NBER WORKING PAPER SERIES

USING PERFORMANCE INCENTIVES TO IMPROVE MEDICAL CARE PRODUCTIVITY AND HEALTH OUTCOMES

Paul Gertler
Christel Vermeersch

Working Paper 19046
http://www.nber.org/papers/w19046

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2013

We gratefully acknowledge funding from the World Bank's Bank-Netherlands Partnership Program, the British Economic and Social Research Council, the Government of Rwanda through a Japanese PHRD grant, and the World Bank's Spanish Impact Evaluation Fund. The authors have no other financial or material interests in the content of the paper. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Using Performance Incentives to Improve Medical Care Productivity and Health Outcomes
Paul Gertler and Christel Vermeersch
NBER Working Paper No. 19046
May 2013
JEL No. I11,J33,O12

ABSTRACT

We nested a large-scale field experiment into the national rollout of the introduction of performance pay for medical care providers in Rwanda to study the effect of incentives for health care providers. In order to identify the effect of incentives separately from higher compensation, we held constant compensation across treatment and comparison groups – a portion of the treatment group's compensation was based on performance whereas the compensation of the comparison group was fixed. The incentives led to a 20% increase in productivity, and significant improvements in child health. We also find evidence of a strong complementarity between performance incentives and baseline provider skill.

Paul Gertler
Haas School of Business
University of California, Berkeley
Berkeley, CA 94720
and NBER
gertler@haas.berkeley.edu

Christel Vermeersch
The World Bank
1818 H Street, NW
Washington DC
cvermeersch@worldbank.org

1. INTRODUCTION

Long-standing concerns about both the cost and effectiveness of health care have led to the relatively recent introduction of performance incentives for medical care providers.[1] These pay for performance (P4P) schemes typically pay bonuses to providers that give higher quality of care to more patients. While performance incentives have long been part of labor contracts, they have only relatively recently become widespread in health care worldwide. While there is some evidence on the effect of P4P on quality and access to services mostly from the US and the UK, there is little rigorous evidence on the impact on health outcomes and provider productivity especially in low and middle-income countries.

In this paper we provide evidence on the effect of incentives on provider productivity and on health outcomes in Rwanda, where we nested a prospective field experiment into the national rollout of a P4P incentive scheme. P4P affects health care provision in two ways; first, through incentives for providers to expend more effort in specific activities and second through an increase in the amount of available financial resources. In order to identify the incentive effect separately from the increase in resources, the traditional input-based budgets of the comparison group were increased by the average amount of incentive payments to treatment facilities. As a result, while treatment and comparison facilities had the same budget, a portion of the treatment facilities' budgets was determined based on their performance whereas the comparison facilities' resources were not. This is important because if P4P achieves its results just from increased financial resources, then the same results could be achieved from a simple increase in budget without incurring the administrative costs associated with implementing the incentive scheme.

We find that the performance incentives translated into large and significant improvements in child health outcomes. Specifically, we find that incentives led to an increase of 0.53 standard deviations in the weight-for-age of children 0-11 months and 0.25 standard deviations in the height-for-age of children 24-49 months. We previously documented the pathways through which the incentives likely affected health outcomes in Rwanda, i.e. increased preventative care utilization and improved clinical quality of care (Basinga et at. 2011).

We also found that the incentives were associated with a substantial improvement in provider productivity. We measured productivity as the gap between provider knowledge and

[1] See Van Herck et al. (2010) and Ashaman et al (2010) for reviews of performance incentives in medical care.

actual practice of appropriate prenatal care clinical procedures, holding the provider's budget constant. Knowledge can be thought of as the provider's production possibilities frontier. Under this definition, the difference between knowledge and practice is a measure of efficiency conditional on holding the provider's budget constant.

In our survey, while providers knew 63 percent of the appropriate clinical protocols for prenatal care, they only delivered about 45 percent of the appropriate protocols. This implies a gap of 18 percentage points between knowledge and practice. It implies that there is substantial distance from the provider practice to their production possibility frontier. The performance incentives reduced the gap by 4 percentage points on average.

Finally, we found large complementarities between provider skill (knowledge of appropriate clinical procedures) and performance incentives in the production of quality. Specifically, we found that higher skilled providers increased quality more than lower skilled providers in response to the same incentives. This means that incentives are more effective in a context where providers have higher skills and suggests that traditional training interventions may yield higher results with incentives than without.

