Scaling-up Results-Based Financing for Community Health

Learning Agenda Full Report

June 2023

LivingGoods

USAID FROM THE AMERICAN PEOPLE

GDI Global Development Incubator

INSTIGLIO TYING FUNDING TO IMPACT

DEERFIELD

Acknowledgements

This Learning Agenda was authored by the following Instiglio staff: Nicole Pflock, Tabitha Ngugi, Melissa Kaminker and Obed Matara. The report also benefited from the contributions and feedback of individuals from the following organizations: USAID DIV, Deerfield Foundation, Living Goods, GDI, and IPA. Their input was instrumental in shaping the insights and lessons highlighted in this report.

June 2023

USAID Grant: 7200AA19FA000147 to GDI
August 12, 2019 - August 11, 2023
Development Innovation Ventures: Stage 3 Award

This report is made possible by the generous support of the American people through the United States Agency for International Development (USAID). The contents are the responsibility of Instiglio, and do not reflect the views of USAID or the United States Government.

INSTIGLIO

Table of Contents

List of Tables

List of Figures

Abbreviations

ANC	Antenatal Care
CHW	Community Health Worker
DESC	Digitally Enabled, Supervised, and Compensation
DQA	Data Quality Assessment
GDI	Global Development Incubator
GOU	Government of Uganda
GPS	Global Positioning System
iCCM	Integrated Community Case Management
IPA	Innovations for Poverty Action
LG	Living Goods
MEL	Monitoring, Evaluation and Learning
PNC	Postnatal Care
RBF	Results-Based Financing
RCT	Randomized Control Trials
TWG	Technical Working Group
USAID DIV	United States Agency for International Development Development Innovation Ventures

1. Introduction

1.1. About Living Goods

Since its founding in 2007, Living Goods (LG) has supported nearly 11,000 digitally empowered community health workers (CHWs) to deliver care, improving the ability of families to access the treatment and care they need.[1] LG has supported CHWs to go door-to-door in their communities, delivering an integrated package of reproductive, maternal, newborn, and child health interventions. Specific services provided include:[2]

1) **Integrated community case management (iCCM):** When a child falls ill, CHWs are guided by iCCM workflows on their smartphone app to provide automated diagnosis and standardized treatment, and to flag acute cases for referral to a qualified health facility.
2) **Pre/Postnatal Care:** CHWs provide early pregnancy diagnoses and education on maternal health and nutrition. They refer high-risk pregnancies, monitor the expected delivery date, and work to ensure that all pregnant women give birth in a health facility.
3) **Immunization:** CHWs capture the immunization status of every child in their community and work closely with health facilities to target defaulters. They use messaging and behavior change to counter barriers to and drive greater demand for vaccinations.
4) **Family Planning:** At clients' requests, CHWs provide comprehensive family planning education and counseling, as well as contraceptives when permitted by law—including condoms, birth control pills, and the 3-month injectable Sayana Press—and referrals for long-term methods.
5) **COVID-19:** CHWs have maintained essential health services in their communities during the pandemic, despite the increasingly challenging operating environment. While visits to health facilities declined from 2019 to 2020, CHW treatments doubled in their areas of operation.
6) **Health Education:** A key part of CHWs' work is providing health education, including on the prevention and treatment of common diseases like malaria; hand washing and other safe water, sanitation, and hygiene practices; and proper nutrition.

LG supports its CHWs to deliver high impact, cost-effective community health services by providing them with a smartphone and diagnostic health app, medicines and health tools, real-time supervision, and compensation for their work, comprising LG's *Digitally Enabled, Equipped, Supervised, and Compensation (DESC)* framework.[3] LG collaborates closely with the governments of Kenya, Uganda, and Burkina Faso to support them to develop their own CHW networks and build a policy environment needed to empower CHWs in the long-term.[4] In their goal to sustainably reach all communities in need, LG has been experimenting with innovative ways of financing community health programs that seek to crowd in additional funding for community health.[5]

1.2. Scaling-up RBF for Community Health

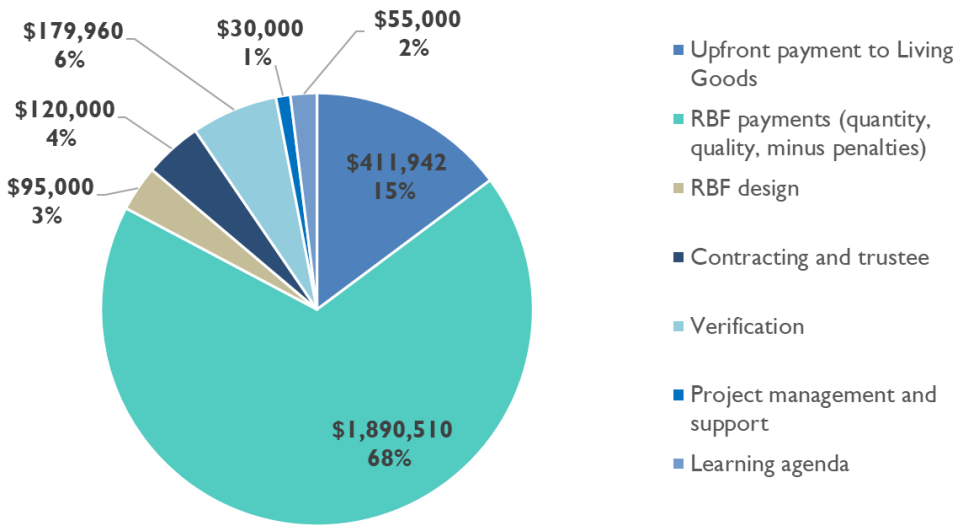
Following the successful implementation of a one-year pilot Results-Based Financing (RBF) program in 2018, funded by Deerfield Foundation in two districts in Uganda, LG decided to scale the RBF to three additional districts. The RBF scale-up program, which began implementation in October 2020 for a duration of 27 months, aimed to **drive improved cost-effectiveness and quality of community health services** targeting underserved and at-risk populations. The scale-up program was implemented in the pilot districts – Masaka and Kyotera – and three additional districts – Mafubira, Lira, and Wobulenzi, targeting approximately 2,000 CHWs. Building on the lessons from the pilot, the RBF scale-up design included a stronger focus on cost-effectiveness by placing greater emphasis on improvements in CHW productivity to achieve results (as opposed to increasing the number of CHWs). In addition, USAID DIV was interested in testing how the design could place a greater emphasis on quality-of-service delivery. As a result, the RBF design

[1] Capacity Statement Living Goods 2022: Living Goods supports digitally enabled community health workers to save lives at scale.
[2] Ibid.
[3] Ibid.
[4] Ibid.
[5] 2022-2026 Strategic Plan. Saving Lives at Scale through Country-Led, Digitally-Enabled Community Health Systems.

included three quality metrics (client knowledge, client satisfaction, and CHW competence) and four quality safeguards (unique households' coverage, CHW supervision, in-facility delivery, and PNC visits).

The scale-up program was structured as an Outcomes Fund, with USAID DIV as the anchor outcome payer (see Figure 1). USAID DIV made a financial commitment of USD 3 million, of which USD 2 million was conditional on meeting matching fund requirements. An additional USD 1 million was expected from other outcome payers/donors. Over the duration of the program, LG raised USD 412,975 from Deerfield Foundation for the RBF scale-up which unlocked an equivalent amount from USAID DIV (1:1 match). In addition, LG secured USD 2 million from a performance-based project LG co-developed with the Government of Kenya which unlocked another USD 1 million from USAID DIV for the RBF (2:1 match). In total, the financial commitment for the scale-up program amounted to USD 2.8 million.[6] Figure 1 shows the breakdown of the total cost by category.[7]

Figure 1. Total Costs

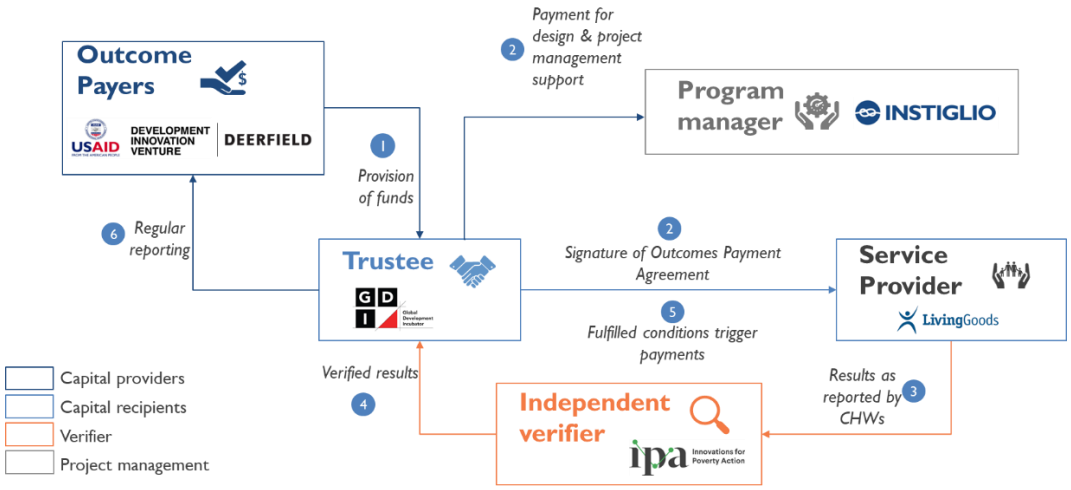


Other stakeholders involved in the scale-up program included: Global Development Incubator (GDI) who acted as the trustee, Innovations for Poverty Action (IPA) who was the Independent Verifier, and Instiglio who designed the RBF mechanism and supported implementation as program manager (see Figure 2).

[6] This includes 1 million from USAID DIV (no match), 1 million from USAID DIV (2:1 match), USD 412,954 from the Deerfield Foundation and an equivalent USD 412,954 in matching funds from USAID DIV (1:1 match).

[7] The 14% upfront payment to Living Goods covered expenses related to program design, preparation, smart health app design configuration, government advocacy, engagement, additional outcome payer engagement.

Figure 2. RBF Actor Relationships



1.3. Objectives of the Learning Agenda

The objective of this Learning Agenda was to evaluate whether and to what extent the RBF drove impact on maternal and child health outcomes and, by extension, whether RBF presents a pathway for improving the scalability and sustainability of results in community health worker programs.

There is an increasing pool of evidence around how RBF can improve health outcomes of children and women, especially when used with health facilities as the incentivized actor.[8] However, there is limited evidence on the effectiveness of RBF for interventions that deliver multiple health services at the community level. Gathering evidence and sharing lessons about when and where to apply RBF and the contextual and technical elements needed to successfully implement RBF mechanisms is therefore crucial for understanding how to best leverage the potential of RBF to deliver impact cost-effectively. In support of this goal, this report aims to contribute to the growing body of evidence on the effectiveness of RBF programs at the community level in improving health outcomes for women and children.

The report is organized as follows: Section 2 outlines the methodological approach used to assess the RBF scale-up program's effectiveness in enhancing cost-effectiveness and quality-of-service delivery.[9] This section comprises key research questions (Section 2.1), data collection and analysis (Section 2.2), and limitations of the methodology (Section 2.3). Section 3 presents the overall results and key insights gleaned from the analysis. Finally, Section 4 concludes with lessons learned, reflections, and recommendations.

[8] One example is the Health Result Innovation Trust Fund which currently supports 35 programs and 33 impact evaluation. See https://www.rbfhealth.org/impact and Bauhoff, S. & Glassman, A. (2017). Health Results Innovation Trust Fund at 10: What Have We Learned So Far? Center For Global Development, January 30, 2017. Retrieved from: https://www.cgdev.org/blog/health-results-innovation-trust-fund-10-what-have-we-learned-so-far..

[9] Quality-of-service delivery refers to the standard of care provided to clients by community health workers, which includes factors such as accessibility, effectiveness, safety, responsiveness, and equity.

2. Methodological approach

2.1. Key research questions

This Learning Agenda aimed to answer one overarching question: "Was the RBF effective in driving impact for the health of children under five and women of reproductive age?" To answer this question, the Learning Agenda evaluated the following:

1) Any changes in results, both those incentivized and non-incentivized,[10] that occurred in addition to any changes in data quality. All dimensions were analyzed in comparison to: i) baseline performance, ii) expected performance, and iii) performance in non-RBF districts (Research Question 1 (RQ1) in Table 1). The objective of this analysis was to identify the main changes in performance and provide the basis for analyzing whether these changes were consistent with expected behavioral patterns in response to an RBF (Research Question 2 (RQ2) in Table 1Considering concerns in the RBF literature for community health,[11] this question paid particular attention to any perverse incentives that might have been triggered by the RBF.

2) Which, and how much, of the changes in performance could be attributed to LG's actions (versus external factors), and, which of those actions may have been in response to the incentives provided by the RBF program (RQ2 in Table 1).

3) Which RBF design features promoted or constrained the observed changes in performance and how the RBF design could be revised to be more impactful (Research Question 3 (RQ3) in Table 1). This question also paid attention to how efficient the design features were at mitigating or preventing perverse incentives (e.g., introduction of quality indicators and safeguards) from materializing. The goal of this analysis was to provide important lessons on how to adapt the design to improve its impact in the future.

In addition, the Learning Agenda analyzed reviewed implementation processes of the RBF program and how those could be adapted to improve efficiency (Research Question 4 (RQ4) in Table 1. Learning Agenda research questions). The analysis looked at four process areas: i) the RBF design process, ii) verification and payment calculations, iii) governance structure and project management, and iv) trustee-held outcomes fund. The main objective was to identify lessons and recommendations to improve the efficiency of RBF programs in the future.

Table 1. Learning Agenda research questions

#	Learning Agenda research questions
RQ1	Did LG's performance change over the duration of the RBF program and what changes in performance were observed? Did the RBF design trigger any perverse incentives?
RQ2	Are the changes in performance attributable to the RBF program or to other factors? Are the observed changes in performance attributable to LG or to external factors (e.g., changes in disease environment or changes in government regulation)?
RQ3	Which design features promoted or constrained the observed changes in performance? How can the RBF design be adapted to be more impactful?
RQ4	Was the RBF program implemented in an efficient way? What lessons can be learnt from the implementation process?

