (2): 70-83.

Oya C, Schaefer F, Skalidou D, McCosker C, Langer L., 2017. Effects of Certification Schemes for Agricultural Production on Socio-Economic Outcomes in Low- and Middle-Income Countries: A Systematic Review, Campbell Systematic Reviews 2017:3.

Perrin, B., 2013. Evaluation of Payment by Results (PBR): Current Approaches, Future Needs, DFID Working Paper No 39. *Available at: http://www.dev-practitioners.eu/fileadmin/Redaktion/Documents/TG_RBA/payment-results-current-approaches-future-needs.pdf.pdf*.

Rosenthal, R., 1979. File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin*, 86: 638–641.

Schoumaker, B., 2014. Quality and Consistency of DHS Fertility Estimates, 1990 to 2012. DHS Methodological Reports No. 12. Rockville, Maryland, USA: ICF International.

Valentine, J. and Cooper, H., 2008. A Systematic and Transparent Approach for Assessing the Methodological Quality of Intervention Effectiveness Research: The Study Design and Implementation Assessment Device. *Psychological Methods*, 13 (2): 130-149.

Waddington, H., White, H., Snilstveit, B., Hombrados, J.G., Vojtkova et al. 2012. How to do a Good Systematic Review of Effects in International Development: A Tool Kit. *Journal of Development Effectiveness*, 4 (3):359-387.

Webster, R., 2016. *Payment by Results: Lessons from the Literature*. Russell Webster. *Available at: http://russellwebster.com/Lessons%20from%20the%20Payment%20by%20Results%20literature%20Russell%20Webster%202016.pdf*.

Witter, S., Fretheim, A., Kessy, F. & Lindahl, A., 2012. Paying for Performance to Improve the Delivery of Health Interventions in Low- and Middle-Income Countries. Cochrane Database of Systematic Reviews, 2. *Available at: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD007899.pub2/full*.

Detailed bibliography for all 100 included studies

Health sector

1. Alonge, O., Gupta, S., Engineer, C., Salehi, A. & Peters, D. 2015. Assessing the pro-poor effect of different contracting schemes for health services on health facilities in rural Afghanistan. *Health Policy and Planning,* 30 (10)**:** 1229-1242.

2. Ashraf, N., Bandiera, O. & Jack, K., 2013, No margin, no mission? Evaluating the role of incentives in the distribution of public goods in Zambia, 3ie Impact Evaluation Report 9. *Available: http://www.3ieimpact.org/media/filer_public/2014/02/20/ie_9-ashraf-no_margin_no_mission_web.pdf.*

3. Barham, T. 2011. A healthier start: The effect of conditional cash transfers on neonatal and infant mortality in rural Mexico. *Journal of Development Economics,* 94 (1)**:** 74–85.

4. Basinga, P., Gertler, P. J., Binagwaho, A., Soucat, A. L. B., Sturdy, J. & Vermeersch, C. M. J. 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.

5. Binagwaho, A., Condo, J., Wagner, C., Ngabo, F., Karema, C., Kanters, S., Forrest, J. I. & Bizimana, J. d. D. 2014. Impact of implementing performance-based financing on childhood malnutrition in Rwanda. *BMC Public Health,* 14 (1132).

6. Bonfrer, I., Soeters, R., Van de Poel, E., Basenya, O., Longin, G., van de Looij, F. & van Doorslaer, E. 2014a. Introduction Of Performance Based Financing In Burundi Was Associated With Improvements In Care And Quality. *Health Affairs,* 33 (12)**:** 2179-2187.

7. Bonfrer, I., Van de Poel, E. & Van Doorslaer, E. 2014b. The effects of performance incentives on the utilization and quality of maternal and child care in Burundi. *Social Science & Medicine,* 123**:** 96-104.

8. Bossuroy, T., Delavallade, C. & Pons, V., 2016, Fighting tuberculosis through community based counsellors: a randomized evaluation of performance based incentives in India, 3ie Grantee Final Report. *Available: http://www.3ieimpact.org/media/filer_public/2016/07/28/gfr-ow31218-tb-health-worker.pdf.*

9. Bowser, D. M., Ramon Figueroab, Laila Natiqc & Okunogbec, A. 2013. A preliminary assessment of financial stability, efficiency, health systems and health outcomes using performance-based contracts in Belize. *Global Public Health,* 8 (9)**:** 1063-1074.

10. Celhay, P., Gertler, P., Giovagnoli, P. & Vermeersch, C., 2015, Long-Run Effects of Temporary Incentives on Medical Care Productivity, World Bank Policy Research Working Paper 7348. *Available: https://openknowledge.worldbank.org/bitstream/handle/10986/22228/Long0run0effec0al0care0productivity.pdf?sequence=1&isAllowed=y.*

11. Chansa, C., Das, A., Qamruddin, J., Friedman, J., Mkandawire, A. & Vledder, M., 2015, Linking Results to Performance : Evidence from a Results Based Financing Pre-Pilot Project in Katete District, Zambia, World Bank Health, Nutrition, and Population (HNP) discussion paper. *Available: https://openknowledge.worldbank.org/bitstream/handle/10986/22390/Linking0result0ete0District00Zambia.pdf?sequence=1&isAllowed=y.*

12. Cornejo-Ovalle, M., Brignardello-Petersen, R. & Pérez, G. 2015. Pay-for-performance and efficiency in primary oral health care practices in Chile. *Revista Clínica de Periodoncia, Implantología y Rehabilitación Oral,* 8 (1)**:** 60-66.

13. de Walque, D., Dow, W. & Nathan, R., 2014, Rewarding Safer Sex: CondlitionalI Cash Transfers for HIV/STI Prevention, World Bank Policy Research Working Paper 7099. *Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2522733.*

14. de Walque, D., Gertler, P. J., Bautista-Arredondo, S., Kwan, A., Vermeersch, C., de Dieu Bizimana, J., Binagwaho, A. & Condo, J. 2015. Using provider performance incentives to increase HIV testing and counseling services in Rwanda. *Journal of Health Economics,* 40**:** 1-9.

15. Eichler, R., Auxila, P., Antoine, U. & Desmangles, B., 2007, Performance-Based Incentives for Health: Six Years of Results from Supply-Side Programs in Haiti, Centre for Global Development Working Paper 121. *Available: http://www.cgdev.org/sites/default/files/13543_file_Haiti_Incentives.pdf.*

16. Engineer, C., Dale, E., Agarwal, A., Agarwal, A., Alonge, O., Edward, A., Gupta, S., Schuh, H., Burnham, G. & Peters, D. 2016. Effectiveness of a pay-for-performance intervention to improve maternal and child health services in Afghanistan: a cluster-randomized trial. *International Journal of Epidemiology,* 45 (2)**:** 451-459.

17. Falisse, J.-B., Ndayishimiye, J., Kamenyero, V. & Bossuyt, M. 2015. Performance-based financing in the context of selective free health-care: an evaluation of its effects on the use of primary health-care services in Burundi using routine data. *Health Policy and Planning,* 30 (10)**:** 1251-1260.

18. Fox, S., Witter, S., Wylde, E., Mafuta, E. & Lievens, T. 2014. Paying health workers for performance in a fragmented, fragile state: reflections from Katanga Province, Democratic Republic of Congo. *Health Policy and Planning,* 29 (1)**:** 96-105.

19. Garcia Prado, A. & Lao Peña, C., 2010, Contracting and Providing Basic Health Care Services in Honduras : A Comparison of Traditional and Alternative Service Delivery Models, World Bank Health, Nutrition and Population (HNP) discussion paper. *Available: https://openknowledge.worldbank.org/bitstream/handle/10986/13609/560080WP0Box341ContractingProviding.pdf?sequence=1&isAllowed=y.*

20. Gertler, P., 2000, The impact of PROGRESA on health, International Food Policy Institute Final Report. *Available: http://www.ifpri.org/publication/impact-progresa-health.*

21. Gertler, P., Giovagnoli, P. & Martinez, S., 2014, Rewarding Provider Performance to Enable a Healthy Start to Life: Evidence from Argentina's Plan Nacer, World Bank Policy Research Working Paper 6884 *Available: https://elibrary.worldbank.org/doi/pdf/10.1596/1813-9450-6884.*

22. Gertler, P. & Vermeersch, C., 2012, Using Performance Incentives to Improve Health Outcomes, World Bank Policy Research Working Paper 6100. *Available: https://openknowledge.worldbank.org/handle/10986/9316.*

23. Gopalan, S. S. & Varatharajan, D. 2012. Addressing maternal healthcare through demand side financial incentives: experience of Janani Suraksha Yojana program in India. *BMC Health Services Research,* 12 (319).

24. Huillery, E. & Sebanz, J., 2014, Performance Based Financing, Motivation and Final Output in the Health Sector: Experimental Evidence from the Democratic Republic of Congo. *Available: http://spire.sciencespo.fr/hdl:/2441/4pmvo3bm7m9claao2gl0337ip4/resources/2014-12.pdf.*

25. Ir, P., Korachais, C., Chheng, K., Horemans, D., Van Damme, W. & Meessen, B. 2015. Boosting facility deliveries with results-based financing: a mixed-methods evaluation of the government midwifery incentive scheme in Cambodia. *BMC Pregnancy and Childbirth,* 15 (170).