Our results are especially important for low income countries who, despite dramatic increases in public spending on health in the last decade, have made little progress towards the 2015 Millennium Development Goal health targets (United Nations 2010 and 2011).[2] One reason is the poor performance of these medical care systems where absenteeism among providers is widespread (Chaudhury et al. 2006), provider knowledge of proper clinical procedures for basic care is low (Das and Hammer 2004), and providers give a standard of care that is well below their clinical knowledge[3]. In this context, incentive payment is a mechanism that can increase health provider productivity in terms of supplying higher quality care to more patients.

Our work contributes to the general literature on P4P in medical care, as it is the first to separately identify the impact of P4P incentives from the associated increase in resources.[4] Our work also contributes to the relatively small literature on the effects of paying medical care

[2] Among the 67 countries with highest child mortality rates, only 10 are on track to reduce mortality by two thirds. And the rate of decline in maternal mortality for all is well short of the 5.5 % needed to achieve 2015 MDG target.
[3] See Das and Gertler (2007); Das, Hammer and Leonard (2008); and Leonard and Masatu (2010a) and (2010b).
[4] See Witter et al., (2012) for a recent systematic review of health care performance incentives in low and middle income countries. Most of the literature that they cite do not have control groups and estimate the impact of P4P as jumps in time trends of the amount of services providers by treatment facilities.

providers for performance in developing countries;[5,6] to our knowledge, ours is the first rigorous evaluation of the impact on health outcomes and productivity in a low-income country.

There are three well-identified related evaluations in other low and middle-income countries. Hospital-based physicians in the Philippines who received extra bonus pay based in part on knowledge of clinical appropriate clinical procedures, reported increases in clinical knowledge (Peabody et al, 2011). In Indonesia, performance incentives to *villages* for improvements health and education outcomes led to an increase in labor supply from health providers (Olken et al. 2011). Finally, Miller et al (2012) found that bonus payments to *schools* significantly reduced anemia among students in China.

This paper is organized as follows. We first describe the P4P scheme introduced in Rwanda. We then use a simple model to layout the behavioral pathways through which the incentives plausibly could lead to improved health outcomes and show that there is evidence to support the model's predictions. We then describe our identification strategy and data. We follow with the results for health outcomes and then productivity, and end with a conclusion.

2. INSTITUTIONAL CONTEXT

Rwanda is one of the poorest countries in the world with a GDP of US$340 per capita in 2007 (World Bank, 2008). Since the end of the 1994 genocide, Rwanda has made remarkable progress in improving maternal and child health. Between 2000 and 2005, infant mortality fell from 107 to 86 deaths per 1,000 live births, while maternal mortality fell from 1,071 to 750 per 100,000 live births (Institut National de la Statistique du Rwanda and ORC Macro 2006). Despite this progress, the health system continues to grapple with serious shortages of qualified personnel and low levels of service delivery, especially in rural areas (World Bank 2010b).

a. P4P

In 2005, the Rwandan Ministry of Health (MoH) used the opportunity of an increase in the health sector budget to scale up nationally a P4P scheme for maternal and child health care services (Logie et al 2008; World Bank 2010a). The decision was based on positive reports of

[5] There is, however, a growing literature on P4P for medical care in the U.S. and other high-income countries with mixed results. See Van Herck et al. (2010), Ashaman et al (2010) and Scott et al. (2011) for reviews.

[6] There is a related literature on performance pay for teachers in low-income countries. For example see Glewwe et al. (2010), Muralidharan and Sundararaman (2011), and Muralidharan (2012).

P4P schemes in two Provinces that had been developed by a number of NGO's (Kalk et al. 2005; Soeters et al. 2005 and 2006). The P4P scheme provides bonus payments to primary care facilities based on the provision of various types of services and the quality of those services. P4P payments go directly to facilities and are used at each facility's discretion. The overall amount of P4P payments is large in comparison to facilities' budgets: a study of 68 facilities receiving P4P payments shows that P4P payments represent an increase in funding of 24.6% above the base budget. On average, 77 percent of P4P funds were used to compensate personnel resulting in an increase of 38 percent in staff compensation (Basinga et al 2011).

b. The Payment Formula

The scheme pays for 14 maternal and child healthcare services conditioned on an overall facility quality assessment score. The formula used for payment to facility i in month t is:

$$Payment_{it} = \left(\sum_j P_j U_{jit} \right) \times Q_{it} \quad \text{with} \quad 0 \leq Q_{it} \leq 1,$$

where P_j is the payment per service unit j (e.g. institutional delivery or child preventive care visit), U_{ijt} is the number of patients using service j in facility i in period t, and Q_{it} is the overall quality index of facility i in period t.