2.2. Data collection and analysis

The Learning Agenda made use of both quantitative and qualitative evidence to answer the research questions. Qualitative data was used to validate, question, and complement findings from the quantitative analysis. For the quantitative analysis, the Learning Agenda used data routinely collected by (i) LG through the mobile application used

[10] Non-incentivized results refer to the outcomes that were not specifically incentivized or rewarded under the RBF program but were still important for the overall quality of health service delivery.
[11] See, for example, GiveWell (2019). A conversation with Dr. Madeleine Ballard of the Community Health Impact Coalition, April 23, 2019. https://docs.google.com/document/d/1OCGUMzXRpbMmwS3Eyl-yEvLYWu1yYpG8ykepvVlfSx0/edit

by CHWs as a job aid and data collection tool as well as (ii) IPA through its independent verification. The analysis leveraged LG's baseline data for the period June 2018 to November 2019 and RBF data for the period October 2020 to January 2023. A trend analysis was also conducted to ascertain whether there were performance disparities between RBF and non-RBF branches. The aim was to identify any performance changes that could be attributed to the RBF program, rather than organizational-wide changes.

For the qualitative analysis, key documents reviewed included: quarterly verification and payment reports; presentations from quarterly performance review meetings, where LG provided contextual information on performance; the RBF Data Generation and Sharing Guide;[12] LG's strategic plan and capacity statement; LG's RBF Annual Reports; a number of LG's proposals and explanatory notes (e.g., revising safeguards proposal, LG memo on unsynced data, and RBF COVID-19 recommendation note); the RBF action plan on addressing the error rate[13] based on the M&E consultant's recommendations;[14] the RBF design memo, and a number of emails and other minutes developed during stakeholder discussions. In addition, 16 semi-structured interviews were conducted with a range of staff from key stakeholders, including LG's Head and regional offices; IPA consultants and enumerators from Mafubira, Lira, Masaka, and Kyotera branches; USAID DIV, Instiglio, GDI and Deerfield Foundation. Additionally, two branch focus group discussions were held with twelve CHWs and twelve branch staff from Kyotera and Mafubira branches. Table 5 and Table 6 summarize all the stakeholders interviewed for this Learning Agenda.

2.3. Limitations of the methodology

This Learning Agenda did not employ experimental study designs, such as randomized control trials (RCT), to evaluate performance. As a result, its ability to generate evidence of attribution to the RBF mechanism is limited. While the study may have observed correlations or associations between variables, it cannot establish causation. Conclusions drawn from the analysis should therefore be understood as not being statistically rigorous in that regard.

Additionally, the analysis may not provide a precise estimate of the impact of the RBF program as it only assesses the outcomes achieved over the program's two-year duration, based on an evaluation conducted within three months of its conclusion. It therefore does not capture any longer-term impact that may manifest after the analytical period. This is a limitation of the analysis as the effect of certain changes implemented during the RBF program may only become visible over an extended period (see Section 3.5).

Finally, the qualitative insights presented in this report were based on a limited number of interviews which exclude some perspectives, most notably those of clients who received LG's services throughout the RBF program.

3. Overall results and key insights

As noted in Section 1.2, the main objective of the LG RBF scale-up program was to improve **the quality and cost effectiveness of community health services for underserved and at-risk populations**. The following section summarizes key findings regarding the RBF's achievement in these areas.

To improve the **quality of community health services**, the RBF design incentivized three metrics relating to the quality-of-services delivered: client knowledge, client satisfaction, and CHW competence. In addition, the RBF established a set of quality safeguards to measure dimensions of service delivery quality that the quantity metrics do not measure. These include coverage of CHW services, frequency of CHW supervision, and the continuity of care from pregnancy visits to in-facility deliveries to timely postnatal care (PNC) visits. Significant weight, up to 15% of outcome payments, was allocated to the quality metrics to incentivize LG to improve their performance in those areas. At the same time, to mitigate the risk of outcome payers paying for low quality results, a penalty of up to 21% would be applied if LG's performance on quality metrics fell below a certain minimum threshold (see Annex 3). In addition to the sliding scale of payments and penalties attached to the quality metric, quality **safeguards** were also introduced that would

[12] The Data Generation and Sharing Guide outlined the processes and considerations around gathering and sharing RBF data. The document was developed to combat challenges with sharing data experienced at the start of the RBF program by providing guidelines on gathering, analyzing, reviewing, requesting, and sharing RBF data.
[13] This refers to the proportion of sampled results that were unverified.
[14] An M&E consultant was contracted to support LG in reviewing their data quality issues.

penalize LG's payment by up to an additional 15% if performance against them consistently fell below a minimum threshold.

To improve **programmatic cost-effectiveness**, the RBF increased performance targets over time for selected metrics and provided strong incentives that targeted improved CHW productivity as the key driver of cost-effectiveness. This was designed as such to mitigate the risk of LG meeting targets through simply increasing the number of CHWs.[15] To support LG to meet the targets, the RBF scale-up was implemented over 27 months instead of just 12 months as had been the case for the RBF pilot. As in the RBF pilot, to promote both quality and cost-effectiveness, the RBF scale-up design included (i) relative prices (see Table 2)[16] to draw LG's attention to the most impactful metrics and (ii) metric-specific caps and price kinks to limit unnecessary visits and encourage improved productivity on all metrics (see Table 2).

Last, also in line with the RBF pilot, unverified results did not count toward the outcome payment and a data quality penalty for verification error rates above 10% was included to provide stronger incentives for LG to improve the accuracy of its data (see Table 2).[17]

Overall, LG earned only 65.3% of the total expected payments. As performance on quality metrics was consistently above target (see Figure 7) and penalties only had a marginal effect on the total payment (see Figure 10), lower than expected payment on quantity metrics was the main reason for the lower-than-expected payments. As shown in Figure 3, and explained in Annex 1, this payment loss was primarily driven by LG deploying fewer CHWs than expected due to challenges LG encountered in securing additional outcomes funding for the RBF. **In terms of CHW productivity, which was the main measure of performance focused on in the Learning Agenda's evaluation, LG achieved 93% of the expected target.** This was achieved despite the COVID-19 pandemic. The Learning Agenda focused primarily on CHW productivity as the main performance indicator as it was the main channel through which the RBF intended to drive impact and cost-effectiveness.

Based on quantitative and qualitative evidence the key insights highlighted in this report are as follows:

1. There is no evidence that the RBF led to improvements in CHW productivity (see Section 3.1).
2. The RBF program led to measurable, scalable, and sustainable improvements in the quality of programmatic data on CHW performance (see Section 3.2).
3. There is insufficient evidence to assess the RBF's contributions to the positive results observed on quality and safeguard metrics. This might be due to the inclusion of quality components in the verification of quantity metrics, which may have made it difficult to isolate the impact of the incentives on quality (see Section 3.3).
4. Evidence suggests that most design features neither promoted nor constrained performance. Instead, design features ensured outcome payers did not pay for results they deemed less valuable (see Section 3.4).
5. Although quantitative evidence indicates that the design and implementation of the RBF mechanism was not cost-effective – in the sense that it incurred costs that may not have seen a return on investment within the timeframe of the Learning Agenda – that assessment does not take into consideration data quality improvements and other benefits that fell outside the scope of this Learning Agenda to evaluate (see Section 3.5).
6. Despite concerns raised about the reliability of the independent verification approach, it was appropriate for its intended objective (see Section 3.6).
7. While the RBF mechanism was not designed or optimized for uptake by the Government of Uganda, the program did contribute to LG's advocacy efforts with the government by supporting to initiate engagements on potential strategies for improving the efficiency of their National Community Health Strategy (see Section 3.7).
8. The design process of the RBF scale-up benefited from a robust RBF pilot design, lessons gathered through a process evaluation, and stronger capacity built within LG (see Section 3.8).
9. The RBF scale-up design was robust and adaptable to the COVID-19 pandemic, which did not trigger significant changes to the design (see Section 3.9).

[15] This was achieved by capping the total number of CHWs that could contribute to the overall results on a quarterly basis.
[16] Relative prices were established based on their estimated impact on maternal and child mortality and the estimated level of effort to deliver results.
[17] The data quality penalty was phased in over the duration of the RBF for new districts and new metrics that were not included in the RBF pilot.

10. The qualitative assessment of the RBF program's governance structure and project management revealed that although stakeholders were generally satisfied, there were some areas for improvement (see Section 3.10).

3.1. Impact of RBF on CHW productivity

There is no evidence that the RBF led to improvements in CHW productivity.

One of the objectives of the RBF scale-up program was to improve the cost-effectiveness of maternal and child health services delivered by LG. This could have been achieved by either decreasing operating costs or increasing outcomes achieved, or a combination of the two. The main channel through which the RBF intended to drive cost-effectiveness was increasing the number of results delivered per CHW, in other words, their productivity.

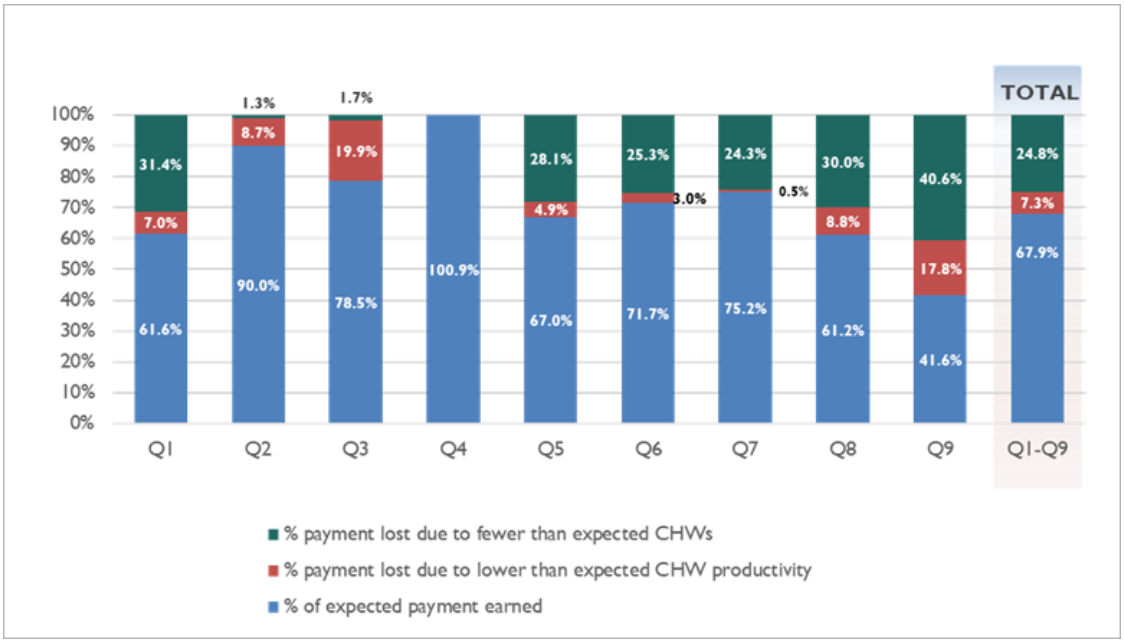
In terms of the first lever, qualitative interviews with LG suggest that there was no evidence that operating costs (overall or per CHW) decreased in the RBF districts. To the contrary, additional investments from both head office and branch staff were required to address the data quality issues raised by the RBF verification (see Section 3.2). In terms of increasing productivity, quantitative analysis shows that while LG maintained a consistent performance on CHW productivity, there was no improvement observed. The data presented below provides a comparison of CHW productivity in RBF branches to historical and expected performance, as well as a comparison to productivity in non-RBF branches. While the comparisons to historical and expected performance (see Annex 1. Quantitative analysis) show that LG's productivity was relatively good despite the COVID-19 pandemic, the comparison to non-RBF branches (see Annex 1. Quantitative analysis) suggests no significant variation in performance between RBF and non-RBF branches. This finding suggests that the RBF did not drive improvements. Other key findings that support this conclusion include:

1. LG achieved 97% of the historical (baseline) performance, however the baseline data was from June 2018 – November 2019 and did not include the period of the COVID-19 pandemic.
2. LG achieved 92.7% of the expected performance (targets) on quantity metrics. I.e., on average CHWs performed 7.3% below the targets across the various metrics (see Figure 3[18]). The new districts (district type 2[19]) were the main driver of lower-than-expected performance (see Annex 1, **Error! Reference source not found.**).
3. A trend analysis to check differences in performance between RBF and comparable non-RBF branches indicates no significant variation in performance in RBF branches compared to non-RBF branches (see Annex 1, Figure 17).

[18] Figure 3 shows the percentage of total expected payment lost on quantity metrics due to fewer-than-expected CHWs and lower-than-expected CHW productivity. The analysis, however, focuses on CHW productivity as this was the key performance driver that the RBF intended to influence. The number of active CHWs operating in the RBF program was, on the other hand, dependent on available funds. Difficulties experienced by LG in fundraising for the RBF resulted in LG not scaling up to the full expected number of CHWs.

[19] This refers to the additional branches included in the scale-up program, i.e., Mafubira, Lira, Wobulenzi.

Figure 3. Overall aggregate performance on quantity metrics



There are several reasons that could explain why an improvement in productivity was not observed. These include: (i) improvements are offset by the negative impact of other factors (e.g., COVID-19), (ii) efforts were made to improve productivity but were ultimately not successful (within the timeframe of the RBF), and (iii) the comparison to non-RBF branches does not show an improvement due to positive spillover effects from RBF to non-RBF branches.

Interviews with LG suggest that COVID-19 significantly disrupted operations, affecting CHW training and supervision and resulting in disruptions to workflows (see Section 3.2). However, qualitative evidence also suggests that LG's focus was not on improving productivity but was instead on understanding the reasons for the high verification error rate and addressing data quality concerns (see Section 3.2.). While some of the changes to address data quality were also rolled out to non-RBF branches (e.g., updates to the SmartHealth app) and may have positively contributed to productivity in those branches, the impact of these changes could not be conclusively measured in this Learning Agenda.

Besides the focus on the verification error rate and data quality, two additional factors may have contributed to the reduced attention on CHW productivity.

First, **stakeholders pursued multiple objectives with the RBF**, including improving cost-effectiveness, testing new design features to promote improved quality-of-service delivery, attracting more funding from donors, and engaging government. The multifarious nature of the RBF resulted in a relatively complex RBF design that included several features and drew attention to different factors. This, combined with the occurrence of COVID-19, and the shift in LG's long-term strategy, likely reduced LG's focus on and bandwidth to improve CHW productivity.