26. Janisch, C., Albrecht, M., Wolfschuetz, A., Kundu, F. & Klein, S. 2010. Vouchers for health: A demand side output-based aid approach to reproductive health services in Kenya. *Global Public Health,* 5 (6)**:** 578-594.

27. Janssen, W., Ngirabega, J., Matungwa, M. & Van Bastelaere, S. 2015. Improving quality through performance-based financing in district hospitals in Rwanda between 2006 and 2010: A 5-year experience. *Tropical Doctor,* 45 (1)**:** 27-35.

28. Kohler, H.-P. & Thornton, R. L. 2012. Conditional Cash Transfers and HIV/AIDS Prevention : Unconditionally Promising? *The World Bank Economic Review,* 26 (2)**:** 165-190.

29. Kumar, M., Lehmann, J., Rucogoza, A., Kayobotsi, C., Das, A. & Schneidman, M., 2016, East Africa Public Health Laboratory Networking Project : Evaluation of Performance-Based Financing for Public Health Laboratories in Rwanda, World Bank Health, Nutrition and Population (HNP) discussion paper. *Available: https://openknowledge.worldbank.org/bitstream/handle/10986/24400/East0Africa0000boratories0in0Rwanda.pdf?sequence=1&isAllowed=y.*

30. Lannes, L. 2015. Improving health worker performance: The patient-perspective from a PBF program in Rwanda. *Social Science & Medicine,* 138**:** 1-11.

31. Lannes, L., B. Meessen, A. Soucat & Basinga, P. 2015. Can performance-based financing help reaching the poor with maternal and child health services? The experience of rural Rwanda. *The International Journal of Health Planning and Management,* 31**:** 309-348.

32. Leroy, J. L., Armando Garcia-Guerra, Raquel Garcia, Clara Dominguez, Juan Rivera & Neufeld, L. M. 2008. The Oportunidades Program Increases the Linear Growth of Children Enrolled at Young Ages in Urban Mexico. *Journal of Nutrition,* 138 (4)**:** 793-798.

33. Lim, S. S., Lalit Dandona, Joseph A Hoisington, Spencer L James, Margaret C Hogan & Gakidou, E. 2010. India's Janani Suraksha Yojana, a conditional cash transfer programme to increase births in health facilities: an impact evaluation. *The Lancet,* 375 (9730)**:** 2009-2023.

34. Liu, X. & Mills, A. 2005. The effect of performance-related pay of hospital doctors on hospital behaviour: a case study from Shandong, China. *Human Resources for Health,* 3 (11).

35. Mac Arthur, I., Nelson, J. & Woodye , M., 2014, Quality improvement of health care in Belize: focusing on results, Inter-American Development Bank Technical Note 661. *Available: https://publications.iadb.org/bitstream/handle/11319/6468/Quality%20Improvement%20of%20Health%20Care%20in%20Belize%20_%20Focusing%20on%20Results.pdf?sequence=1.*

36. Matsuoka, S., Obara, H., Nagai, M., Murakami, H. & Lon, R. 2014. Performance-based financing with GAVI health system strengthening funding in rural Cambodia: a brief assessment of the impact. *Health Policy and Planning,* 29 (4)**:** 456-465.

37. Moyo, I., Gandidzanwa, C., Tsikira, T., Mabhena, T., Dieleman, M. & Kane, S., 2015, Process Monitoring and Evaluation II of Zimbabwe's Results-Based Financing Project: The Case of Mutoko, Chiredzi, Nkayi and Kariba Districts, DFID *Available.*

38. na, 2016, Rewarding Provider Performance to Improve Quality and Coverage of Maternal and Child Health Outcomes - Evidence to Inform Policy and Management Decisions, DFID Report. *Available.*

39. na, na, Rwanda Community Performance-Based Financing Impact Evaluation, DFID *Available.*

40. Nahimana, E., Iyer, H., Manzi, A., Uwingabiye, A., Gupta, N., Uwilingiyemungu, N., Drobac, P. & Hirschhorn, L. 2015. The race to the top initiative: towards excellence in health-care service delivery. *Global Health Action,* 9.

41. Obare, F., Okwero, P., Villegas, L., Mills, S. & Bellows, B., 2016, Increased Coverage of Maternal Health Services among the Poor in Western Uganda in an Output-Based Aid Voucher Scheme, World Bank Policy Research Working Paper 7709. *Available: https://openknowledge.worldbank.org/bitstream/handle/10986/24626/Increased0cove0d0aid0voucher0scheme.pdf?sequence=1&isAllowed=y.*

42. Olken, B. A., Onishi, J. & Wong, S. 2014. Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia. *American Economic Journal: Applied Economics,* 6 (4)**:** 1-34.

43. Peabody, J. W., Riti Shimkhada, Stella Quimbo, Orville Solon, Xylee Javier & McCulloch, C. 2014. The impact of performance incentives on child health outcomes: results from a cluster randomized controlled trial in the Philippines. *Health Policy and Planning,* 29 (5)**:** 615-621.

44. Powell-Jackson, T. & Hanson, K. 2012. Financial incentives for maternal health: Impact of a national programme in Nepal. *Journal of Health Economics,* 31 (1)**:** 271-284.

45. Powell-Jackson, T., Sumit Mazumdar & Mills, A. 2015. Financial incentives in health: New evidence from India's Janani Suraksha Yojana. *Journal of Health Economics,* 43**:** 154-169.

46. Regalía, F. & Castro, L., 2007, Performance-based Incentives for Health: Demand- and Supply-Side Incentives in the Nicaraguan Red de Protección Social, Center for Global Development Working Paper 119. *Available: http://www.cgdev.org/sites/default/files/13541_file_Nicaragua_final.pdf.*

47. Renaud, A. & Semasaka, J.-P., 2014, Verification of Performance in Results-Based Financing : The Case of Community and Demand-Side RBF in Rwanda, World Bank Health, Nutrition and Population (HNP) Discussion paper. *Available: https://openknowledge.worldbank.org/bitstream/handle/10986/20791/917720WP0Verif00Box385343B00PUBLIC0.pdf?sequence=1&isAllowed=y.*

48. Robyn, P. J., Bärnighausen, T., Souares, A., Traoré, A., Bicaba, B., Sié, A. & Sauerborn, R. 2014. Provider payment methods and health worker motivation in community-based health insurance: A mixed-methods study. *Social Science & Medicine,* 108**:** 223-236.

49. Rusa, L., Ngirabega, J., Janssen, W., Van Bastelaere, S., Porignon, D. & Vandenbulcke, W. 2009a. Performance-based financing for better quality of services in Rwandan health centres: 3-year experience. *Tropical Medicine & International Health,* 14 (7)**:** 830-837.

50. Rusa, L., Schneidman, M., Fritsche, G. & Musango, L., 2009b, Rwanda: Performance-Based Financing in the Public Sector, Center for Global Development Case Study. *Available: http://www.nvag.nl/afbeeldingen/Bibliotheek/Rwanda/Rwanda%20PBF.pdf.*

51. Sherry, T. B., Sebastian Bauhoff & Mohanan, M. 2017. Multitasking and Heterogeneous Treatment Effects in Pay-for-Performance in Health Care: Evidence from Rwanda. *American Journal of Health Economics,* 3 (2)**:** 196-226.

52. Skiles, M. P., Curtis, S. L., Basinga, P. & Angeles, G. 2012. An equity analysis of performance-based financing in Rwanda: are services reaching the poorest women? *Health Policy and Planning,* 28 (8)**:** 825-837.

53. Skiles, M. P., S.L. Curtis, P. Basinga, G. Angeles & Thirumurthy, H. 2015. The effect of performance-based financing on illness, care-seeking and treatment among children: an impact evaluation in Rwanda. *BMC Health Services Research,* 15 (375).

54. Soares, F. V., Ribas, R. P. & Hirata, G. I., 2008, Achievements and Shortfalls of Conditional Cash Transfers: Impact Evaluation of Paraguay's Tekopora Programme, UNDP - International Poverty Centre Evaluation Note Number 3. *Available: http://www.ipc-undp.org/pub/IPCEvaluationNote3.pdf.*

55. Soeters, R., Habineza, C., Peerenboom, P. & Rietsema, A. 2007. Performance-based financing and changing the district health system: experience from Rwanda. *Tropical Medicine & International Health,* 12.

56. Soeters, R., Peerenboom, P., Mushagalusa, P. & Kimanuka, C. 2011. Performance-Based Financing Experiment Improved Health Care In The Democratic Republic Of Congo. *Health Affairs,* 30 (8)**:** 1518-1527.

57. Sood, N., Bendavid, E., Mukherji, A., Wagner, Z., Nagpal, S. & Mullen, P. 2014. Government health insurance for people below poverty line in India: quasi-experimental evaluation of insurance and health outcomes. *BMJ,* 349 (g5114)**:** 1-13.

58. Spisak, C., Morgan, L., Eichler, R., Rosen, J., Serumaga, B. & Wang, A. 2016. Results-Based Financing in Mozambique's Central Medical Store: A Review After 1 Year. *Global Health Science and Practice,* 4 (1)**:** 165-177.