The 14 service indicators (U_{jit}) and associated payment rates (P_j) are listed in Table 1. The Rwanda Ministry of Health (MoH) defined these indicators and payments based on national health priorities, available budget and previous NGO experience with P4P (Ministère de la Santé du Rwanda 2006 and 2008). The first 7 indicators consist of the number of visits to the facility for various types of service such as prenatal care and institutional delivery, while the second set of 7 indicators refers to the content of care provided during those visits. This second set includes the number of children who were fully vaccinated, the number of pregnant women who received tetanus vaccines and malaria prophylaxis during prenatal care, the number of at-risk pregnancies that were referred to hospitals for delivery during prenatal care, the number of severely malnourished children who were referred to treatment facilities, and the number of general emergencies that were referred to the appropriate facility for care. The health literature considers these to be measures of aspects of the process or clinical quality of care (Donabedian, 1988).

The actual amount paid to the facility is adjusted based on the overall facility quality, Q_{it}. The facility's overall quality enters the payment formula as a multiplicative factor that proportionately raises or lowers the payment for the 14 output indicators. The quality index is bounded between zero and one. If the facility meets all of the quality criteria, then the index equals one and the facility receives the full P4P payment. However, if the facility is deficient in some of the quality criteria, then all of the payments are discounted. For example, if the facility only scores 0.80 on the quality index, then it only receives 80 percent of the payment for the 14 output indicators. In this way, the P4P scheme pays for both facility output and facility quality.

The quality index Q_{it} is a function of structural and process measures of quality specified in the Rwandan preventive and clinical practice guidelines (Ministère de la Santé du Rwanda 1993, 1997, and 2003). Structural measures are the extent to which the facility has the equipment, drugs, medical supplies and personnel necessary to be able to deliver a specific medical service. Process measures are the clinical content of care provided for specific services. [7]

The formula for the quality index is:

$$Q_{it} = \sum_k \omega_k S_{kit} \quad \text{with} \quad \sum_k \omega_k = 1,$$

where S_{ikt} is the share of indicators for service k that are met by facility i in period t, and ω_k is the weight for service k. If a facility has perfect structural and process quality, then all the S_{ikt} take on value one and the overall quality index is equal to one; in this case, the facility is paid the maximum possible bonus for the services provided. By contrast, if the quality index is less than one, P4P payments are discounted for *all* services provided at the facility. Table 2 details the services that are included in the quality index, their weights and the relative importance of structural and process indicators in the computation of the score for each service k.

c. Full Prices

There is a large range of payment rates (P_j) for the U_{ijt}'s (Table 1). The largest are for institutional deliveries and emergency referrals to hospitals for obstetric services; both services pay $4.59 per case. The next largest, at $1.83, is for new contraceptive user visits, referral of at-risk pregnancies to hospitals for delivery and referral of malnourished children to higher-level

[7] Clinical practice guidelines are "systematically developed statements to assist practitioners and patient decisions about appropriate health care for specific circumstances." (Field and Lohr, 1990)

facilities for treatment. Facilities are paid about half as much for a child who is fully vaccinated on time and about half as much again for pregnant women who receive the tetanus vaccine and the same for malaria prophylaxis. Curative and contraceptive re-supply visits are paid the small sum of $0.18 per visit. Finally, prenatal care visits are paid only $0.09 per visit with a bonus of $0.37 for every woman who completes 4 visits.

The incentives implicit in the payment scheme are more complicated than they appear. The incentive structure focuses not just on treating more patients, but also on providing more patients with higher quality of care. This happens not only through the multiplicative scaling factor Q discussed above, but also because direct payment for content of care services in the U_i's. The full payment for seeing a patient depends on the services provided during that visit. In fact, payment rates for visits are much higher if the provider supplies better content of care.

Consider for example, the payment for prenatal care. Providers receive $0.18 for every pregnant women who starts prenatal care, an additional $0.37 if the women completes at least 4 visits, an additional $0.92 if they give the patient a tetanus shot and malaria prophylaxis during a prenatal care visit, and an additional $1.83 if they assess the delivery to likely be risky and refer the mother to deliver at the district hospital. Hence, a provider will receive $0.55 for four prenatal care visits of low quality versus $1.47 for providing high quality.