Second, during the first four quarters, **performance was reported at the aggregated level** (total results achieved by metric),[20] which made it more challenging for stakeholders, particularly outcome payers, to pinpoint whether lower than expected performance was driven by having fewer-than-expected CHWs or lower-than-expected productivity by CHWs. From Q5, Instiglio presented these two drivers of performance separately during the quarterly review meetings. Nevertheless, interviews suggest that this may not have been as impactful as the amount of information that was shared in verification and payment reports and quarterly review meetings may have been too much to absorb.

[20] Total results achieved per metric equals productivity per CHW multiplied by total number of CHWs.

3.2. Impact of RBF on quality of programmatic data

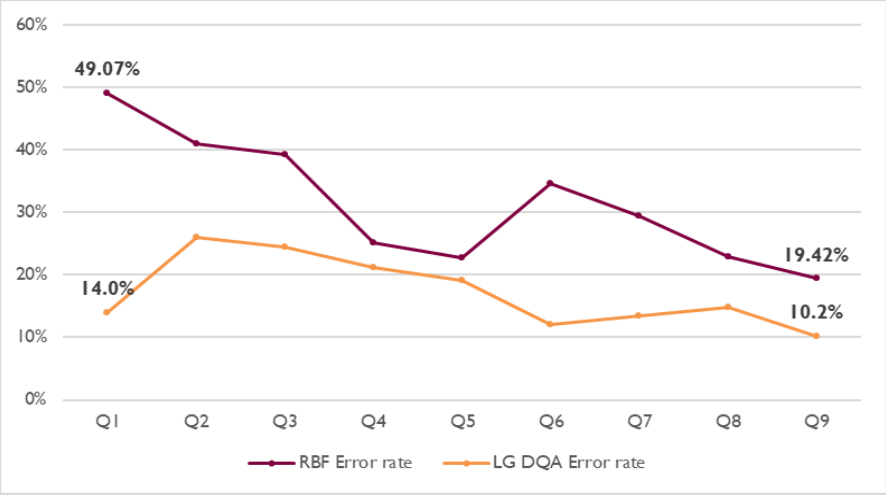
The RBF program led to measurable, scalable, and sustainable improvements in the quality of programmatic data on CHW performance.

The RBF verification revealed significant data quality issues, particularly during the first three quarters of the RBF with 49% (Q1) and 40% (Q2 and Q3) of LG's reported data deemed ineligible based on verification findings. In addition to impacting payments to LG, understanding the drivers of the high error rate was critical to LG as accurate data forms the basis for (i) generating reliable performance insights and informing programmatic course corrections and (ii) rewarding CHWs accurately. Reliable data is furthermore central to LG's credibility with donors and governments, and to the design and implementation of its Digitally enabled, Equipped, Supervised, and Compensated (DESC) approach which aims to improve CHWs' performance by incentivizing and enabling them to deliver high-quality health care at a low cost. In response, LG invested significant time and resources to understand the causes of the high error rate and resolve data quality issues which ultimately led to measurable, scalable, and sustainable improvements in the quality of LG's programmatic performance data.

While LG's internal DQA also identified data quality challenges (see Figure 4), the RBF data verification was able to draw attention to them, and their magnitude, quickly due to the following **key differences between the RBF verification and LG's DQA**:

1) **A higher level of disaggregation.** The RBF verification disaggregated error rates across nine quantity metrics instead of the five verified under LG's internal DQA. For example, the RBF evaluated U5 sick child assessments and U5 referrals separately while LG' DQA evaluated these as one metric thus making it difficult to identify data quality challenges driven by U5 referrals versus U5 sick child assessments. For example, in Q1 the error rate on the U5 sick child assessment was about 44.7%, whereas for U5 referrals, it was considerably higher at 83.9%. An investigation into this revealed a workflow issue caused by a protocol change during the COVID-19 pandemic, where CHWs were prompted to refer clients even when they were ineligible for referrals. CHWs accepted these prompts without realizing the implications, resulting in a significant number of inaccurate U5 referrals in the reported data, which contributed to the high error rate.
2) **Conditions for verification.** The RBF verification assessed more detailed conditions for a result to be verified. For example, the verification was not only concerned with whether a pregnancy visit was done but also probed on the specific services offered by CHWs during these visits e.g., whether during a pregnancy visit the CHW talked about the benefits of going to a facility for an ANC visit. Because of this, the RBF verification approach was more likely to count results as unverified and report higher error rates.
3) **Higher frequency of verification.** The RBF verification employed a two-week verification cycle which reduced the risk of recall bias and provided information on unverified results in a timelier manner. In contrast, LG conducted verification on a quarterly basis.

Figure 4. RBF error rate versus LG's internal DQA



Feedback gathered from LG indicates that the information generated by the rigorous verification unearthed the magnitude of the data quality challenges. Due to the financial and programmatic implications of these findings, a strong response was elicited from LG's management. This was further supported by the fact that this issue also attracted attention from outcome payers. For example, following the Q1 meeting, there was a request from USAID DIV for LG to outline strategies to address the verification/data challenges experienced in the first quarter, as well as for LG to provide regular updates on data challenges henceforth.

As a result, LG made changes to their DQA protocols (e.g., LG shifted to reviewing all 9 RBF metrics in their DQA, incorporated in-person verification to investigate cases flagged as unverified, and began conducting monthly data verification to investigate outliers) and developed a data quality optimization plan structured along three key pillars:

1) **Simplification:** Includes a series of reforms aimed at simplifying reporting tools and processes by supervisors and CHWs and focusing on better integration of workflows (e.g., LG integrated immunization with the U5 sick child assessment workflow and enhanced audits and automation).
2) **Process optimization,** including developing data quality indicators and corresponding dashboards that can be cascaded to all levels of the organization.
3) **Competency building for all staff** to ensure that everyone interacts with the data so that LG can improve quality of care. From stakeholder interviews with LG, it was indicated that the focus on data quality was already driving this as it resulted in an increased intensity in the use of data for decision making. Branch teams began requesting data reports which included information on CHW performance as well as outcomes from the independent verification. This enabled more targeted support towards CHWs such as in terms of supervision.

More specifically, LG employed several strategies to address data quality challenges. These included:

1) **Improved training and capacity building of CHWs to address knowledge gaps that, according to LG, were the main contributor to the poor quality of data.** Prior to COVID-19, LG held quarterly in-service trainings with all CHWs and monthly meetings between supervisors and CHWs to ensure they had the necessary knowledge and support to provide quality care to their communities. However, due to COVID-19, in-person engagements were halted. This created a challenge for CHWs whose workload had just been expanded to include the new workflows of family planning and immunization services, requiring them to adapt with decreased support as all activities moved to remote implementation. This resulted in some data quality challenges. As the pandemic waned, LG gradually resumed in-person support and training, and by the fourth quarter of the project, LG had implemented a more targeted approach to training and capacity building. This approach resulted in substantial improvements in the quality of trainings, as reported by Kyotera branch staff and CHWs, who noted a stronger focus on data quality than before. LG also started conducting general reorientations for branch staff, as well as branch-specific reorientations, to enhance supervisors' capacity to address recurring data quality challenges and improve the effectiveness of supportive supervision.
2) **Changes in the target-setting strategy for the Uganda LG program.** The Uganda LG program modified its target-setting strategy by shifting from program-wide targets to branch-specific targets that better reflected expected performance based on contextualized factors such as disease burden, population coverage, seasonality of issues, and historical performance. Feedback gathered during the interviews noted that program-wide targets, which had been set without taking into consideration the specific context at each branch, may have put CHWs under pressure resulting in some of them falsifying entries to meet targets. As a result, the shift to branch-specific targets served to reduce this pressure to some extent.
4) **Adopting stricter penalties at both CHW and supervisor level to motivate a stronger focus on data quality.** At the supervisor level, this aimed to shift attention towards focusing on quality-of-services provided by CHWs and ensuring CHWs had adequate guidance and support. At CHW level, stricter penalties, which also included possible terminations, aimed to ensure CHW compliance with the data quality agreements signed between LG and CHWs.[21]

[21] Despite knowing that their care for children should be focused on those under the age of 5 years, LG discovered that CHWs were treating children over 5 years. This resulted from multiple factors, including parents lying about the age of their children, CHWs feeling compassion for sick children, especially those they had cared for prior to their 5th birthday, and fear of retribution or tarnishing their image as a trusted care provider in their community.

5) **Updating SmartHealth app workflows to address challenges that contributed to the high error rate. Workflow challenges specifically affecting the U5 referral, ANC visit, and immunization metrics contributed to the high error rate.** For instance, between Q1 and Q4, the app experienced glitches, leading to CHWs being inaccurately prompted to make U5 referrals which they accepted but did not follow through.[22] While CHWs did not understand the implications of accepting these prompts, this resulted in unverified referrals that contributed to the high error rate. In Q5, LG successfully rolled out a solution that addressed this challenge.

In the case of the ANC metric, LG reported that updates made to the ANC workflow in Q6, which aligned the metric's tracking on the app to how it is verified, led to improved accuracy of reporting. [23] Similarly, resolving the challenges in the immunization workflow in Q4/Q5 ensured that CHWs received timely reminders to complete immunization visits. Both of these findings are also supported by the increases in the reported number of ANC visits per CHWs in Q6 and the number of immunization visits per CHWs in Q5, as shown in Figure 5 and Figure 6.

Figure 5. Number of ANC visits per CHW

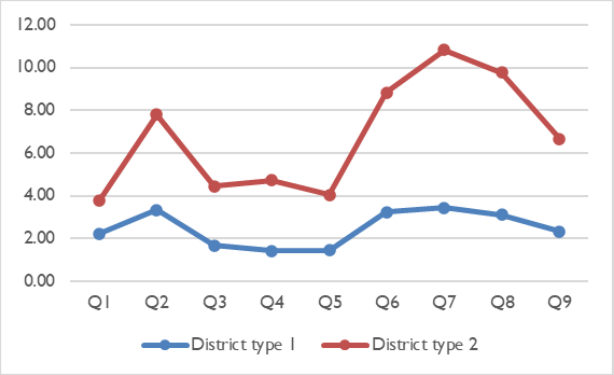
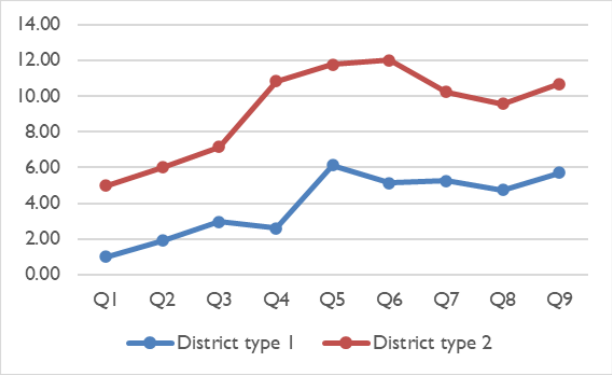


Figure 6. Number of Immunization defaulter referral and follow-ups per CHW



6) **Upgrading the SmartHealth App and revising protocols around syncing of data and hardware use to minimize errors.** One factor that contributed to the high verification error rate was the delayed syncing of data, which resulted in data being submitted long after a service was offered. As IPA verified results based on the date reported on the app, some results were unverified because of the mismatch between the app reporting date and the actual date the service was provided. Underlying reasons for delayed syncing of data included (i) server overload (e.g., resulting in approximately 29% of data not being submitted in Q1), (ii) limited cell service and (iii) CHWs lacking phones when they are taken for repairs or are non-functional, hence preventing CHWs from uploading data at the time the service was provided or in a timely manner.

To account for delays in syncing of data, an agreement was reached with the RBF partners that from Q2, IPA would consider for verification unsynced data if it synced within the next three verification cycles (a cycle is two weeks). Internally, LG implemented several actions to promote more timely syncing:

a) **LG instituted syncing windows – a period when CHWs were required to sync data – to prevent server overload and reduce the risk of unsynced data.** The version of the SmartHealth App that was being used in the first year of the project managed data generated through the app in an inefficient way. As a result, by June 2021, servers began to get overloaded, slowing data syncing. As a stop gap, LG instituted syncing windows with RBF branches given priority on this. At the same time, LG worked to increase server capacity as it planned, tested, and subsequently implemented an App upgrade across all LG branches, thus alleviating this issue.

b) **LG developed phone guidelines that outlined what a CHW should do if their phone was non-functional.** Concurrently, LG initiated phone pick-ups for CHWs operating in poor connectivity areas.

[22] RBF Annual Report – Year 2
[23] 170622_RBF Q6 result presentation

By Q3, these efforts ensured that approx. 97% of data synced within the last two verification cycles and by Q4, the error rate had significantly declined. The results of all these changes can be seen in the decline in the error rate between Q1 (49.07%) and Q9 (19.42%) where the error rate declined by approx. 60.4% (see Figure 4).[24] By the end of the scale-up program, the error rate closely resembled verification error rates observed during the RBF pilot, which had an average error rate of 18% (see Figure 9). Most of all, these changes were impactful and cost effective despite the high investment, because they are:

1) **Scalable and sustainable.** For example, interviews with LG suggest that the changes made to address data quality were scaled to other non-RBF branches and are viewed as sustainable as they have led to LG changing/revamping protocols e.g., adopting a data quality optimization plan, to ensure a continued focus on data quality.
2) **Anticipated by LG to improve quality-of-service delivery.** For example, by ensuring that CHWs always have access to a functional phone, this will ensure that they have access to workflows to guide them in service provision. As a result, this would minimize risks of incorrect diagnosis and ensure high-quality service provision.

3.3. RBF impact on quality-of-service provision

There is insufficient evidence to assess the RBF's contributions to the positive results observed on quality and safeguard metrics. This might be due to the inclusion of quality components in the verification of quantity metrics, which may have made it difficult to isolate the impact of the incentives on quality

One of the objectives of the RBF scale-up program was to incentivize LG to maintain a high quality-of-service delivery. The RBF mechanism was designed to achieve this result by I) incentivizing performance on three quality metrics (client knowledge, client satisfaction, and CHW competence) and ii) reducing payments earned through quantity metrics if performance on the quality metrics fell below a minimum threshold (see Annex 3). In addition, penalties were applied if a minimum threshold was violated on four quality safeguard indicators addressing coverage, supervision, in-facility delivery, and PNC visits. Based on quantitative evidence, performance on client knowledge and client satisfaction – the two quality metrics measured every quarter[25] – consistently exceeded targets (see Figure 7) with limited room for further improvement. Safeguard penalties were triggered only once on the in-facility delivery safeguard indicator in Q9 (see Figure 8).[26] While this was overall a positive outcome, there is limited evidence regarding the extent to which the RBF mechanism influenced this result.