59. Sun, X., Xiaoyun Liu, Qiang Sun, Winnie Yip, Adam Wagstaff & Meng, Q. 2016. The impact of a pay-for-performance scheme on prescription quality in rural China. *Health Economics,* 25 (6)**:** 706-722.

60. Sylvia, S., Luo, R., Zhang, L., Shi, Y., Medina, A. & Rozelle, S. 2013. Do you get what you pay for with school-based health programs? Evidence from a child nutrition experiment in rural China. *Economics of Education Review,* 37**:** 1-12.

61. Urquieta, J., Angeles, G., Mroz, T., Lamadrid-Figueroa, H. & Hernández, B. 2009. Impact of Oportunidades on Skilled Attendance at Delivery in Rural Areas. *Economic Development and Cultural Change,* 57 (3)**:** 539-558.

62. Valadez, J., Jeffery, C., Brant, T., Vargas, W. & Pagano, M., 2015, Final impact assessment of the results-based financing programme for Northern Uganda, DFID Final Report. *Available.*

63. Van de Poel, E., Flores, G., Ir, P. & O'Donnell, O. 2016. Impact of Performance-based financing in a low-resource setting: a decade of experience in Cambodia. *Health Economics,* 25 (6)**:** 688-705.

64. Wei, X., Zou, G., Yin, J., Walley, J., Yang, H., Kliner, M. & Mei, J. 2012. Providing financial incentives to rural-to-urban tuberculosis migrants in Shanghai: an intervention study. *Infectious Diseases of Poverty,* 1 (9).

65. Witter, S., Zaman, R., Scott, M. & Misty, R., 2016, Delivering Reproductive Health Results Through Non-State Providers in Pakistan, DFID Impact Evaluation Report. *Available.*

66. World Bank, 2013, Turkey - Performance Based Contracting Scheme in Family Medicine : Design and Achievements, World Bank Report. *Available:*

https://openknowledge.worldbank.org/bitstream/handle/10986/16532/770290Revised0box377292B00PUBLIC00.pdf?sequence=1&isAllowed=y.

67. Yao, H., Wei, X., Liu, J., Zhao, J., Hu, D. & Walley, J. 2008. Evaluating the effects of providing financial incentives to tuberculosis patients and health providers in China. *The International Journal of Tuberculosis and Lung Disease,* 12 (10)**:** 1166-1172.

68. Yip, W., Powell-Jackson, T., Wen Chen, Min Hu, Eduardo Fe, Mu Hu, Weiyan Jian, Ming Lu, Wei Han & Hsiao, W. C. 2014. Capitation Combined With Pay-For-Performance Improves Antibiotic Prescribing Practices In Rural China. *Health Affairs,* 33 (3)**:** 502-510.

69. Zeng, W., Cros, M., Wright, K. & Shepard, D. 2013. Impact of performance-based financing on primary health care services in Haiti. *Health Policy and Planning,* 28 (6)**:** 596-605.

70. Zeng, W., Rwiyereka, A., Amico, P., Avila-Figueroa, C. & Shepard, D. 2014. Efficiency of HIV/AIDS Health Centers and Effect of Community-Based Health Insurance and Performance-Based Financing on HIV/AIDS Service Delivery in Rwanda. *American Journal of Tropical Medicine and Hygiene,* 90 (4)**:** 740-746.

71. Zhang, L., Rozelle, S. & Shi, Y., 2013, Paying for performance in China's battle against anaemia, 3ie Impact Evaluation Report 8. *Available: http://www.3ieimpact.org/media/filer_public/2014/02/20/ie_8-zhang-china_anaemia_final.pdf.*

All other sectors

72. Angelucci, M. & Attanasio, O. 2009. Oportunidades: Program Effect on Consumption, Low Participation, and Methodological Issues. *Economic Development and Cultural Change,* 57 (3)**:** 479-506.

73. Bagyendera, J., Asiimwe, J., Twinamatsiko, A. & Gumisiriza, D., 2016, "Keeping Marginalised Girls in School by Economically Empowering their Parents" Project, GEC Midline Report. *Available.*

74. Barrera-Osorio, F. & Raju, D., 2015, Teacher Performance Pay : Experimental Evidence from Pakistan, World Bank Policy Research Working Paper 7307. *Available: https://openknowledge.worldbank.org/bitstream/handle/10986/22180/Teacher0perfor0idence0from0Pakistan.pdf?sequence=1&isAllowed=y.*

75. Cambridge Education, 2015, Evaluation of the pilot project results-based aid in the education sector in Ethiopia, *Available.*

76. Camfed International, 2016, A new “equilibrium” for girls, GEC Midline Report. *Available.*

77. Clist, P., Whitty, B., Holden, J., Abbot, P., Latimer, K., Reid, K. & Boyd, C., 2015, Evaluation of results based aid in Rwandan education, *Available.*

78. DFID, 2016, Support to the Employment Fund Nepal, DFID Project completion review. *Available: https://devtracker.dfid.gov.uk/projects/GB-1-201489/documents.*

79. Duflo, E., Hanna, R. & Ryan, S. P. 2012. Incentives Work: Getting Teachers to Come to School. *American Economic Rreview,* 102 (4)**:** 1241-1278.

80. Duvendack, M., Camfield, L., Thorpe, M., Zango, V., Ringot, V., Dib, G. & Fenton Villar, P., 2016, PAGE-M: Programme for the Advancement of Girls Education in Mozambique, GEC Midline Report. *Available.*

81. Glewwe, P., Nauman Ilias & Kremer, M. 2010. Teacher Incentives. *American Economic Journal: Applied Economics,* 2 (3)**:** 205-227.

82. Gulesci, S. & Jonga, M., 2016, BRACMT, GEC Midline Reportv. *Available.*

83. HEART, 2013, Independent Verification of Educational Data for a Pilot of Results-Based Aid (RBA) in Rwanda, DFID Baseline Report. *Available: http://www.heart-resources.org/wp-content/uploads/2013/06/Rwanda-Baseline-Report-education-data.pdf?x30250.*

84. Ksoll, C., Johnston, J., Sanchez Guiu, S., Jumpah, J., Carver, G., Issifu, M. & Dowley, L., 2015, MGCubed, GEC Midline Report. *Available.*

85. Linn, R. e. a., 2016, PEAS Girls’ Enrolment, Attendance, Retention & Results (GEARR) Project, GEC Midline Report. *Available.*

86. Marimira, K., Jaison, L., Chikaura, F. & Mutowa, G., 2015, Improving Girls’ Access through Transforming Education GEC Midline Report. *Available.*

87. Miske Witt & Associates, 2016, Improving Girls’ Access through Transforming Education GEC Midline Report. *Available.*

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

89. Murphy, B., Grasso, A. & Haddis, L., 2016, Pastoralist Afar Girls Education Support (PAGES) Project, GEC Midline Report. *Available.*

90. 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.

91. Opportunity International, 2016, Opportunity International Girls' Education Challenge Project, GEC Midline Report. *Available.*

92. Randall, J., Garcia, A., Boyer, M., Nordtveit, B. H., Diame, M., Berthet-Valdois, J., Ohara, K., Ranivoarianjy, R. & Ouedraogo, M., 2016, VAS-Y-Filles!, GEC Midline Report. *Available.*

93. Sinha, N. & Yoong, J., 2009, Long-Term Financial Incentives and Investment in Daughters: Evidence from Conditional Cash Transfers in North India World Bank Policy Research Working Paper 4860. *Available: https://elibrary.worldbank.org/doi/pdf/10.1596/1813-9450-4860.*

94. 95. 96. Vinuela, L. & Zoratto, L., 2015, Do Performance Agreements Help Improve Service Delivery? : The Experience of Brazilian States, World Bank Policy Research Working Paper 7375. *Available: https://openknowledge.worldbank.org/bitstream/handle/10986/22459/Do0performance00of0Brazilian0states.pdf?sequence=1&isAllowed=y.* (Note: Counted as 3 references as 3 different PbR projects dicussed in the same study, hence data extracted for each of the 3 projects separately).

97. Woldehanna, T., Fujiwara, D., Lawton, R., Tafere, Y., Reidy, E., Germana, E. & Kiwanuka, A., 2016, From early marriage, risky migration, domestic work and street life into transformative education, GEC Midline Report. *Available.*

98. Women Educational Researchers of Kenya, 2016a, Jielimishe GEC project, GEC Midline Report. *Available.*

99. Women Educational Researchers of Kenya, 2016b, Wasichana Wote Wasome (Let All Girls Learn) Project, GEC Midline Report. *Available.*

100. Ziwetz, L., Morris, M. & Lungu, D., 2016, Child-Centred Schooling: Innovation for the Improvement of Learning Outcomes for Marginalised Girls in Zambia, GEC Midline Report. *Available.*

8. Appendices

Appendix 1: Number of records returned for each search term

Science Direct		Google Scholar	
Search term		**Search term**	**Additional search term**
big results now		“big results now” (259)	“big results now” AND Tanzania (254)
“cash on delivery aid” (186)	{cash on delivery aid} (143)	“cash on delivery aid” (175)	
“financial incentives” (23,260)		“financial incentives” (413,000)	“financial incentives” AND development (52)
“incentive contracts” (1,964)	{incentive contracts} (1,406)	“incentive contracts” (27,600)	“incentive contracts” AND international development (0)
“output based aid” (45)	{output based aid} (10)	“output based aid” (1,590)	“output based aid” AND international development (1,400)
“outcome-based commissioning” (2)	{outcome-based commissioning} (0)	“outcome-based commissioning” (223)	
“outcome-based contracting” (85)	{outcome-based contracting} (10)	“outcome-based contracting” (316)	
outcome-based payment (45)	{outcome-based payment} (21)	“outcome-based payment” (325)	
payment for performance†		“payment for performance” (2,000)	“payment for performance” AND development aid (807)
“pay for performance” (4,394)	{pay for performance} (1984)	“pay for performance” (60,400)	“pay for performance” AND development aid (18,500)
“pay for quality” (497)	{pay for quality} (387)	“pay for quality” (6,060)	“pay for quality” AND development aid (2,050)
	{pay for success} (23)	“pay for success” (1,700)	“pay for success” AND development aid (751)
“payment by outcome” (2)	{payment by outcome} (1)	“payment by outcome” (75)	“payment by outcome” AND aid (12)
“payment by results” (618)		“payment by results” (14,300)	“payment by results” AND aid (4,010)
“paying for results” (55)	{paying for results} (18)	“paying for results” (645)	“paying for results” AND aid

			(240)
“payment to results” (10)	{payment to results} (0)	“payment to results” (48)	“payment to results” AND development aid (32)
“performance-based contracting” (477)	{performance-based contracting} (159)	“performance-based contracting” (5,000)	“performance-based contracting” AND international development (3,570)
“performance-based incentive” (521)	{performance-based incentive} (168)	“performance-based incentive” (3,590)	“performance-based incentive” AND international development (2,300)
“performance-based aid” (24)	{performance-based aid} (22)	“performance-based aid” (474)	
“performance-based financing” (88)	{performance-based financing} (71)	“performance-based financing” (1,710)	“performance-based financing” AND development aid (1,400)
	{performance incentives} (1,146)	“performance incentives” (23,800)	“performance incentives” AND international development (15,100)
“performance related pay” (620)	{performance related pay} (218)	“performance related pay” (19,200)	“performance related pay” AND development aid (13,200)
“program-for-results” (16)	{program-for-results} 2	“program-for-results” (379)	“program-for-results” AND world bank (287)
	{results-based financing} (56)	“results-based financing” (1,550)	“results-based financing” AND development aid (1,110)
“results based aid” (5)		“results based aid” (319)	

Notes: Number in brackets () indicates number of records returned from the search if conducted.

Appendix 2: Total number of records returned from each search

	WoS	SD	CGD	DIE	WB	ODI	IDS	3ie	R4D	OECD	AfBD	ADB	IDB	GS
big results now			1	4	1	4	0	0	1	0	7	0	0	254
cash on delivery aid		143	249	1	0	9	0	0	1	0	0	0	0	175
financial incentives		246	70	1	342		5	27	37	76	73	33	1	34
incentive contracts			2	0	8	6	0	3	0	0	0	0	0	0
output based aid	12	45	11	0	56		0	0	3	2	7	6	0	
outcome-based commissioning	1	2	0	0	0	0	0	0	0	0	0	0	0	
outcome-based contracting	6	85	5	0	0	0	0	0	0	0	0	0	0	
outcome-based payment	4	45	5	0	0	2	0	0	0	1	0	0	0	
payment for performance			49	0	3	9	0	0	0	0	2	0	0	
pay for performance			102	0	21	9	0	5	8	8	23	0	0	
pay for quality		30**	0	0	3	1	0	0	0	1	0	1	0	
pay for success	6	23	26	0	0	2	0	0	0	0	0	0	0	751*
payment by outcome	1	2	0	0	0	0	0	0	0	0	0	0	0	
payment by results	175	618	49	2	7		4	0	5	0	4	0	0	4010*
paying for results	7	55	54	2	1	0	1	1	0	0	2	0	0	240
payment to results	0	10	1	3	7	1	4	0	0	0	4	0	0	
performance-based contracting	103	64**	4	0	18		0	1	4	2	10	11	0	
performance-based incentive		168	66	0	5		0	4	4	3	5	1	0	
performance-based aid	6	24	7	1	1		0	0	2	0	5	1	0	
performance-based financing	73	88	58	3	15		0	3	2	0	4	0	0	
performance incentives		73**	51	0	42		0	6	5	9	12	4	0	
performance related pay		620	0	0	5		0	1	4	14	2	0	0	
program-for-results	1	16	35	2	12	8	0	1	0	0	40	1	0	
results-based financing	28	56	67	5	40		0	0	7	0	31	11	2	
results based aid	2	5	88	28	1		1	0	5	1	0	0	0	
Total	**425**	**2418**	**1000**	**52**	**588**	**51**	**15**	**52**	**88**	**117**	**231**	**69**	**3**	**5464**

Notes: i) Only publications were searched in ODI and the WB repository, ii) Only the impact evaluations were searched in the 3ie repository, iii) SD and GS search terms were altered as outlined in Appendix 1, iii) ***** Denotes only first 40 (ordered by relevance) screened when more than 500 hits returned, iv) ****** Denotes search limited to Title, Abstract and Keywords, v) blank cells indicate term has not been screened.

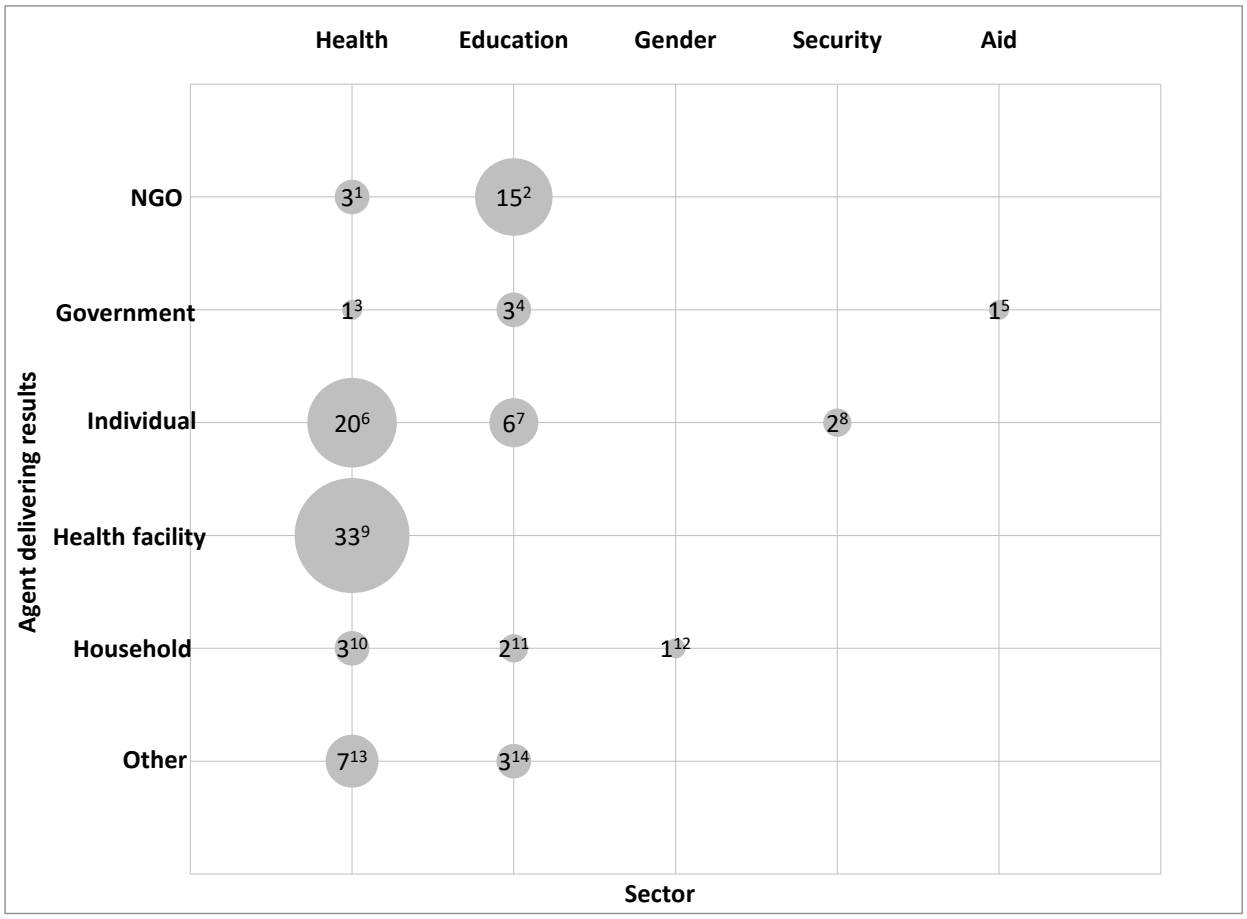
Appendix 3: Missing data

Table A3: Missing data for the following studies for key variables: principal (12 studies of a sample of 100), agent (12 studies of a sample of 100) and measure (17 studies of a sample of 100)