The Learning Agenda revealed that the design of the verification of quantity metrics – which was dependent on evidence of quality-of-service delivery – may explain the limited variation of performance on the quality metrics. For example, for a quantity metric result, such as the number of pregnancy visits, to be verified, it had to meet two criteria. First, it had to be accurately recorded on a technological device and submitted on time. Second, it needed to be accompanied by evidence that the service was delivered in a quality manner, such as the CHW covering specific topics like the benefits of going to a facility for an ANC visit (client knowledge). As a result, performance on a quantity metric already included a quality component, which means that – by design – the RBF did, to some extent, incentivize quality-of-service delivery. However, because quality service delivery was implicitly incentivized by the quantity metric, it was difficult to assess whether the incentives attached specifically to quality metrics had an effect on quality. Moreover, it remains unclear whether the two metrics selected as measures of quality were, in fact, the best ones to use.

A key recommendation for how to improve the measurement of quality metrics is, therefore, to ensure that they are disentangled in the verification from quantity metrics (see section 4.5). This would generate better performance insights into quantity and quality as separate aspects of service delivery. Investing in further research to identify suitable ways to measure and incentivize quality-of-service delivery in RBF mechanisms should also be a priority (see Section 4.5).

[24] The increase in Q6 may have been due to the addition of Wobulenzi since it takes time for branches and CHWs to acclimate to the RBF.
[25] CHW competence, the third quality metric, measures the capabilities of CHWs active in the RBF program through an annual recertification exam administered by LG and monitored by IPA at the end of Q4 and Q8. Unfortunately, LG was only able to report results for the Q4 recertification exams with results for the Q8 exams delayed. In the first year, LG's performance on CHW competence was at 93%, an achievement of 98% of the expected target (95%).
[26] The in-facility delivery safeguard indicator was defined as percentage of women who delivered in a facility following a pregnancy visit. Every quarter, a penalty is applied if Living Goods' performance goes below the agreed threshold on a safeguard indicator and has received a warning in a previous quarter. LG had received a warning on the in-facility safeguard indicator in Q5.

Figure 7. Performance on client knowledge and client satisfaction

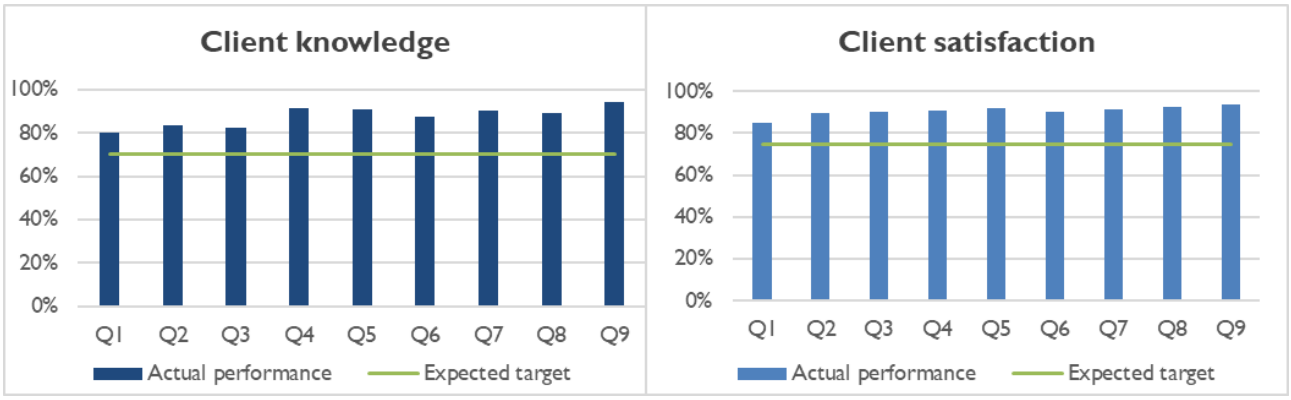


Figure 8. In-facility delivery safeguard

3.4. Impact of RBF design features on performance

Evidence suggests that most design features neither promoted nor constrained performance. Instead, design features ensured outcome payers did not pay for results they deemed less valuable.

The scale-up program included nine key design features outlined in Table 2. While all design features, excluding the quarterly payment cap and the renegotiation of targets, had an effect on LG's payment, based on qualitative and quantitative evidence, other than the data quality penalty, none of the other design features seem to have promoted a change in behavior or performance. On the other hand, there is no evidence to suggest that design features caused any perverse incentives (i.e., an overly strong focus on certain results) and their main contribution appears to be that they successfully mitigated the risk of outcome payers (over-)paying for results they deemed less valuable during the design of the RBF. For example, price kinks on the family planning metric were found to be effective in ensuring outcome payers did not overpay for performance on family planning.

Deep diving into the design features, there is, for example, no evidence that increasing targets over time or the impact premium included in the prices for ANC visits, in-facility deliveries, PNC visits, or U5 sick child assessments promoted a stronger focus on these results as even baseline targets on these metrics were not achieved (see Figure 15 and Figure 16**Error! Reference source not found.**)

Qualitative feedback from LG suggests that one reason design features did not incentivize performance could be because LG did not align incentives at the CHW and Supervisor-level with RBF performance targets. Furthermore, LG noted that responding to incentives provided by the design features was challenging without first understanding the sources of the high error rate and the uncertainty in estimating the expected quarterly error rates. This led LG to deprioritize

the focus on design features, such as metric specific caps and price kinks, as it was unknown how these would affect payment since they were contingent on the results considered for payment after discounting based on the error rate.

In conclusion, while design features did not drive a change in behavior, they mitigated risks that ensured value for money for outcome payers, as outlined in the table below.

Table 2. RBF design features

Design feature	Description	Rationale for including it
Quarterly payment cap	A payment cap is included every quarter, which is equivalent to 105% of the expected payment for this quarter. The quarterly payment cap will be increased by the aggregate sum of foregone payments, defined as the difference between expected payments and earned payments before the application of penalties.	Encourage performance while ensuring that the RBF program does not run out of money before the end of the program. Provide flexibility by allowing LG to make up for underperformance in one quarter with overperformance in subsequent quarters. This, however, also lowered the strength of the incentives to improve performance as LG was able to earn respective lost payments in subsequent quarters.
Metric-specific caps	Some metrics have caps to limit unnecessary results and reduce the risk that relative prices are not well calibrated. Metric-specific caps are typically set at the client level and/or significantly above the performance targets to provide flexibility.	Provide flexibility to respond to changes in the disease burden and client needs by allowing LG to choose how to allocate efforts across metrics. Encourage cost-effectiveness by preventing unnecessary visits and provide incentives to perform well on all metrics. Provide financial incentives to focus on the most impactful results by paying a higher price (per level of effort) for results with the greatest impact on maternal and child mortality.
Price kinks	Price kinks allow LG to earn payments for performance above the targets but the price decreases as more results are achieved so that it becomes harder to compensate underperformance on other metrics with overperformance on these metrics. Price kinks were included for U5 assessments and family planning visits as these metrics drive a large portion of the total expected payment and the metric-specific cap is set well above the expected target	
Relative prices	Relative prices for quantity metrics were established based on (i) the estimated level of effort to deliver the results and (ii) the estimated impact on maternal and child mortality.	
CHW adjustment factor	The outcome payers only pay for verified results achieved by a pre-agreed number of CHWs as defined in the CHW scaling plan. If more CHWs are hired by LG to conduct visits, payment tied to results will be pro-rated to the number of expected CHWs.	Encourage cost-effectiveness by incentivizing LG to increase the productivity of CHWs instead of achieving the total number of expected results by adding additional CHWs (which would increase cost).
Payment function linking quality and quantity	The payment function does not pay separately for performance on the quality metrics. Instead, performance on the quality metrics translates into a payment factor that increases or decreases the payment earned for performance on the quantity metrics (see Annex 3). Overall, the payment function	Encourage performance on quality-of-service delivery and avoid a situation where LG gets paid a lot for performance on quantity, even though quality is very low.

	ensures that 1. the payment/penalty for quality is proportional to the performance on quantity metrics and 2. for the same performance on the quantity metrics, LG receives a higher (lower) payment, the higher (lower) their performance on quality.	
Safeguards	Safeguards are intended to ensure that a basic standard of quality is being met across all the quantity payment metrics. Each safeguard indicator has a set minimum threshold where performing below the threshold first triggers a warning followed by a penalty with a maximum penalty cap of 15%.[27]	Encourage performance on quality-of-service delivery and prevent underperformance on some key quality indicators
Data quality penalty	To provide a stronger incentive for LG to improve the accuracy of its data, on top of not paying for unverified results, a penalty for each result that cannot be verified is applied (above a minimum threshold of 10%)	Improve data quality by ensuring that unverified results are not only not paid for but adding an additional penalty.
Renegotiation of targets	The renegotiation of targets leaves room for adjusting targets by taking into account the strong preference to set ambitious targets during the design phase yet also conscious of the uncertainty brought by the long duration of the RBF program which increases exposure to external factors thus increasing the uncertainty of setting realistic but ambitious targets.	Encourage performance by ensuring that targets and prices are set in a way that provides incentives for LG to improve performance.

3.5. Cost-effectiveness of the RBF mechanism

Although quantitative evidence indicates that the design and implementation of the RBF mechanism was not cost-effective – in the sense that it incurred costs that may not have seen a return on investment within the timeframe of the Learning Agenda – that assessment does not take into consideration data quality improvements and other benefits that fell outside the scope of this Learning Agenda to evaluate.

Cost-effectiveness is most often a measure of improved outcomes to cost ratio, yet it has many dimensions. According to its standard definition, quantitative analysis suggests that the RBF scale-up program was not cost effective since its outcomes – measured through CHW productivity at the time of analysis for this Learning Agenda – did not improve compared to either non-RBF districts or performance at baseline (status quo, see Section 3.1). There is also no evidence that operational costs decreased. To the contrary, the RBF's design, verification, and implementation all required additional cost. The estimated additional costs incurred to engage LG, GDI, IPA and Instiglio for the RBF design, verification, and implementation amounted to **USD 613,307**[28] and does not include significant additional time that USAID DIV (as well as other stakeholders) invested to manage the RBF and troubleshoot issues. This included:

- **RBF design:** USD 95,000
- **Program design and preparation:** USD 188,347 (this included costs related to the RBF design and smart health app configuration but also other activities to mobilize the program which are unrelated to the RBF)
- **Verification:** USD 179,960
- **Contracting and trustee:** USD 120,000
- **Project management and support:** USD 30,000

However, there is potential for increased cost-effectiveness in the long-term if the following factors are considered:

1) **Disproportionately high start-up costs:** A large proportion of RBF costs are fixed or are mainly incurred at the start of a program to design and roll-out the mechanism. With a longer RBF duration and larger scale, some of these costs (per result), such as those incurred by GDI, Instiglio, IPA would decrease due to economies of scale.

[27] See design memo for application of penalties for safeguard indicators.
[28] Additional costs related to the Learning Agenda or government engagement were not included here as not related to the RBF.

For example, verification costs per branch in the pilot program were approximately 40% more than verification costs per branch in the scale up program.

2) **Delayed return on investment:** The opposite is true of returns on investment, which can take time to yield. As explored in previous insights, there is strong evidence that the RBF supported improvements in programmatic data quality. While this did not lead to measurable improvements in CHW productivity, these improvements were highly valued by LG and provided greater visibility into factors affecting productivity or quality-of-service delivery which may lead to improvements in the future as follow-up actions come into effect.
3) **Positive spillover effects:** In addition to the delayed returns on investments, the RBF program had positive spillover effects on non-RBF branches. Specifically, the changes made to address data quality issues were scaled to other branches, which could lead to future improvements in performance.
4) **Unquantified benefits:** The RBF also generated other benefits that are harder to quantify. This includes other lessons experienced by LG throughout the implementation of the program or lessons documented in this Learning Agenda which can inform LG's engagement with government or use of RBF by other actors. Furthermore, an unquantified benefit of the RBF program for outcome payers was they only paid for verified results, thus reducing the risk of funding activities that may not result in the desired impact.

The learnings generated by this program did, nonetheless, highlight some potential ways in which the RBF related costs could potentially be reduced in the future, by, for example, streamlining the RBF contracting and governance structure or exploring ways to reduce verification cost. Section 4.4 further elaborates additional recommendations for improving cost-effectiveness.

3.6. Reliability of the independent verification approach

Despite concerns raised about the reliability of the independent verification approach, it was appropriate for its intended objective.

The goal of the independent verification was to verify results in a cost-effective manner,[29] leading to accurate payments for results achieved at the overall contract level.

In the initial quarters of the RBF program, the independent verification approach reported high error rates of 49% (Q1) and 40% (Q2 and Q3). This raised concerns from stakeholders over the reliability of the independent verification. More specifically, three main concerns were raised: (1) RBF verification not being sufficiently fit for context given error rates were much higher compared to LG's internal DQA process, (2) the sample size being too small and (3) the sampling methodology being biased. These concerns were possibly compounded by an imprecision in the RBF Design Memo, which cited improvements in data quality from 60% to 96% during the RBF pilot without clarifying that these improvements referred to LG's DQA process. This likely contributed to USAID DIV's expectations of much lower verification error rates during the RBF scale-up when in fact, verification error rates during the pilot were on average 18%.

As further elaborated on below, an analysis of the verification approach and these concerns revealed that:

1) The higher RBF verification error rates were likely driven by a greater rigor of the RBF verification, primarily in terms of what was considered as a verified result. This increased rigor was one of the benefits of the RBF as it drew LG's attention to data quality issues that the internal DQA process had not picked up on (see Section 3.2). The decision to use stricter requirements for a result to count as verified was, nevertheless, factored into the RBF design, specifically in setting targets and prices and in the application of data quality penalties.
2) The sample size was okay given its objective but too small in light of LG's desire to understand what was driving the high error rate.
3) The sampling methodology unlikely had an impact on verification error rates.

[29] For example, the RBF Design Memo states that "(…) many choices in the design of the verification process were deliberately made to allow for cost-effectively scaling this approach."