Principal	Agent	Measure
(Alonge et al., 2015, Angelucci and Attanasio, 2009, Barham, 2011, de Walque et al., 2014, Huillery and Sebanz, 2014, Kohler and Thornton, 2012, Olken et al., 2014, Robyn et al., 2014, Soares et al., 2008, Yao et al., 2008, Yip et al., 2014, Zhang et al., 2013)	(Angelucci and Attanasio, 2009, Binagwaho et al., 2014, DFID, 2016, Lannes, 2015, Lannes et al., 2015, Moyo et al., 2015, na, na, Soeters et al., 2007, Sood et al., 2014, Urquieta et al., 2009, Vinuela and Zoratto, 2015, Zeng et al., 2014)	(Alonge et al., 2015, Angelucci and Attanasio, 2009, Lannes, 2015, Lannes et al., 2015, Moyo et al., 2015, na, na, Rusa et al., 2009b, Soeters et al., 2007, Sood et al., 2014, Urquieta et al., 2009, Valadez et al., 2015, Vinuela and Zoratto, 2015, Wei et al., 2012, Witter et al., 2016, World Bank, 2013, Yip et al., 2014, Zeng et al., 2014)

Appendix 4: Evidence gap maps with details of studies

Map A1: All studies, all risk of bias levels, by sector and by agent, details of studies in footnotes

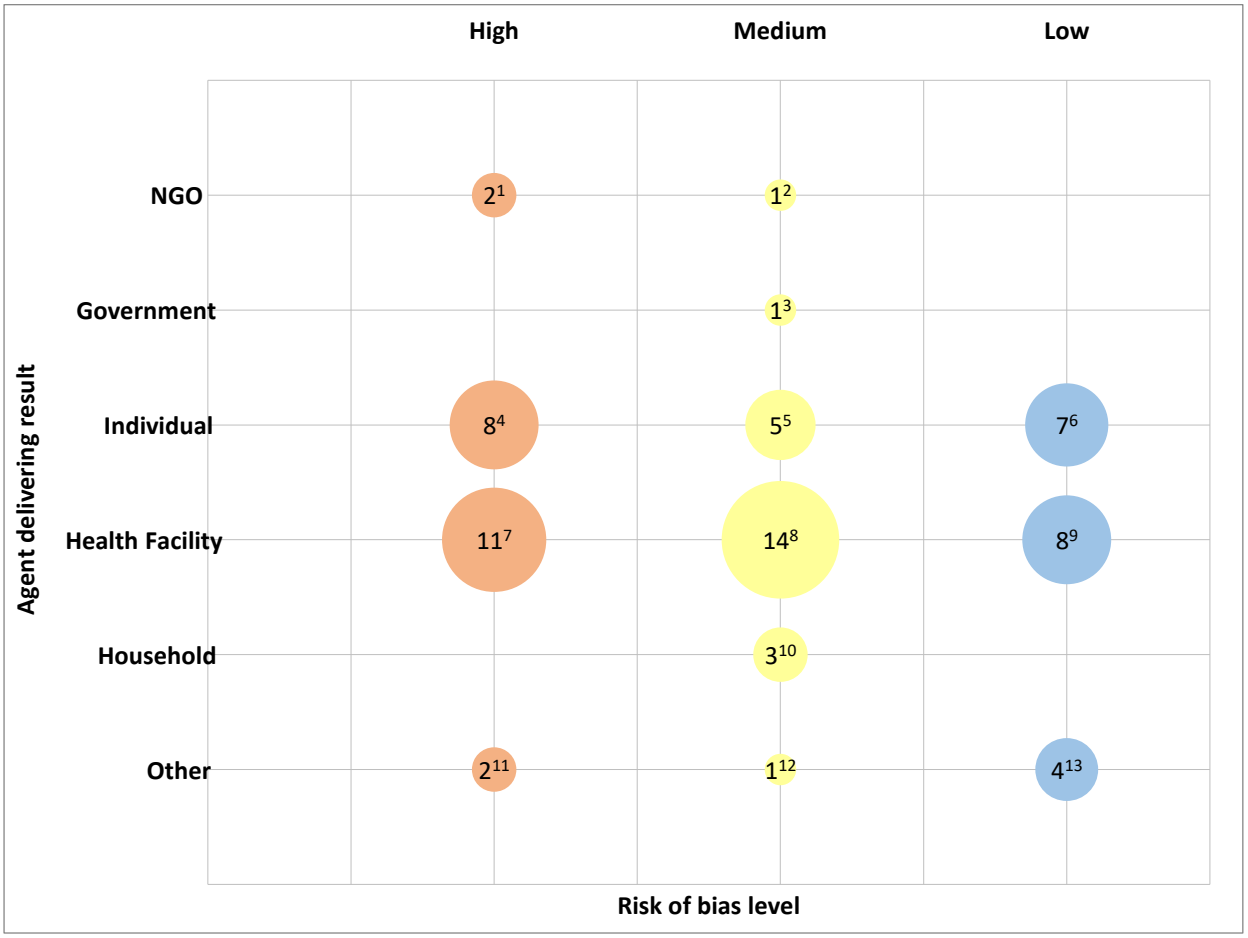


Notes: Map A1 lists 67 rather than 71 health studies because 9 health studies do not provide any information on 'Who gets paid?' - the agent - while 5 studies report 2 different agents (these appear twice in the map), hence 67 studies. The section 'Other' includes very study specific agents, e.g. villages, cooperatives, central medical stores, schools or school principals, local civic organisations.

[1] (Alonge et al., 2015, Eichler et al., 2007, Garcia Prado and Lao Peña, 2010)
[2] (Bagyendera et al., 2016, Camfed International, 2016, Duvendack et al., 2016, Gulesci and Jonga, 2016, Ksoll et al., 2015, Linn, 2016, Marimira et al., 2015, Miske Witt & Associates, 2016, Murphy et al., 2016, Opportunity International, 2016, Randall et al., 2016, Woldehanna et al., 2016, Women Educational Researchers of Kenya, 2016b, Women Educational Researchers of Kenya, 2016a, Ziwetz et al., 2016)
[3] (Gertler et al., 2014)
[4] (Cambridge Education, 2015, Clist et al., 2015, HEART, 2013)
[5] (Öhler et al., 2012)
[6] (Ashraf et al., 2013, Barham, 2011, Bossuroy et al., 2016, de Walque et al., 2014, Engineer et al., 2016, Fox et al., 2014, Gertler, 2000, Gopalan and Varatharajan, 2012, Ir et al., 2015, Kohler and Thornton, 2012, Lim et al., 2010, Liu and Mills, 2005, Peabody et al., 2014, Powell-Jackson and Hanson, 2012, Powell-Jackson et al., 2015, Renaud and Semasaka, 2014, Rusa et al., 2009b, Wei et al., 2012, World Bank, 2013, Yao et al., 2008)
[7] (Barrera-Osorio and Raju, 2015, Duflo et al., 2012, Glewwe et al., 2010, Muralidharan and Sundararaman, 2011, Vinuela and Zoratto, 2015)
[8] (Vinuela and Zoratto, 2015)
[9] (Alonge et al., 2015, Basinga et al., 2011, Bonfrer et al., 2014a, Bonfrer et al., 2014b, Bowser et al., 2013, Celhay et al., 2015, Chansa et al., 2015, Cornejo-Ovalle et al., 2015, de Walque et al., 2015, Falisse et al., 2015, Gertler and Vermeersch, 2012, Huillery and Sebanz, 2014, Janssen et al., 2015, Kumar et al., 2016, Mac Arthur et al., 2014, Matsuoka et al., 2014, na, 2016, Nahimana et al., 2015, Obare et al., 2016, Powell-Jackson and Hanson, 2012, Regalía and Castro, 2007, Robyn et al., 2014, Rusa et al., 2009a, Rusa et al., 2009b, Sherry et al.,

2017, Skiles et al., 2012, Skiles et al., 2015, Soeters et al., 2011, Sun et al., 2016, Valadez et al., 2015, Van de Poel et al., 2016, Yip et al., 2014)
[10] (Leroy et al., 2008, Regalía and Castro, 2007, Soares et al., 2008)
[11] (Leroy et al., 2008, Soares et al., 2008)
[12] (Sinha and Yoong, 2009)
[13] (Olken et al., 2014, Renaud and Semasaka, 2014, Spisak et al., 2016, Sylvia et al., 2013, Witter et al., 2016, Zeng et al., 2013, Zhang et al., 2013)
[14] (Olken et al., 2014, Vinuela and Zoratto, 2015)

Map A2: Health studies by risk of bias levels and by agent, details of studies in footnotes

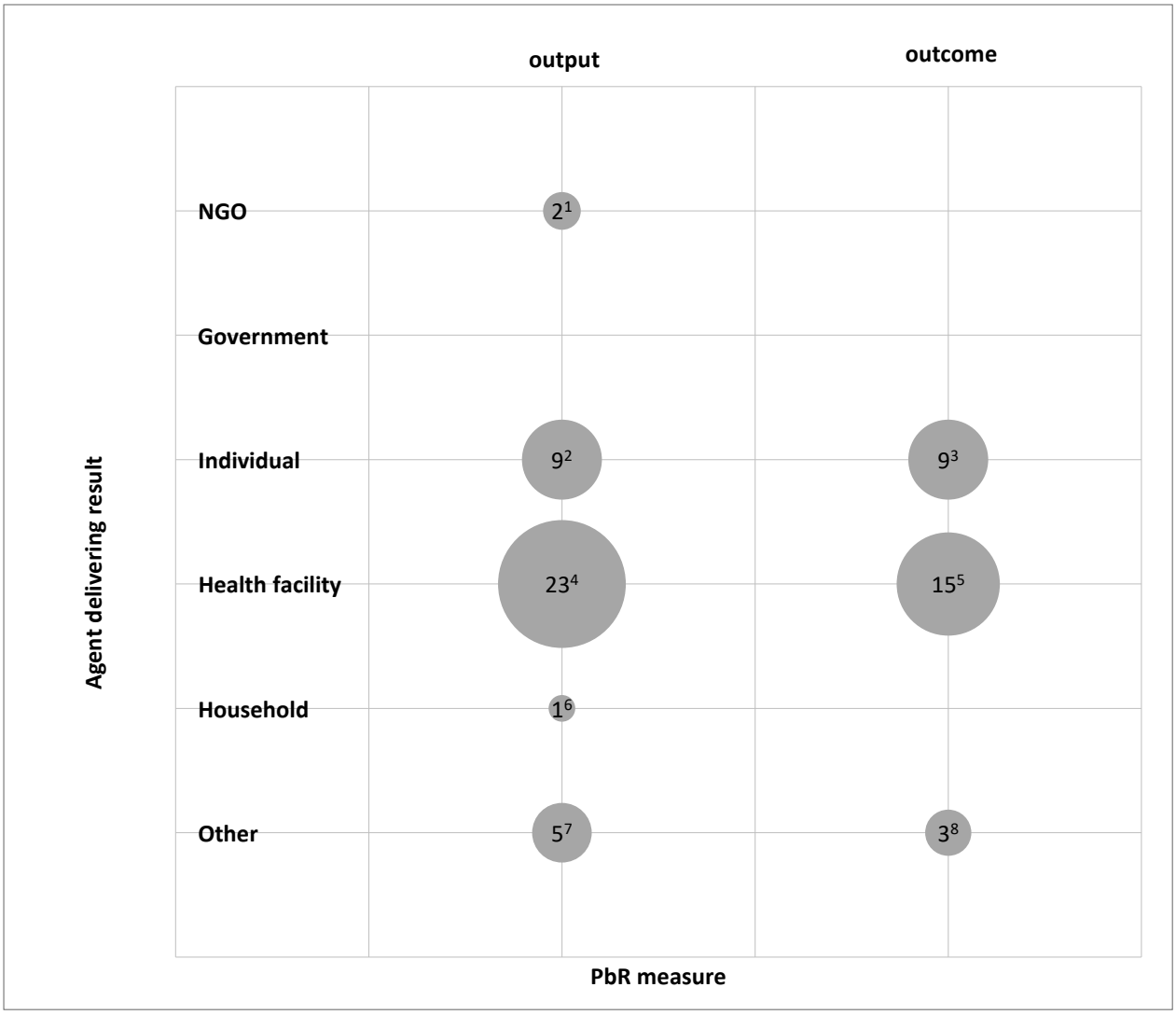


Notes: As in the case of Map A1, 67 rather than 71 health studies are captured because 9 health studies do not provide any information on 'Who gets paid?' - the agent - while 5 studies report 2 different agents (these appear twice in the map), hence 67 studies. The section 'Other' includes very study specific agents, e.g. villages, cooperatives, central medical stores, schools or school principals, local civic organisations.

[1] (Eichler et al., 2007, Garcia Prado and Lao Peña, 2010)
[2] (Alonge et al., 2015)
[3] (Gertler et al., 2014)
[4] (Fox et al., 2014, Gopalan and Varatharajan, 2012, Ir et al., 2015, Liu and Mills, 2005, Rusa et al., 2009b, Wei et al., 2012, World Bank, 2013, Yao et al., 2008)
[5] (Engineer et al., 2016, Lim et al., 2010, Powell-Jackson and Hanson, 2012, Powell-Jackson et al., 2015, Renaud and Semasaka, 2014)
[6] (Ashraf et al., 2013, Barham, 2011, Bossuroy et al., 2016, de Walque et al., 2014, Gertler, 2000, Kohler and Thornton, 2012, Peabody et al., 2014)
[7] (Bonfrer et al., 2014a, Chansa et al., 2015, Cornejo-Ovalle et al., 2015, Janisch et al., 2010, Janssen et al., 2015, Mac Arthur et al., 2014, Matsuoka et al., 2014, Nahimana et al., 2015, Robyn et al., 2014, Rusa et al., 2009a, Rusa et al., 2009b)

[8] (Alonge et al., 2015, Bonfrer et al., 2014b, Bowser et al., 2013, Celhay et al., 2015, Falisse et al., 2015, Kumar et al., 2016, na, 2016, Obare et al., 2016, Powell-Jackson and Hanson, 2012, Regalía and Castro, 2007, Sherry et al., 2017, Sun et al., 2016, Valadez et al., 2015, Van de Poel et al., 2016)
[9] (Basinga et al., 2011, de Walque et al., 2015, Gertler and Vermeersch, 2012, Huillery and Sebanz, 2014, Skiles et al., 2012, Skiles et al., 2015, Soeters et al., 2011, Yip et al., 2014)
[10] (Leroy et al., 2008, Regalía and Castro, 2007, Soares et al., 2008)
[11] (Spisak et al., 2016, Zeng et al., 2013)
[12] (Renaud and Semasaka, 2014)
[13] (Olken et al., 2014, Sylvia et al., 2013, Witter et al., 2016, Zhang et al., 2013)

Map A3: Health studies by PbR measure and by agent, all risk of bias levels, details of studies in footnotes



Notes: As in the case of Map A1, 67 rather than 71 health studies are captured because 9 health studies do not provide any information on 'Who gets paid?' - the agent - while 5 studies report 2 different agents (these appear twice in the map), hence 67 studies. 15 studies do not report any information on the pre-agreed measure but of the 56 studies reporting pre-agreed measures 32 report 2 or even 3 measures. The section 'Other' includes very study specific agents, e.g. villages, cooperatives, central medical stores, schools or school principals, local civic organisations.

[1] (Eichler et al., 2007, Garcia Prado and Lao Peña, 2010)
[2] (Ashraf et al., 2013, Barham, 2011, Bossuroy et al., 2016, Engineer et al., 2016, Gertler, 2000, Gopalan and Varatharajan, 2012, Liu and Mills, 2005, Renaud and Semasaka, 2014, Yao et al., 2008)

[3] (de Walque et al., 2014, Engineer et al., 2016, Gopalan and Varatharajan, 2012, Ir et al., 2015, Kohler and Thornton, 2012, Lim et al., 2010, Powell-Jackson and Hanson, 2012, Powell-Jackson et al., 2015, Renaud and Semasaka, 2014)
[4] (Basinga et al., 2011, Bonfrer et al., 2014a, Bonfrer et al., 2014b, Celhay et al., 2015, Chansa et al., 2015, Cornejo-Ovalle et al., 2015, de Walque et al., 2015, Falisse et al., 2015, Gertler and Vermeersch, 2012, Janisch et al., 2010, Janssen et al., 2015, Mac Arthur et al., 2014, Matsuoka et al., 2014, na, 2016, Obare et al., 2016, Regalía and Castro, 2007, Robyn et al., 2014, Rusa et al., 2009a, Sherry et al., 2017, Skiles et al., 2012, Soeters et al., 2011, Sun et al., 2016, Van de Poel et al., 2016)
[5] (Basinga et al., 2011, Bonfrer et al., 2014a, Chansa et al., 2015, Janisch et al., 2010, Kumar et al., 2016, Mac Arthur et al., 2014, na, 2016, Nahimana et al., 2015, Obare et al., 2016, Powell-Jackson and Hanson, 2012, Rusa et al., 2009a, Sherry et al., 2017, Skiles et al., 2012, Soeters et al., 2011, Van de Poel et al., 2016)
[6] (Regalía and Castro, 2007)
[7] (Liu and Mills, 2005, Olken et al., 2014, Renaud and Semasaka, 2014, Sylvia et al., 2013, Zeng et al., 2013, Zhang et al., 2013)
[8] (Olken et al., 2014, Renaud and Semasaka, 2014, Spisak et al., 2016)

Appendix 5: Health synthesis: Description of the sample with details of studies

Table A5.1: Health sub-themes, **across all RoB**, individual and health facility only