1) RBF verification error rates being higher compared to LG's DQA process

An analysis of the two methodologies suggests that the **main driver of the discrepancies were differences in the conditions for verification**. As driving quality of health services was a key objective of the RBF scale-up program, the verification methodology assessed more detailed conditions in order for a result to be verified. For example, the verification was not only concerned with whether a pregnancy visit was done but also probed on the specific services offered by CHWs during these visits e.g., whether during a pregnancy visit the CHW talked about the benefits of going to a facility for an ANC visit. Because of this, the RBF verification approach was more likely to count results as unverified and report higher error rates. The level of specificity employed by the independent verification approach could have also contributed to the error rate by increasing the risk of recall bias or respondents giving incorrect responses due to the script being longer as a result. Nevertheless, there were several mitigations put in place to account for these risks as outlined below:

a) For verification questions that probed on specific services CHWs offered, respondents were not required to recall all the services. A visit was verified if the respondents recalled at least one service offered.
b) The frequency of the independent verification also limited recall bias. If, for example, the RBF verified results on a quarterly basis instead, then this would be a much larger concern.
c) IPA conducted backcheck visits on at least 6% of the sampled results to ensure consistency between what was reported by the enumerators and the respondent's answers.

Concerns were also raised regarding the RBF verification protocol for family planning, pregnancy visits, and children under 5 metrics, which was seen as insufficiently adapted to the context. For family planning and pregnancy visits, LG noted that some respondents gave inaccurate responses due to discomfort, e.g., if a husband was present when the enumerator called. For children under 5 metrics, some results may have been unverified due to IPA enumerators not speaking to the caretaker at the time the CHW visit happened. Though these issues were mostly accounted for in the verification protocol,[30] minor adjustments were made to it in Q8. However, as the sample size in the last quarters was small, it was difficult to assess the impact of these changes.

Overall it is important to note that the RBF design had factored in that (1) the verification error rate would be higher than LG's DQA error rate (by factoring in an average error rate of 18% in setting targets and determining prices) and (2) LG was only penalized for error rates above 10% to account for results being unverified due to challenges that were not in their control such as incorrect responses from respondents or enumerator data entry mistakes.

Despite the mitigations put in place, one consideration for future verification as noted in Section 4.5, is to differentiate between verification of quality (i.e., how a service was provided) and quantity (i.e., was a service provided) components of a service. This can provide a clearer understanding of an implementer's performance on each of these dimensions and enable implementers gain a better understanding of the factors driving underperformance i.e., is it because actual productivity is low or is it due to poor quality-of-service delivery. Further, this could contribute to research on how to effectively measure and incentivize quality-of-service delivery, an area that the scale-up program did not conclusively assess (Section 4.5).

2) Sample size

Concerns regarding the sample size being too small were likely due to competing objectives. Due to the high RBF verification error rate, it became of interest for LG to understand which metrics were driving the high error rate to identify potential root causes. However, the **RBF's sample size** was by design not sufficient to provide a precise estimate at the metric-level to investigate root causes of the high error rate. For that purpose, a larger sample size calculated at the metric level would have been required. This, however, would have significantly increased the cost of

[30] Per IPA protocols, verification of sensitive metrics like family planning were treated carefully and with the required sensitivities. The enumerators only mentioned the metric being verified when it was confirmed that they were speaking to the primary respondent. Therefore, even if a family planning visit was being discussed over the phone, the husband would not know anything about the topic of conversation. For the children under 5 metrics, IPA typically treated such instances as replacement cases (see reasons for replacement from the Quarterly Verification and Payment Reports.)

the verification. As such, like the RBF pilot, to achieve a reliable estimate of the results at a reasonable cost, the sample size of the scale-up program was determined at the overall program-level. Specifically, the sample size was calculated to achieve a high precision of 2 percent with a confidence interval of 99%, meaning that the average error rate over the duration of the RBF should be approximately within +/- 2% of the error rate that would have been found if all results had been verified (as opposed to a sample-based approach). Even at the quarterly level, the sample size achieved a precision of 5% at a confidence interval of 95%.

3) Sampling methodology

Concerning the sampling methodology, the analysis determined that the sampling methodology unlikely led to biased verification error rates that impacted payments. The RBF's verification approach adopted a simple random sampling methodology to ensure that, on average, the sample at metric level was proportional to the reported data shared by LG, thereby ensuring accurate payment for results.[31] To reduce verification costs, one modification was made. In each of the 2-week verification cycles only two out of five districts were selected. Given the geographical dispersion of district type 1 and district type 2, one district per district type was chosen (i.e., either Kyotera or Masaka was chosen for district type 1 or Lira, Mafubira (Jinja and Buikwe), or Wobulenzi for district type 2).[32]Given the higher number of expected results in district type 2 over the course of the RBF, this methodology was by design going to lead to a small under-representation of district type 2. This was considered acceptable as error rates were expected to be higher in the new districts that had not operated under an RBF before. While this was indeed the case for Lira and Wobulenzi, Mafubira (the largest district) presented overall lower error rates.[33] However, overall, this is unlikely to have impacted the error rate as the difference between the weighted average error rate – taking into account the proportional contribution of districts to overall results (2) – is only slightly below the weighted average error rate considering the proportional contribution of districts to the sample (4) (30.9% vs 31.1%) (see Table 3).

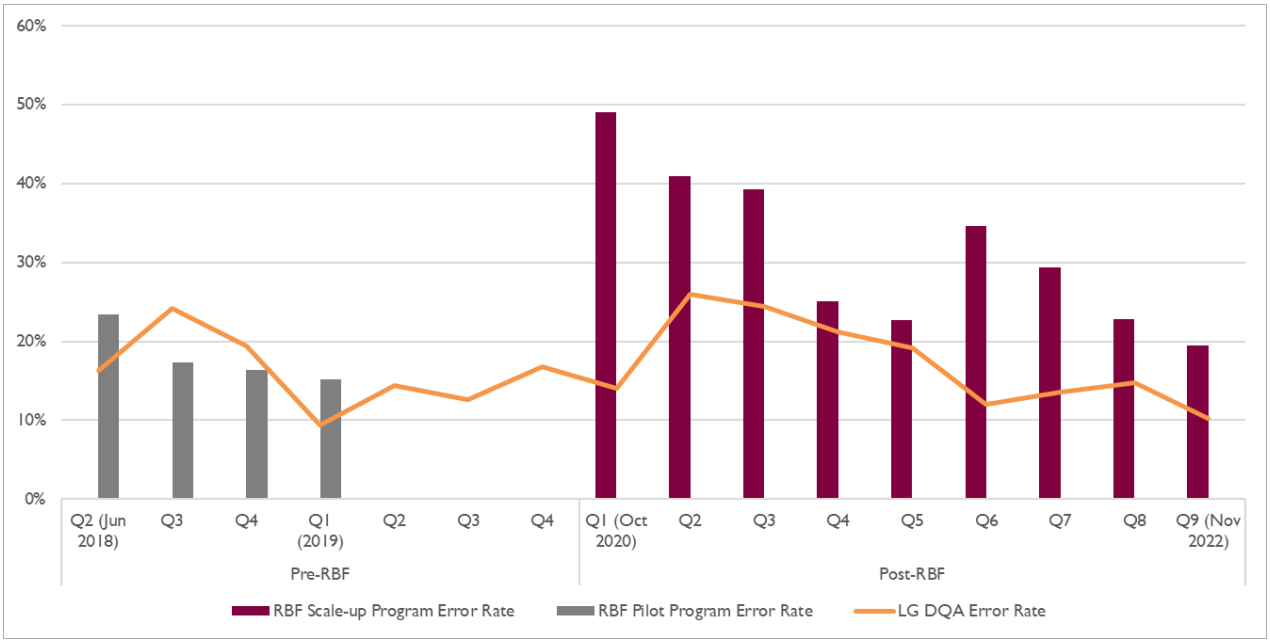
Table 3. Distribution of results by district and average error rater per district

	Masaka	Kyotera	Lira	Mafubira	Wobulenzi*
(1) Total results by district	234,371	274,062	280,381	494,529	118,574
(2) Percentage of total results	17%	20%	20%	35%	8%
(3) Total sample size by district	1020	753	612	944	343
(4) Percentage sample size by district	28%	21%	17%	26%	9%
(5) Average error rate per district	30.47%	29.04%	33.03%	27.84%	29.48%

[31] This means that metrics that are represented more in the LG data (the basis for payment) are also represented more in the sample, hence avoiding the perverse incentive that data quality is driven by certain metrics that are overrepresented in the sample.
[32] This sampling methodology reduces verification costs by limiting the geographical scope that enumerators need to cover within a 2-week cycle as well as travel between the two district types.
[33] Though this needs to be taken with caution since estimates of the error rate at the district level are less precise.

In conclusion, the verification approach was appropriate for its intended objective. Nonetheless, the concerns raised highlight useful lessons for the future. Particularly regarding: (a) the importance of ensuring greater alignment on the objectives and level of rigor of verification and being more explicit about the purpose of providing a breakdown of the sample size at, for example, the metric level; (b) being flexible to adapt the design to new or changing objectives and enabling this by setting aside budget to support adjustments or investments; (c) aligning on expected error rates more explicitly and providing more information on valid reasons for the error rate not to be 0% and, (d) ensuring survey instruments are sufficiently comprehensive to consider factors outside of the program's control.

Figure 9. RBF pilot and scale-up error rates compared to LG DQA error rates



3.7. Impact of RBF on government engagement

While the RBF mechanism was not designed or optimized for uptake by the Government of Uganda, the program did contribute to LG's advocacy efforts with the government by supporting to initiate engagements on potential strategies for improving the efficiency of their National Community Health Strategy.

An important goal of the RBF scale-up program, noted during stakeholder interviews, was to test the potential for elements of the RBF model to be adopted by the Government of Uganda (GoU). However, a clear theory of change on how the scale-up program would influence government adoption was not developed and the RBF design was also not developed for adoption by government. Nonetheless, the RBF program was useful in supporting LG to initiate discussions with the GoU on innovative approaches to improving the efficiency of CHW programs. By the second year of the scale-up program, LG had made substantial progress in promoting the RBF approach with the GoU, partly through leveraging LG's experiences implementing the RBF scale-up program. By Q4 of the RBF scale-up program, LG was asked to act as an advisor to the MoH RBF Unit and the technical working group (TWG), that was formed to lead RBF efforts.[34] LG's engagement strategy also contributed to the government's inclusion of elements of LG's Digitally Enabled, Equipped, Supervised and Compensated (DESC) strategy into Uganda's first national community health strategy. This is evidenced most notably by one of the National Strategy's Strategic Objectives which directly references LG's approach in its goal to operationalize "a performance management framework for CHWs using the DESC approach."[35]

[34] Draft_RBF Annual Report-Year 2.
[35] *National Community Health Worker Strategy (NCHWS), Strategic Objective 2.2.* Government of Uganda, Ministry of Health, February 2023

Some of the key lessons noted during stakeholder interviews that have facilitated LG's advocacy efforts with the GoU include:

1) Assessing performance through the use of data collected by CHWs through the digital app further reinforced the need for digitally enabling government CHWs. LG drew on this lesson to continue advocating for digitizing CHWs across the country.
2) Lessons learnt from LG's approach to providing performance-based incentives to CHWs have contributed to refining the implementation of a simplified DESC approach with an RBF component with one district in Uganda. In turn, learnings from this district model and a robust advocacy effort by LG has resulted in the inclusion of DESC as part of Uganda's first national community health strategy. Advocacy by LG for an RBF model for community health has also brought attention to the need to recognize CHW contributions when assessing performance in areas where RBF is used at the facility level.

3.8. Efficiency of the RBF scale-up design process

The design process of the RBF scale-up benefited from a robust RBF pilot design, lessons gathered through a process evaluation, and stronger capacity built within LG.

Qualitative evidence gathered from staff involved in both the design process of the RBF pilot and the scale-up program suggest that the scale-up design process was more efficient due to the following:

1) The scale-up program leveraged a lot of the RBF pilot design elements (e.g., payment metrics, targets, pricing methodology, verification methodology) and tools created (e.g., financial model).
2) Lessons and recommendations[36]gathered from the RBF pilot were used to refine the design of the scale-up program.
3) The scale-up program leveraged LG's organizational expertise and understanding of RBF, especially among key leadership, owing to their involvement in designing the pilot program and implementing it.

Despite the efficiency of the scale-up program design process, the integration of novel design elements, such as quality metrics and the addition of new districts (district type 2), required additional effort. For quality metrics, additional effort was dedicated to defining the metrics, developing the verification protocol, and determining appropriate targets due to insufficient historical data. With regards to new districts, variation in the performance levels increased the complexity of designing the financial model, determining prices, and determining how the CHW adjustment factor (see Table 2) would be applied. Overall, however, the scale-up design process was far more efficient than the pilot program.

3.9. Effect of COVID-19 on the RBF scale-up design

The RBF scale-up design was robust and adaptable to the COVID-19 pandemic, which did not trigger significant changes to the design.

The scale-up program was originally planned to start on March 31st, 2020, but was delayed due to the COVID-19 pandemic. The pandemic created uncertainty regarding LG's ability to meet performance targets because of anticipated disruptions to LG's operations. For example, the pandemic was expected to divert CHW attention to the COVID-19 response, decrease productivity due to illness or fear of infection among staff and CHWs, and delay LG's plan to scale-up the number of CHWs due to restrictions on gathering.[37] Following discussions with stakeholders, it was decided to delay the start of the RBF program. The program was eventually launched in October 2020, with the following small modifications, to account for the ongoing pandemic and the risks it posed on performance and the implementation of the RBF program:

1) To address challenges in scaling to the expected number of CHWs, outcome payers (i) approved the inclusion of an additional district (Wobulenzi), (ii) extended the timeline of the RBF program by one quarter, and (iii) approved

[36]Living Goods' RBF Pilot Internal Review; October, 2019
[37] Restrictions on gathering were expected to hinder the recruitment and training of new cohorts of CHWs.

a small modification to the CHW adjustment factor (see Table 2) to provide greater flexibility to LG[38] while still ensuring that the objective of this design feature was maintained i.e., that outcome payers only paid for results achieved by a pre-agreed number of CHWs as defined in the CHW scaling plan.