Health subtheme	**% of studies**	**Details of studies**
Maternal and child health	47.4	(Barham, 2011, Basinga et al., 2011, Bonfrer et al., 2014b, Celhay et al., 2015, Chansa et al., 2015, Engineer et al., 2016, Falisse et al., 2015, Gertler, 2000, Gertler and Vermeersch, 2012, Gopalan and Varatharajan, 2012, Huillery and Sebanz, 2014, Ir et al., 2015, Janisch et al., 2010, Lim et al., 2010, Matsuoka et al., 2014, Obare et al., 2016, Powell-Jackson and Hanson, 2012, Powell-Jackson et al., 2015, Regalía and Castro, 2007, Renaud and Semasaka, 2014, Rusa et al., 2009b, Sherry et al., 2017, Skiles et al., 2012, Skiles et al., 2015, Van de Poel et al., 2016, World Bank, 2013)
HIV/STI	13.6	(Ashraf et al., 2013, Chansa et al., 2015, de Walque et al., 2015, Falisse et al., 2015, Obare et al., 2016, Rusa et al., 2009b)
Nutrition	5.1	(Gertler, 2000, Nahimana et al., 2015, Regalía and Castro, 2007)
Health general /Other	33.9	(Alonge et al., 2015, Bonfrer et al., 2014a, Bossuroy et al., 2016, Bowser et al., 2013, Cornejo-Ovalle et al., 2015, de Walque et al., 2014, Fox et al., 2014, Janssen et al., 2015, Kohler and Thornton, 2012, Kumar et al., 2016, Liu and Mills, 2005, Mac Arthur et al., 2014, Nahimana et al., 2015, Peabody et al., 2014, Robyn et al., 2014, Rusa et al., 2009a, Soeters et al., 2011, Sun et al., 2016, Valadez et al., 2015, Wei et al., 2012, Yao et al., 2008, Yip et al., 2014)
Total	**100**	

Table A5.2: Geographical region, **across all RoB**, individual and health facility only

Region	**% of studies**	**Details of studies**
Sub-Sahara Africa	52.9	(Ashraf et al., 2013, Basinga et al., 2011, Bonfrer et al., 2014a, Bonfrer et al., 2014b, Chansa et al., 2015, de Walque et al., 2014, de Walque et al., 2015, Falisse et al., 2015, Fox et al., 2014, Gertler and Vermeersch, 2012, Huillery and Sebanz, 2014, Janisch et al., 2010, Janssen et al., 2015, Kohler and Thornton, 2012, Kumar et al., 2016, na, 2016, Nahimana et al., 2015, Obare et al., 2016, Renaud and Semasaka, 2014, Robyn et al., 2014, Rusa et al., 2009a, Rusa et al., 2009b, Sherry et al., 2017, Skiles et al., 2012, Skiles et al., 2015, Soeters et al., 2011, Valadez et al., 2015)
Latin America	13.8	(Barham, 2011, Bowser et al., 2013, Celhay et al., 2015, Cornejo-Ovalle et al., 2015, Gertler,

		2000, Mac Arthur et al., 2014, Regalía and Castro, 2007)
South Asia	9.8	(Bossuroy et al., 2016, Gopalan and Varatharajan, 2012, Lim et al., 2010, Powell-Jackson and Hanson, 2012, Powell-Jackson et al., 2015)
East Asia	9.8	(Liu and Mills, 2005, Sun et al., 2016, Wei et al., 2012, Yao et al., 2008, Yip et al., 2014)
Southeast Asia	7.8	(Ir et al., 2015, Matsuoka et al., 2014, Peabody et al., 2014, Van de Poel et al., 2016)
Middle East	5.9	(Alonge et al., 2015, Engineer et al., 2016, World Bank, 2013)
Total	**100**	

Risk of bias	**% of studies**	**Details of studies**
High	35.3	(Bonfrer et al., 2014a, Chansa et al., 2015, Cornejo-Ovalle et al., 2015, Fox et al., 2014, Gopalan and Varatharajan, 2012, Ir et al., 2015, Janisch et al., 2010, Janssen et al., 2015, Liu and Mills, 2005, Mac Arthur et al., 2014, Matsuoka et al., 2014, Nahimana et al., 2015, Robyn et al., 2014, Rusa et al., 2009a, Rusa et al., 2009b, Wei et al., 2012, World Bank, 2013, Yao et al., 2008)
Medium	35.3	(Alonge et al., 2015, Bonfrer et al., 2014b, Bowser et al., 2013, Celhay et al., 2015, Engineer et al., 2016, Falisse et al., 2015, Kumar et al., 2016, Lim et al., 2010, na, 2016, Obare et al., 2016, Powell-Jackson and Hanson, 2012, Powell-Jackson et al., 2015, Regalía and Castro, 2007, Renaud and Semasaka, 2014, Sherry et al., 2017, Sun et al., 2016, Valadez et al., 2015, Van de Poel et al., 2016)
Low	29.4	(Ashraf et al., 2013, Barham, 2011, Basinga et al., 2011, Bossuroy et al., 2016, de Walque et al., 2014, de Walque et al., 2015, Gertler, 2000, Gertler and Vermeersch, 2012, Huillery and Sebanz, 2014, Kohler and Thornton, 2012, Peabody et al., 2014, Skiles et al., 2012, Skiles et al., 2015, Soeters et al., 2011, Yip et al., 2014)
Total	**100**	

Table A5.3: Risk of bias, **individual and health facility only**

Appendix 6: Health synthesis: Success and failure of PbR with details of studies

Table A6.1: Success and failure of all PbR measures (output and outcome), **across all risk of bias levels,** details of studies in footnotes

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	30[1]	4[2]	34
Not significant	24[3]	3[4]	27
Total	54	7	61

Notes: Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures, all risk of bias levels.

[1] (Ashraf et al., 2013, Barham, 2011, Basinga et al., 2011, Bonfrer et al., 2014a, Bonfrer et al., 2014b, Bossuroy et al., 2016, Celhay et al., 2015, Chansa et al., 2015, Cornejo-Ovalle et al., 2015, Eichler et al., 2007, Garcia Prado and Lao Peña, 2010, Gertler, 2000, Ir et al., 2015, Kumar et al., 2016, Lim et al., 2010, Liu and Mills, 2005, Matsuoka et al., 2014, na, 2016, Nahimana et al., 2015, Obare et al., 2016, Olken et al., 2014, Powell-Jackson and Hanson, 2012, Powell-Jackson et al., 2015, Sherry et al., 2017, Skiles et al., 2012, Soeters et al., 2011, Sylvia et al., 2013, Van de Poel et al., 2016, Zeng et al., 2013, Zhang et al., 2013)
[2] (Liu and Mills, 2005, Sherry et al., 2017, Sylvia et al., 2013, Yao et al., 2008)
[3] (Ashraf et al., 2013, Basinga et al., 2011, Bonfrer et al., 2014a, Bonfrer et al., 2014b, Bossuroy et al., 2016, de Walque et al., 2014, de Walque et al., 2015, Engineer et al., 2016, Falisse et al., 2015, Garcia Prado and Lao Peña, 2010, Gertler and Vermeersch, 2012, Kumar et al., 2016, Lim et al., 2010, na, 2016, Obare et al., 2016, Olken et al., 2014, Powell-Jackson et al., 2015, Regalía and Castro, 2007, Sherry et al., 2017, Skiles et al., 2012, Soeters et al., 2011, Sun et al., 2016, Van de Poel et al., 2016, Zhang et al., 2013)
[4] (Engineer et al., 2016, Kohler and Thornton, 2012, Sherry et al., 2017)

Table A6.2: Success and failure of all PbR measures (output only**), across all risk of bias levels,** details of studies in footnotes

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	24[1]	4[2]	28
Not significant	23[3]	2[4]	25
Total	47	6	53

Notes Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures, all risk of bias levels. The figures of Tables A6.2 and A6.3 do not add up to the ones reported in Table A6.1 as some studies report both output and outcome measures, if this is the case the study is only counted once in Table A6.1 but appears in both Tables A6.2 and A6.3.

[1] (Ashraf et al., 2013, Barham, 2011, Basinga et al., 2011, Bonfrer et al., 2014a, Bonfrer et al., 2014b, Bossuroy et al., 2016, Celhay et al., 2015, Chansa et al., 2015, Cornejo-Ovalle et al., 2015, Eichler et al., 2007, Garcia Prado and Lao Peña, 2010, Gertler, 2000, Liu and Mills, 2005, Matsuoka et al., 2014, na, 2016, Obare et al., 2016, Olken et al., 2014, Sherry et al., 2017, Skiles et al., 2012, Soeters et al., 2011, Sylvia et al., 2013, Van de Poel et al., 2016, Zeng et al., 2013, Zhang et al., 2013)
[2] (Liu and Mills, 2005, Sherry et al., 2017, Sylvia et al., 2013, Yao et al., 2008)
[3] (Ashraf et al., 2013, Basinga et al., 2011, Bonfrer et al., 2014a, Bonfrer et al., 2014b, Bossuroy et al., 2016, de Walque et al., 2015, Engineer et al., 2016, Falisse et al., 2015, Garcia Prado and Lao Peña, 2010, Gertler and Vermeersch, 2012, na, 2016, Obare et al., 2016, Olken et al., 2014, Regalía and Castro, 2007, Sherry et al., 2017, Skiles et al., 2012, Soeters et al., 2011, Sun et al., 2016, Van de Poel et al., 2016, Zhang et al., 2013)
[4] (Engineer et al., 2016, Sherry et al., 2017)

Table A6.3: Success and failure of all PbR measures (outcome only), **across all risk of bias levels,** details of studies in footnotes

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	16[1]	1[2]	17
Not significant	14[3]	3[4]	17
Total	30	4	34

Notes: Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures, all risk of bias levels. The figures of Tables A6.2 and A6.3 do not add up to the ones reported in Table A6.1 as some studies report both output and outcome measures, if this is the case the study is only counted once in Table A6.1 but appears in both Tables A6.2 and A6.3.