2) In addition, to mitigate payment risks for LG, the option of re-negotiating targets[39] (see Table 2) which had been included as an option at the end of Q6 was amended, thus providing LG the opportunity to renegotiate targets up to two times, between Q3 and Q7, conditional on providing compelling reasons for adjusting targets.
3) Lastly, the verification protocol was amended, shifting from a 50-50 split between in-person and phone verification to fully conducting verification over the phone up until Q6. The implication of this was that reported results that lacked a phone number could not be verified and therefore were replaced in the verification sample.

Ultimately, the minimal adjustments made to the RBF design in response to the COVID-19 pandemic indicate that the design was quite robust and could effectively adapt to the challenges posed by the crisis. It also reflects the stakeholders' commitment to find effective solutions that would uphold the program's objectives and ethos. This is evidenced by the fact that stakeholders opted not to have funds disbursed as a regular grant in the first two quarters – one option that was explored in the advent of COVID-19 – which would have reduced the time for testing and learning from the RBF. Instead, other solutions that still maintained the objectives of the program were explored and implemented.

3.10. Effectiveness of the RBF program's governance structure and project management

The qualitative assessment of the RBF program's governance structure and project management revealed that although stakeholders were generally satisfied, there were some areas for improvement.

Interviews with the stakeholders involved, revealed the following key insights regarding the RBF program's governance structure, project management, and reporting processes:

Stakeholders reported that **roles and responsibilities** of various stakeholders were clearly defined which facilitated efficient program management. However, since the program did not attract additional outcome payers beyond the Deerfield Foundation, the role of the trustee – which was created to streamline the contracting process through the involvement of additional outcome payers – was somewhat diluted. Another area that could have been improved was providing more clarity regarding **which decisions** were under the purview of the trustee versus that of the outcome payers.

Communications and reporting processes were generally deemed satisfactory. USAID DIV, for example, found the quarterly review meetings useful for gaining more insight into the context, progress, and challenges experienced by LG, which were not covered in the quarterly verification and payment reports. However, stakeholders held mixed views regarding the information provided in the quarterly verification and payment calculation reports. While LG appreciated the level of detail provided, such as metric-specific error rates, outcome payers sometimes felt overwhelmed with the amount of information presented and struggled to determine which issues to prioritize.

Collaboration and problem-solving: Similarly, stakeholders appreciated the high level of responsiveness from all parties, which included willingness to openly discuss challenges and collaboratively find solutions. For example, as noted in Section 3.9, stakeholders closely collaborated to find solutions to the COVID-19 pandemic. USAID DIV also provided LG with an M&E Consultant to support them in understanding the root causes of the high error rate. LG and IPA furthermore worked closely to both address challenges with the data sharing process[40] experienced at the start of the program and to update the verification protocol (see Section 3.6). Nonetheless, IPA noted the importance of ensuring that despite this collaboration, a sufficient level of independence or a clear protocol on what modifications need to be approved by GDI/outcome payers is maintained.

[38] The revisions in the CHW adjustment factor allowed LG to make up for fewer than expected CHWs in one district type with more CHWs in the other district type.
[39] LG had the option to renegotiate targets to account for performance risks that were beyond their control e.g., changes in external factors and updates in the assumptions used in the design phase to calculate ultimate targets.
[40] Initially, there was a lack of clarity regarding the process for sharing data for verification, leading to delays, incomplete or erroneous data. To overcome this challenge, a data generation and sharing guide was developed. In addition, a bi-weekly data sharing check-in was introduced to address any questions, concerns, or considerations related to the bi-weekly data shared by LG. These interventions led to a more streamlined and effective data sharing process between IPA and LG.

4. Lessons, reflections, and recommendations

Based on the analysis, this section presents some key lessons, reflections, and recommendations for RBF mechanisms.

4.1. RBF mechanisms can deliver value for money through their ability to accelerate learnings.

RBF mechanisms can function as incubators to enable organizations to **identify challenge areas, generate efficiencies and create positive spillover effects that can improve an organization's performance beyond a specific program.** For example, the RBF enabled LG to identify and address data quality challenges, which led to measurable, scalable, and sustainable improvements in the quality of programmatic data on CHW performance for RBF branches and non-RBF branches (see Section 3.2). While addressing these challenges was costly for LG, the changes implemented – for example the development and adoption of a data quality optimization plan – may result in a positive effect on both productivity and quality-of-service delivery or reduce service delivery cost in the long-term.

On the other hand, the design, verification, and implementation of an RBF all require additional cost and staff time of all parties that should be factored into cost-effectiveness calculations. For example, the estimated additional costs for the design, verification, and implementation of the RBF during the scale-up amounted to USD 613,307 – despite the efficiencies leveraged from the initial pilot (see Figure 3). These costs may, however, decrease over time as efficiencies are gained, improvements and simplifications to the RBF processes are made, and economies of scale are realized.

4.2. The complexity of RBF mechanisms is contingent on stakeholder needs and objectives.

A key learning from the scale-up program is that when designing RBF mechanisms, stakeholders should carefully consider how their objectives could impact the design and consider approaching objectives in phases instead of simultaneously to avoid the need for a complex and thus expensive mechanism. It is especially important for the objectives of a program to be aligned with the capacity of the implementer and the maturity of a program.

The RBF scale-up program pursued multiple objectives which led to a relatively complex design (see Section 3.1). For example, the design included several design features to drive cost-effectiveness and enable testing and learning about how to incentivize quality-of-service delivery (see Section 3.4). During the design stage, the complexity was deemed acceptable given the experiences and lessons from the RBF pilot and LG's high capacity. During implementation, however, its complexity may have made it harder for LG to identify how to increase performance, particularly when navigating unexpected challenges, such as COVID-19 and the high verification error rates.

With regards to the maturity of the RBF mechanism, during the pilot and refinement phases of an RBF, the emphasis for stakeholders may be on generating a broad base of learnings and knowledge on the mechanism, testing different modalities while comprehensively mitigating key risks for multiple parties. These goals can drive up both the complexity and cost of a design. The focus of learning and evidence-generation may then become more targeted as a program matures and trust in programmatic data is strengthened, reducing the amount of rigor required and subsequently reducing the number of design features that an RBF needs to include. In such cases, more reliance can be put on complementary management strategies that are less costly to implement, for example relying on less frequent audits of data and data systems instead of resource-intensive periodic verification approaches.

4.3. RBF mechanisms should seek in their design to balance the risks of underpayment to the service provider with the risk of overpayment by the outcome funders.

While a key objective of RBF mechanisms is to mitigate against the risk of outcome payers overpaying for results, a crucial goal is to also ensure that service providers are paid fairly, i.e., that the risk of which is assumed by service providers, is mitigated. Many services providers may be unaccustomed to taking on the risk that they may not recover all, or a substantial portion, of their costs if they do not meet quantity or quality targets, or if there are challenges with verification. As most verification approaches rely on service provider data, RBF designers and service providers need to assess and factor in potential inaccuracies in service provider's historical data and other reasons why the verification approach may not verify a result (e.g., recall bias of respondents) when setting performance targets and assessing payment risks. The RBF scale-up program considered this, for example, by factoring in the verification error rates of the RBF pilot when setting targets/prices and ensuring that LG was not penalized for error rates below 10%. However,

a crucial lesson is that this needs to be clearly communicated to all stakeholders to ensure that they sufficiently inform decision-making and achieve their intended effect.

4.4. Different strategies should be explored to improve the cost-effectiveness of RBF verification at scale.

One of the main costs of an RBF mechanism is often its verification mechanism, which is crucial for generating robust results evidence to calculate payments. In the case of the RBF scale-up program, the cost of verification accounted for 6% of the overall expenses of the program (see Figure 1). Inherent in the design of a verification mechanism is the trade-off that almost always has to be made between the cost of verification and its rigor. While stakeholders during the piloting, testing, and refining phases of an RBF mechanism may be willing to invest more in verification in order to maximize learnings and meet the needs of low-risk appetites, these costs can become prohibitive as an RBF scales. As a result, identifying alternative effective strategies that reduce the cost of verification methodologies as a program matures is crucial for the sustainability of an RBF design. Some strategies that could be considered include:

1) Leveraging technology such as GPS mapping to confirm that CHWs visited a household or locate survey respondents, which could reduce verification costs by minimizing the effort required to find respondents when doing in-person verification.
2) Incentivizing CHWs to add and repeatedly verify phone numbers for all clients to allow for greater reliance on phone-verification which is less costly.
3) Auditing the implementor's verification approach (through system, process, and data assessments and/or random spot checks) to ensure it meets a minimum standard so that the implementors' own administrative data can be relied upon to calculate results payments. This can be an interesting strategy when implementers already have strong internal controls and data quality processes in place and are determined to be operating at a high capacity.

4.5. Understanding how to measure and incentivize the quality of performance is an area that requires further research.

One of the objectives of the RBF scale-up program was to test and learn about how to incentivize quality-of-service delivery. However, the design of the verification of the quantity metrics, which to some extent ensured that only 'quality' results were paid for (see Section 3.6) may have resulted in the strong performance on quality, but limited evidence on RBF's role in influencing this performance (see Section 3.3). Consequently, this means that the program generated less evidence regarding whether quality metrics were well designed and effective (or could be effective) in motivating a focus on quality. The following are recommendations for how future programs could improve how quality performance is measured and incentivized:

1) The verification of quality components of a program (how well a service was delivered) should be disentangled from the verification of quantity metrics (how often a service was delivered). This could mean, for example, that the verification of quantity metrics only considers whether a visit happened, while any details regarding what was discussed during the visit or client satisfaction would be assessed separately and not influence the analysis of the quantity metric. This may provide greater performance insights into quantity and quality aspects of service delivery.
2) Invest in further research to identify ways to measure and incentivize quality of service delivery in RBF mechanisms that accurately reflects improvements while managing the risks of over- or under-paying a service provider. Most RBF approaches in global health focus payment metrics on the delivery of services (i.e., activities or outputs) as opposed to the outcomes of those services, such as mortality and morbidity, despite outcomes potentially being an important measure of the quality of services delivered. This is because it is often prohibitively costly, particularly at scale, to measure such outcomes. In addition, outcomes are usually further along in the results chain from specific CHW activities that it becomes challenging to confidently establish causation for any improvement in outcomes observed. For example, an increase in in-facility deliveries or vaccinations completed depends on an action to be taken by the person receiving care, which falls outside of the manageable control of CHWs. Placing a payment metric on that type of outcome as a way of incentivizing quality would therefore introduce a risk that the service provider may not be paid despite delivering a high-quality service. On the other hand, incentivizing outputs or activities also presents risks as the outcomes of health services arguably matter as much or more than the volume of health services delivered and should therefore also be considered. Figuring out this dichotomy is therefore critical for ensuring that an RBF can be effectively used and scaled.

4.6. Understanding how RBF governance structures can best establish clear protocols for decision-making while enabling flexibility and collaboration is an area that could benefit from additional research and development.

Stakeholders' commitment to learning and a governance structure that fostered collaboration and collective problem-solving was critical to the success of the RBF (see section 3.10). However, too much flexibility can also lead to inefficiencies or undermine the integrity of the RBF by focusing stakeholders' attention on what adaptations of the RBF mechanism to make instead of the core objective of improving performance. To mitigate these risks, clear protocols should be established to identify which issues merit conversations on adapting the RBF design and who should approve which modifications.

While issues are often difficult to anticipate, stakeholders can and should use learnings from a 'pilot' or 'test/refine' phases of an RBF mechanism to inform protocols for at scale implementation when arguably, efficiency becomes more important. In addition, the space would benefit from more best practices regarding governance structures and protocols that are efficient while not undermining flexibility/collaboration.

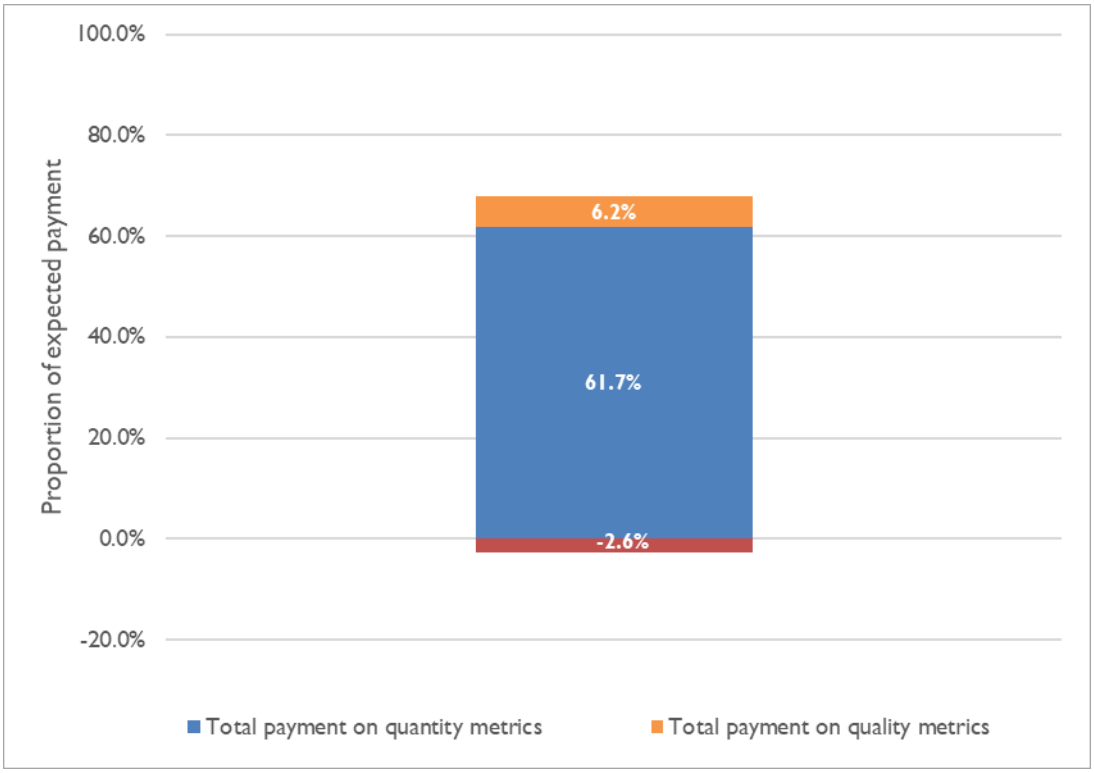
Annex

Annex 1. Quantitative analysis

Comparison to expected performance

Performance in the RBF, measured using total actual payment earned aggregated across the different metrics, was a function of three factors: (i) achievement on quantity metrics, (ii) achievement on quality metrics, (iii) penalties from unverified results and from violating safeguards. Overall, LG earned only 65.3% of the expected payment, which includes, 61.7% of payment earned on quantity metrics, 6.2% of payment on quality metrics less penalties of 2.6% incurred on safeguard metrics and data quality (see Figure 10). As performance on quality metrics was consistently above target (see Figure 7) and penalties only had a marginal effect on total payment, lower than expected payment on quantity metrics was the main reason for the overall underperformance.

Figure 10. Composition of actual payment as a share of total expected payment



Payment on quantity metrics aggregated (i) productivity per CHW, (ii) number of CHWs operating in the RBF, and (iii) the price per metric.[41] In this report, productivity per CHW has been used as the main indicator to evaluate LG's performance in the RBF, as well as the RBF's impact on performance, as it was the main factor the RBF intended to influence. Hence, the conclusion in Section 3.1 that LG achieved 92.7% of the expected target on quantity metrics (based on Figure 3. Considering all the factors impacting payment, as described above, LG earned on aggregate 67.9% of the total expected quantity payment[42], with Q4 being the best performing quarter and Q9 the worst performing (see Figure

[41] Since prices are constant, any variations between actual and expected payments would be driven by differences between actual and expected productivity per CHW and/or differences between the actual and expected number of CHWs. This means that actual payment would increase if CHWs were more productive (keeping the number of CHWs constant) or if the actual number of CHWs was higher than expected (keeping productivity constant), and vice versa. In practice, while both factors can result in lower-than-expected payment if below expected levels, in the reverse, LG can only exceed expected payment by overperforming on productivity. This is because the RBF was designed to encourage LG to improve CHW productivity.

[42] The 32% lost on quantity payment is equivalent to USD 906,742. This value is based on the total expected payment which was modelled under the assumption that LG would raise all the matching funds.

11). On average, district type 1, Kyotera and Masaka, performed significantly better than district type 2,[43] with targets being met in three quarters and an average target achievement of 75.6% compared to 64.1% in district type 2 (see Figure 11).

Figure 11. Total actual payment on quantity metrics compared to expected

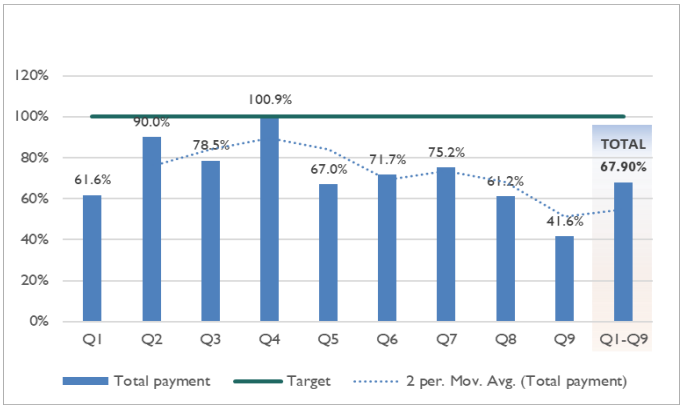
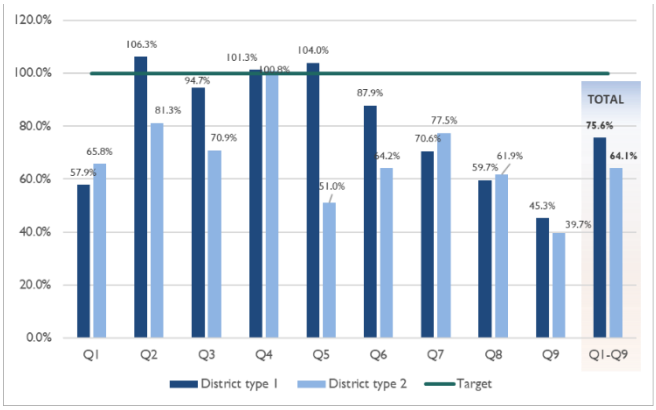


Figure 12. Actual payment compared to expected by district type



Having fewer CHWs than anticipated rather than low CHW productivity was the main driver of payment losses in both districts, with most of the payment losses occurring in the second half of the RBF (see Figure 13 and Figure 14). LG had fewer-than-expected CHWs in this period due to a decision not to scale up the number of CHWs to the expected level due to uncertainties in securing the targeted financial commitments from additional outcome payers and thus avoid the risk of being unable to sustainably deploy CHWs beyond the duration of the RBF.

Nevertheless, while fewer CHWs than expected was the main driver of payment losses in both district types, 22.8% in district type 1 and 30.5% in district type 2, low CHW productivity also contributed to some payment losses, especially in district type 2. On average, underperformance due to low CHW productivity in district type 2 was more than double underperformance due to productivity in district type 1 (see Figure 13 and Figure 14).

Overperformance in Q4 across both districts was attributed to the inclusion of unsynced cases resulting from a glitch in the app experienced in Q3. Approximately, 11,000 additional cases were submitted for payment in Q4.

[43] This refers to Lira, Mafubira, and Wobulenzi.

Figure 13. District type 1: Overall aggregate performance on quantity metrics

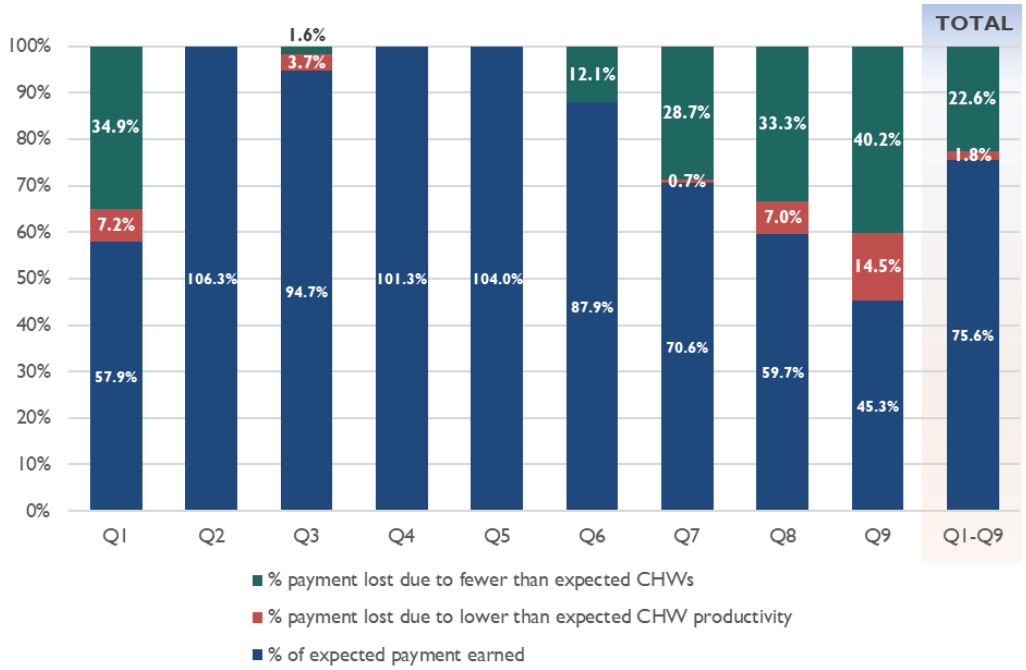
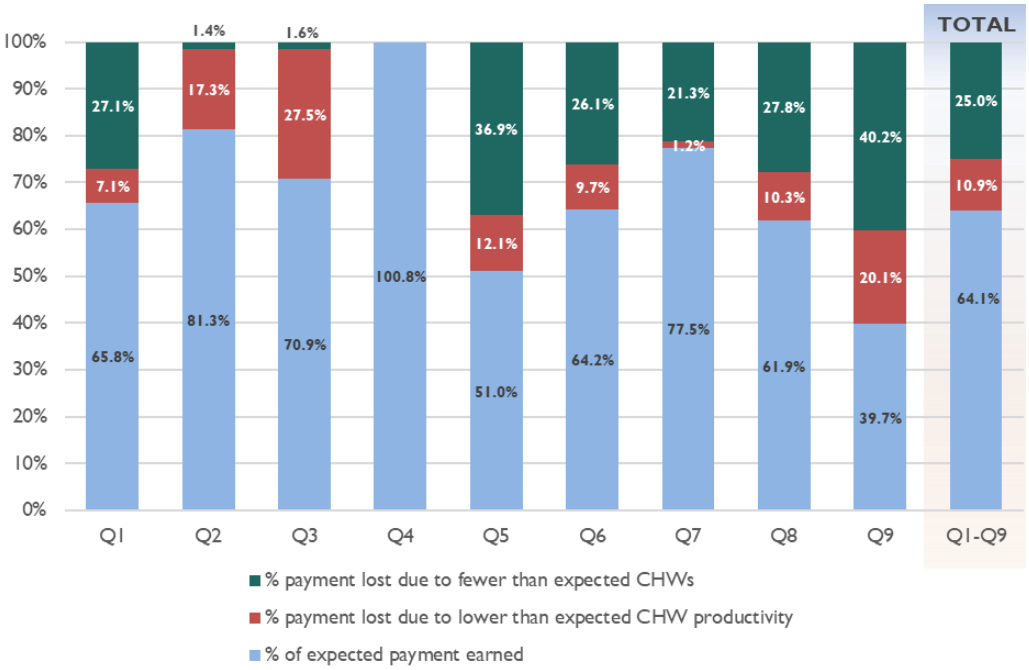


Figure 14. District type 2: Overall aggregate performance on quantity metrics



Comparison to historical performance

This section compares average productivity in the RBF from Q1 to Q9 to productivity in the baseline period (June 2018 to November 2019 for most metrics, October 2019-November 2019 for the immunization metric, and March 2019-November 2019 for the family planning metric).

On average there was significant variance in performance across metrics compared to historical performance (baseline) with most metrics, particularly the maternal and child health metrics, failing to reach baseline targets (see **Error! Reference source not found.** and **Error! Reference source not found.**).

Figure 15. District type 1: Comparison of actual performance

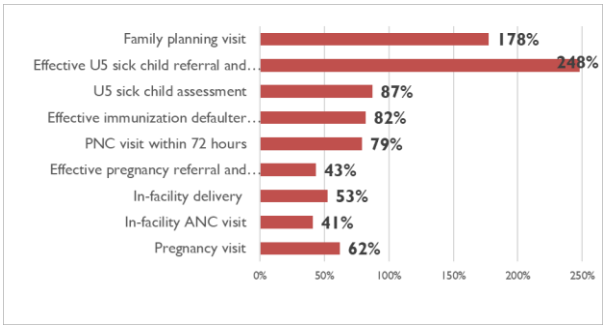
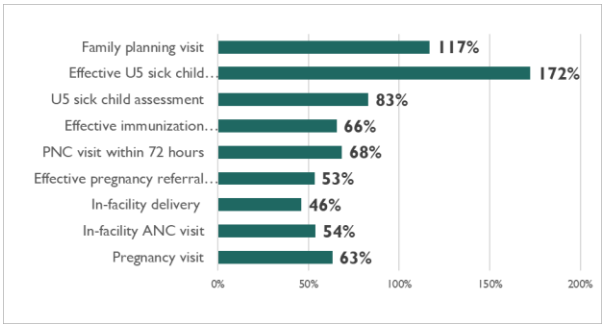


Figure 16. District type 2: Comparison of actual performance



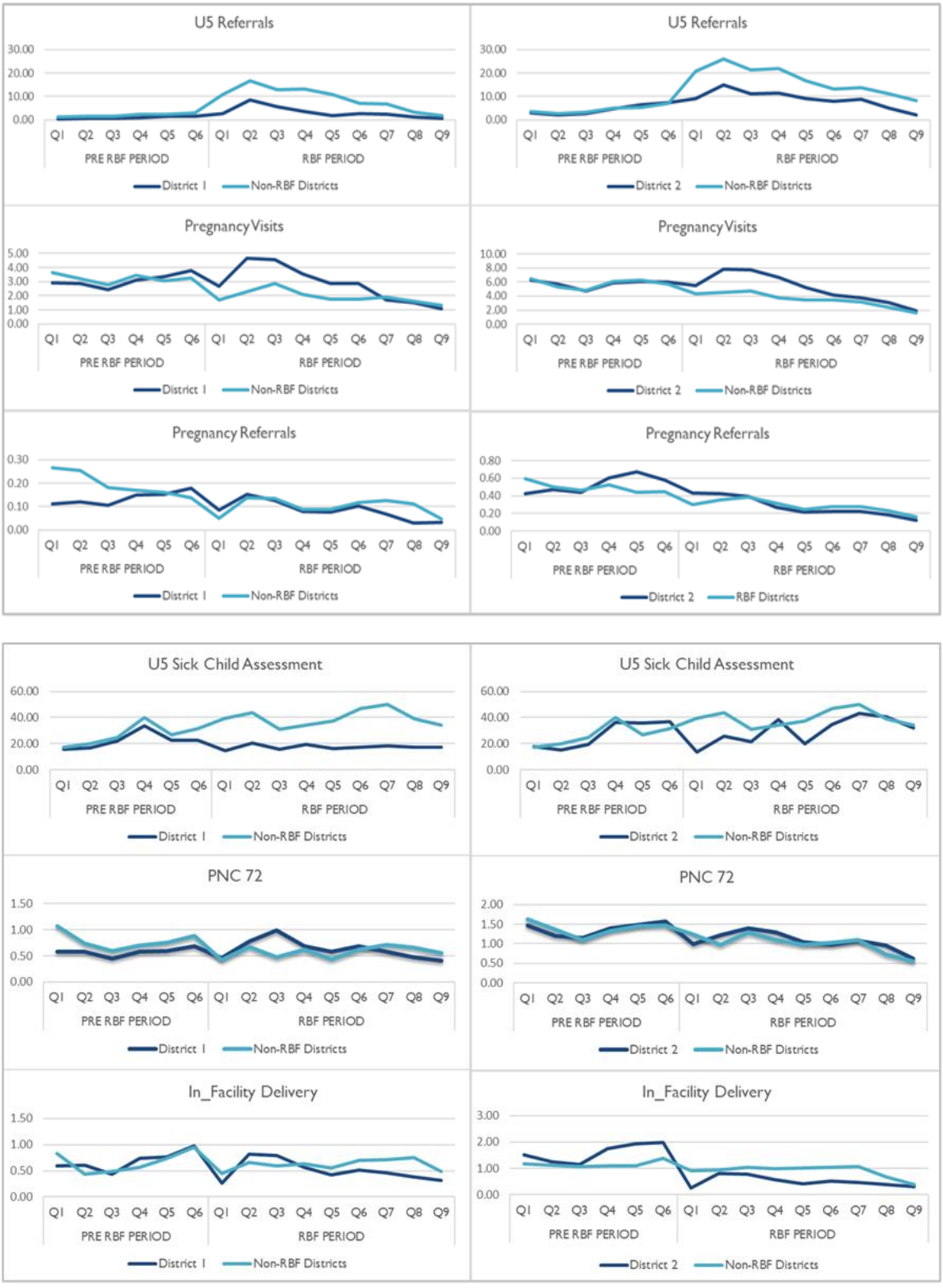
Only family planning and U5 sick child referral met or surpassed baseline targets. Performance on these metrics helped offset some of the losses on the other metrics. However, this performance could be due to targets not being ambitious enough due to having (i) limited historical data to set targets for the family planning metric[44] and (ii) less experience implementing both metrics – the U5 referral metric was not measured in the same way during the pilot.