[1] (Basinga et al., 2011, Bonfrer et al., 2014a, Chansa et al., 2015, Ir et al., 2015, Kumar et al., 2016, Lim et al., 2010, na, 2016, Nahimana et al., 2015, Obare et al., 2016, Olken et al., 2014, Powell-Jackson and Hanson, 2012, Powell-Jackson et al., 2015, Sherry et al., 2017, Skiles et al., 2012, Soeters et al., 2011, Van de Poel et al., 2016)
[2] (Sherry et al., 2017)
[3] (Basinga et al., 2011, Bonfrer et al., 2014a, de Walque et al., 2014, Engineer et al., 2016, Kumar et al., 2016, Lim et al., 2010, na, 2016, Obare et al., 2016, Powell-Jackson et al., 2015, Sherry et al., 2017, Skiles et al., 2012, Soeters et al., 2011, Van de Poel et al., 2016)
[4] (Engineer et al., 2016, Kohler and Thornton, 2012, Sherry et al., 2017)

Appendix 7: Sub-group analysis: Success and failure of PbR measures by risk of bias levels

Table A7.1: Success and failure of all PbR measures (output and outcome), **high risk of bias levels**

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	10[1]	2[2]	12
Not significant	2[3]	0	2
Total	12	2	14

Notes: Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures.

[1] (Bonfrer et al., 2014a, Chansa et al., 2015, Cornejo-Ovalle et al., 2015, Eichler et al., 2007, Garcia Prado and Lao Peña, 2010, Ir et al., 2015, Liu and Mills, 2005, Matsuoka et al., 2014, Nahimana et al., 2015, Zeng et al., 2013)
[2] (Liu and Mills, 2005, Yao et al., 2008)
[3] (Bonfrer et al., 2014a, Garcia Prado and Lao Peña, 2010)

Table A7.2: Success and failure of all PbR measures (output only), **high risk of bias levels**

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	9[1]	2[2]	11
Not significant	2[3]	0	2
Total	11	2	13

Notes: Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures. The figures of Tables A7.2 and A7.3 do not add up to the ones reported in Table A7.1 as some studies report both output and outcome measures, if this is the case the study is only counted once in Table A7.1 but appears in both Tables A7.2 and A7.3.

[1] (Bonfrer et al., 2014a, Chansa et al., 2015, Cornejo-Ovalle et al., 2015, Eichler et al., 2007, Garcia Prado and Lao Peña, 2010, Liu and Mills, 2005, Matsuoka et al., 2014, Yao et al., 2008, Zeng et al., 2013)
[2] (Liu and Mills, 2005, Yao et al., 2008)
[3] (Bonfrer et al., 2014a, Garcia Prado and Lao Peña, 2010)

Table A7.3: Success and failure of all PbR measures (outcome only), **high risk of bias levels**

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	4[1]	0	4
Not significant	1[2]	0	1
Total	5	0	5

Notes: Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures. The figures of Tables A7.2 and A7.3 do not add up to the ones reported in Table A7.1 as some studies report both output and outcome measures, if this is the case the study is only counted once in Table A7.1 but appears in both Tables A7.2 and A7.3.

[1] (Bonfrer et al., 2014a, Chansa et al., 2015, Ir et al., 2015, Nahimana et al., 2015)
[2] (Bonfrer et al., 2014a)

Table A7.4: Success and failure of all PbR measures (output and outcome), **medium risk of bias levels,** details of studies in footnotes

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	10[1]	1[2]	11
Not significant	12[3]	2[4]	14
Total	22	3	25

Notes: Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures.

[1] (Bonfrer et al., 2014b, Celhay et al., 2015, Kumar et al., 2016, Lim et al., 2010, na, 2016, Obare et al., 2016, Powell-Jackson and Hanson, 2012, Powell-Jackson et al., 2015, Sherry et al., 2017, Van de Poel et al., 2016)

[2] (Sherry et al., 2017)

[3] (Bonfrer et al., 2014a, Engineer et al., 2016, Falisse et al., 2015, Kumar et al., 2016, Lim et al., 2010, Obare et al., 2016, Powell-Jackson and Hanson, 2012, Powell-Jackson et al., 2015, Regalía and Castro, 2007, Sherry et al., 2017, Sun et al., 2016, Van de Poel et al., 2016)

[4] (Engineer et al., 2016, Sherry et al., 2017)

Table A7.5: Success and failure of all PbR measures (output only), **medium risk of bias levels,** details of studies in footnotes

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	6[1]	1[2]	7
Not significant	9[3]	2[4]	11
Total	15	3	18

Notes: Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures. The figures of Tables A7.5 and A7.6 do not add up to the ones reported in Table A7.4 as some studies report both output and outcome measures, if this is the case the study is only counted once in Table A7.4 but appears in both Tables A7.5 and A7.6.

[1] (Bonfrer et al., 2014b, Celhay et al., 2015, na, 2016, Obare et al., 2016, Sherry et al., 2017, Van de Poel et al., 2016)

[2] (Sherry et al., 2017)

[3] (Bonfrer et al., 2014b, Engineer et al., 2016, Falisse et al., 2015, na, 2016, Obare et al., 2016, Regalía and Castro, 2007, Sherry et al., 2017, Sun et al., 2016, Van de Poel et al., 2016)

[4] (Engineer et al., 2016, Sherry et al., 2017)

Table A7.6: Success and failure of all PbR measures (outcome only), **medium risk of bias levels,** details of studies in footnotes

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	8[1]	1[2]	9
Not significant	8[3]	2[4]	10
Total	16	3	19

Notes: Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures. The figures of Tables A7.5 and A7.6 do not add up to the ones reported in Table A7.4 as some studies report both output and outcome measures, if this is the case the study is only counted once in Table A7.4 but appears in both Tables A7.5 and A7.6.

[1] (Kumar et al., 2016, Lim et al., 2010, na, 2016, Obare et al., 2016, Powell-Jackson and Hanson, 2012, Powell-Jackson et al., 2015, Sherry et al., 2017, Van de Poel et al., 2016)

[2] (Sherry et al., 2017)

[3] (Engineer et al., 2016, Kumar et al., 2016, Lim et al., 2010, na, 2016, Obare et al., 2016, Powell-Jackson et al., 2015, Sherry et al., 2017, Van de Poel et al., 2016)

[4] (Engineer et al., 2016, Sherry et al., 2017)

Table A7.7: Success and failure of all PbR measures (output and outcome), **low risk of bias levels,** details of studies in footnotes

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	10[1]	1[2]	11
Not significant	10[3]	1[4]	11
Total	20	2	22

Notes: Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures.

[1] (Ashraf et al., 2013, Barham, 2011, Basinga et al., 2011, Bossuroy et al., 2016, Gertler, 2000, Olken et al., 2014, Skiles et al., 2012, Soeters et al., 2011, Sylvia et al., 2013, Zhang et al., 2013)
[2] (Sylvia et al., 2013)
[3] (Ashraf et al., 2013, Basinga et al., 2011, Bossuroy et al., 2016, de Walque et al., 2014, de Walque et al., 2015, Gertler and Vermeersch, 2012, Olken et al., 2014, Skiles et al., 2012, Soeters et al., 2011, Zhang et al., 2013)
[4] (Kohler and Thornton, 2012)

Table A7.8: Success and failure of all PbR measures (output only), **low risk of bias levels,** details of studies in footnotes

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	10[1]	1[2]	11
Not significant	9[3]	0	9
Total	19	1	20

Notes: Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures. The figures of Tables A7.8 and A7.9 do not add up to the ones reported in Table A7.7 as some studies report both output and outcome measures, if this is the case the study is only counted once in Table A7.7 but appears in both Tables A7.8 and A7.9.

[1] (Ashraf et al., 2013, Barham, 2011, Basinga et al., 2011, Bossuroy et al., 2016, Gertler, 2000, Olken et al., 2014, Skiles et al., 2012, Soeters et al., 2011, Sylvia et al., 2013, Zhang et al., 2013)
[2] (Sylvia et al., 2013)
[3] (Ashraf et al., 2013, Basinga et al., 2011, Bossuroy et al., 2016, de Walque et al., 2015, Gertler and Vermeersch, 2012, Olken et al., 2014, Skiles et al., 2012, Soeters et al., 2011, Zhang et al., 2013)

Table A7.9: Success and failure of all PbR measures (outcome only), **low risk of bias levels,** details of studies in footnotes

PbR measure	Positive (# of estimates)	Negative (# of estimates)	Total (# of estimates)
Significant	4[1]	0	4
Not significant	5[2]	1[3]	6
Total	9	1	10

Notes: Health studies only, not all studies reported information on success or failure but some studies reported multiple effects due to using multiple measures. The figures of Tables A7.8 and A7.9 do not add up to the ones reported in Table A7.7 as some studies report both output and outcome measures, if this is the case the study is only counted once in Table A7.7 but appears in both Tables A7.8 and A7.9.

[1] (Basinga et al., 2011, Olken et al., 2014, Skiles et al., 2012, Soeters et al., 2011)
[2] (de Walque et al., 2014, Engineer et al., 2016, Murphy et al., 2016, Olken et al., 2014, Woldehanna et al., 2016)
[3] (Kohler and Thornton, 2012)