Comparison of performance of RBF branches to non-RBF branches

To identify whether the RBF program affected productivity, a trend analysis was conducted to check whether there were differences in performance between RBF branches and non-RBF branches that performed similarly in the pre-RBF period (June 2018- November 2019). On average, the trend analysis showed no significant differences in performance (see Figure 17).[45] This could mean that the RBF did not motivate improvements in performance or that there were positive spillover effects to non-RBF branches. Based on qualitative insights, the former reason seems more likely as during implementation, LG's focus was on understanding the reasons behind the high verification error rate and addressing data quality concerns rather than improving productivity (see Section 3.2).

[44] LG had just rolled out this service with historical data covering the period March 2019-November 2019.
[45] Some variance in performance was seen on the U5 referrals metric for both district types though unclear what was causing this.

Figure 17. Performance trend-lines for RBF branches and non-RBF branches on various quantity metrics



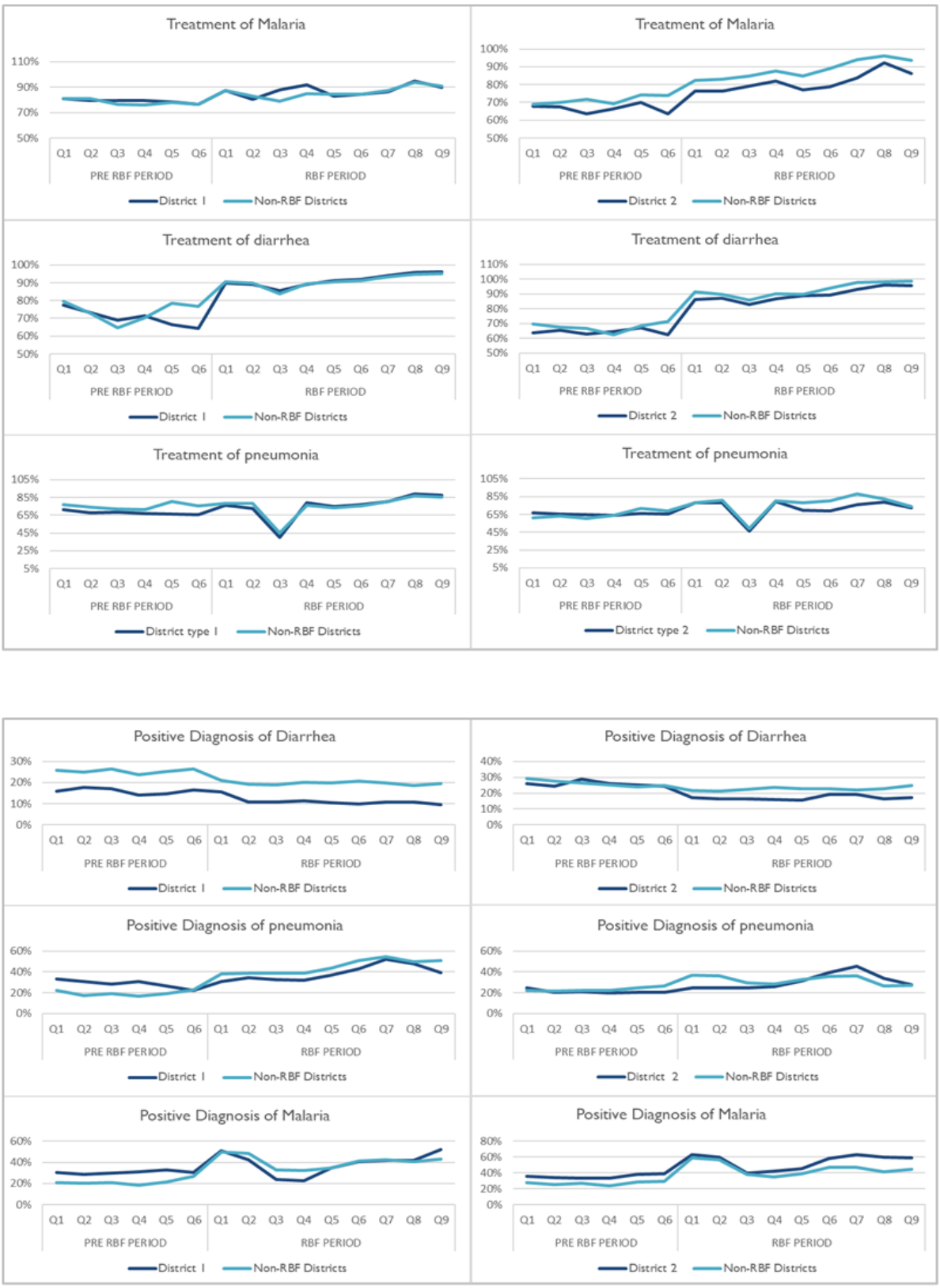
Performance on non-incentivized results

Non-incentivized results refer to results that do not directly have financial incentives tied to them in the RBF program. Analyzing performance on non-incentivized results was done to measure the broader impact of the RBF program, including any positive or negative side-effects that occurred during the RBF program or might occur in the future (sustainability).

For the non-incentivized results measuring (i) percentage of positive diagnosis of malaria, pneumonia, and diarrhea as a share of total number of U5 assessments and (ii) percentage of treatment provided as a share of total number of U5 assessments resulting in a positive diagnosis, performance in the RBF districts was compared to performance in non-RBF districts to see if there were improvements in the quality-of-service delivery in RBF districts. However, results, based on the trend analysis (see Figure 18) show that there were no major differences in performance between RBF and non-RBF branches.

Further, an analysis of CHW attrition shows that attrition has been relatively low except for the spike in Q9. No reason was given for the spike in attrition in Q9.

Figure 18. Performance trend-lines for RBF branches and non-RBF branches on non-incentivized results



Composition of services

An analysis was also done to assess changes in the composition of services offered by CHWs in response to the RBF incentives. This analysis aimed to assess whether LG and CHWs prioritized some services in response to the strength of the relative incentives or other factors such as changing disease burden. The analysis measured the percentage of services provided as a share of total services provided. Table 4 highlights periods when there were major changes in the composition of services from the previous period. Most changes coincided with periods when there was: (i) reported syncing issues for U5 assessment (Q2), (ii) reported workflow issues[46] for family planning (Q4) and immunization (Q3-Q4), and (iii) when major workflow updates were completed for immunization (Q5-Q6) and ANC visits (Q6) metrics. The increase in immunization and family planning services in Q2 follows the period after LG had fully rolled out these services.[47]

Table 4. Percentage of individual services as a share of total services provided

Payment metric	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Pregnancy visit	7.3%	7.9%	10.6 %	9.0%	8.4%	6.1%	4.6%	4.6%	4.0%
In-facility ANC visit	2.2%	3.1%	2.1%	1.8%	1.6%	4.1%	4.4%	4.4%	3.8%
In-facility delivery	0.5%	1.3%	1.7%	1.6%	1.4%	1.1%	1.1%	1.2%	1.2%
Effective pregnancy referral and follow-up visit	0.5%	0.4%	0.5%	0.3%	0.3%	0.3%	0.3%	0.2%	0.2%
PNC visit within 72 hours	1.3%	1.3%	2.0%	1.8%	1.7%	1.5%	1.4%	1.5%	1.4%
Effective immunization defaulter referral and follow-up visit	0.5%	0.9%	0.5%	0.3%	0.8%	1.5%	0.9%	0.6%	0.5%
U5 sick child assessment	64.2 %	43.6 %	39.4 %	50.1 %	51.2 %	53.4 %	54.1 %	57.0 %	65.2 %

[46] In these periods, the app failed to send reminders for follow-up tasks.
[47] The evaluation was unable to establish reasons for changes in the composition of services for metrics such as pregnancy visit, in-facility delivery, and pregnancy referral as well as the decline in the composition of services for immunization and U5 referral in the last quarters of the RBF program.

Effective U5 sick child referral and follow-up visit	10.6 %	15.5 %	15.1 %	14.8 %	12.0 %	10.3 %	10.1 %	7.0%	4.0%
Family planning visit	13.0 %	26.1 %	28.1 %	20.3 %	22.6 %	21.7 %	23.2 %	23.4 %	19.7 %
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%

	>20% decline in the composition of services
	>20% increase in the composition of services
	>75% increase in the composition of services

Annex 2. Stakeholders interviewed and branch focus group discussions

Table 5. Stakeholders interviewed

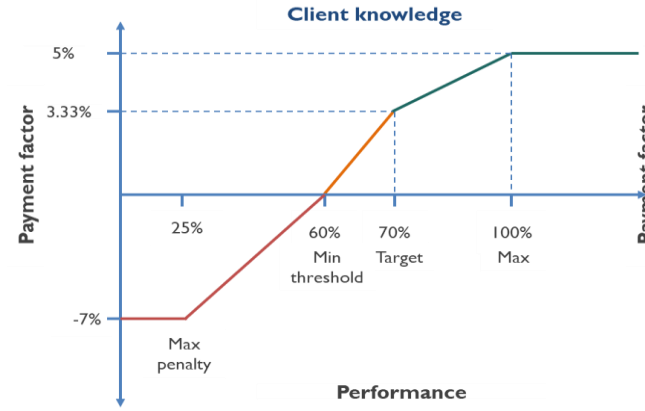
Organization	Name	Title/Role	Interview date
Living Goods	**Sarah Riczo**	Senior Manager, Business Development,	February 24
	Edward Zzimbe	Global Director Program Strategy and Excellence	March 2
	Afra Nuwasiima	MEL Manager	March 1
	Grace Nakibaala	Innovations Manager	February 17
	Monica Mugisha	Senior Manager, Digital Health	March 16
	Evelyn Kusiima	Head of Direct Ops	March 1
	Catherine Namutaawe	Quality Assurance Coordinator	March 1
	Amy Kakiza	Director of Partnership, Advocacy and Communications for Uganda	March 1
	Warren Blessing Tukwasibwe	Sr Manager, Partnership and Stakeholder Engagement for Uganda	March 1
IPA	**Jan Will**	Research Consultant/IPA lead on the RBF program verification	March 1
	Morris Bwambale	Verification field coordinator (*Q5-Q9)	March 1
	Jackie Namubiru	Verification field coordinator (Q1-Q4)	March 1
	IPA enumerators (Betty-Mafubira branch, Francis-Lira branch, Jolly-Kyotera and Masaka branch)	Enumerators	March 1
GDI	**Steven Bergen**	Development Impact Bond Compliance Manager	March 8
USAID DIV	**Paul Hamlin**	USAID DIV	March 15

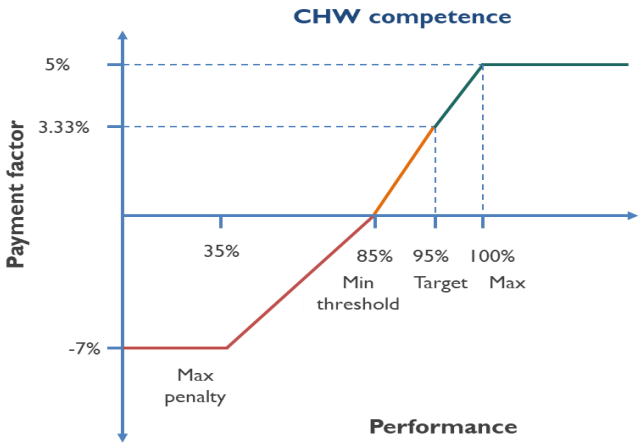
	Amelia McDonough	USAID DIV	March 15
	Michael Cretz	USAID DIV	March 15
Deerfield Foundation	**Dharan Kadiyala**	Deerfield	March 7

Table 6. Branch focus group discussions

Mafubira (Focus group discussion conducted on Feb 27th)	Balisanyuka Sarah	CHWs
	Saziri Baje	CHWs
	Nalubaga Stella	CHWs
	Kwagala Esther	CHWs
	Namusoke Betty	CHWs
	Woty Proscovia	CHWs
	Okello Emmanuel	branch team
	Ojambo Clare	branch team
	Nakamya Christine	branch team
	Mwesigwa Kenneth	branch team
	Nakaima Silvia	branch team
	Kasada Jonathan	branch team
	Bagira Teddy	branch team
Kyotera (Focus group discussion conducted on Feb 28th)	Nabasese Grace	CHWs
	Nakabazzi Mildred	CHWs
	Nalubega Dorothy	CHWs
	Kasagga Jude	CHWs
	Nangozi Harriet	CHWs
	Kikyonkyo Agatha	CHWs
	Tibenda Ruth	branch team
	Nalubwama Rashidah	branch team
	Wataka Peter	branch team
	Ggingo Benjamin	branch team
	Kaketo Ronald	branch team

Annex 3. Payment for performance on quality metrics





Green line: Living Goods receives an additional payment for performance on the quality metrics, and the payment is equal or higher than expected (assuming expected performance on the quantity metrics).

Orange line: Living Goods receives an additional payment for performance on the quality metrics, but the payment is lower than expected (assuming expected performance on the quantity metrics).

Red line: Living Goods receives a penalty for performance on the quality metrics. The penalty reduces payment for the quantity metrics proportionally.