The same is true for institutional delivery and child growth monitoring. If the provider detects a high-risk pregnancy and refers the woman to the hospital for delivery, payments for this high-quality care increase to $3.30. In the case of growth monitoring, the payment to the provider is $0.18 per visit plus an additional $1.83 if the child is malnourished and she refers her to the hospital for treatment.

d. Administration and Auditing of Payments

P4P payments are administered by district steering committees comprised of members of the Government, Providers and civil society (Ministry of Health Republic of Rwanda, 2006; Fritsche et al 2010). Facilities submit monthly activity reports (U_{jit}) and quarterly requests for payment to the steering committee, which is responsible for verifying the data and authorizing payment. For the referral indicators, the facility must also submit verification from the hospital that the referral was appropriate and the referred patient was treated. The committee verifies the

reports by sending auditors to facilities on a quarterly basis on an unannounced randomly chosen day. The auditors review the utilization registry and facility records to verify the data reported is the same as the data recorded in facility records. During the 2006-2008 period, MoH conducted a survey of face-to-face interviews with approximately 1000 patients to verify the accuracy of the records. False reporting on patients or services was less than 5 percent (HDP 2008).

Information used to compute each facility's overall quality score is collected under the existing national monitoring system that requires all district hospitals to monitor and supervise the quality of health centers in their districts. Every quarter, a district hospital team from different services (e.g. prenatal, curative care, preventive care) visits each facility on an unannounced randomly chosen day to assess the facility's quality through direct observation and review of patient records using a standardized tool. At the end of the visit, the team discusses their findings with the facility's personnel and provides recommendations to improve quality. In P4P districts, the data are used to construct the overall quality score for the facility each quarter.

3. PATHWAYS

Here we discuss how the introduction of incentives induces providers to supply more effort in ways that increase utilization and quality of care. We argue that these are the pathways through which incentives would lead to improved health outcomes and increased productivity.

Without incentives a provider is paid a fixed amount and her income is not affected by seeing more patients or providing them with better care. Hence, the provider treats all patients who show up and provides them with a minimum level of care defined by her ethical standards.[8]

The P4P scheme introduces a new dimension by linking part of provider income to the provision of certain services and to quality of care. Taking into account the basic structure of the P4P formula, the profit function is:

$$V = I + [\sum_i P_i U_i(\varepsilon_i)]Q(\epsilon_q) - C(\varepsilon) \quad (3)$$

where I is the fixed salary, P_i is the P4P payment for service i, U_i is the total quantity of service i provided to patients, Q is the overall quality of care, ε is total effort (i.e. $\varepsilon = \sum_i P_i U_i(\varepsilon_i) + \varepsilon_q$), and

[8] Some argue that medical providers care about their patients' health and are not just motivated by money. We incorporate this into our framework through minimum effort constraints as defined by the provider's ethics. See Reinikka and Svensson (2010) for an investigation of objective functions of clinics in a rural African setting.

$C(*)$ is the cost of effort. The service production functions $U_i(\cdot)$ and the quality production function $Q(\cdot)$ are concave in effort, and $C(\cdot)$ is a convex function of total effort.

The provider chooses effort levels to maximize income subject to effort levels being no less than the ethical minimum levels. In the case of an interior solution, effort is allocated in such a way that marginal revenue of effort is equalized across the three types of effort and that it is equal to the marginal cost of effort:

$$P_1 U_1'(\epsilon_1) = P_2 U_2'(\epsilon_2) = \cdots = [\textstyle\sum_i P_i U_i(\varepsilon_i)] Q'(\epsilon_q) = C'(\varepsilon) \quad (4)$$

Note that the marginal return to effort supplied to quality depends on all prices. Hence, an increase in either price always raises the return to effort supplied to quality. Marginal return to effort supply to service *i* depends not only on the price but also on the extent to which more effort increases quantity. Effort supplied to anything raises the marginal cost of effort because the cost of effort is a function of total effort.

The introduction of P4P raises all prices simultaneously. This increases the allocation of effort to quality because increases in any price raise the marginal return to supplying effort to quality. The largest allocations of effort to specific services are to those services for which the relative price increases are the largest and the marginal productivity of effort is the highest.

The effect of the introduction of the P4P payments depends not only on the relative payment rates, but also on how hard it is to increase the levels of services. In general, it takes more work to increase services that depend on patient choices than services that are completely in the provider's control. For example, it takes more work to convince a pregnant woman to come to the clinic for prenatal care than to give the woman a tetanus shot once she is there. Hence, even if payments were equal for an additional patient visit as for a tetanus shot, one would expect to see larger increases in the number of tetanus shots than in the number of visits to the facility. Moreover, initiation of care takes more effort than its continuation. For example, it will take a provider substantial amounts of effort to go out to the community to find pregnant women to bring them in for prenatal care. By contrast, it is a relatively easier task to use an existing prenatal care visit to lobby women already in prenatal care to deliver in the facility.

We may see no increases in effort at all for some services. If for a particular service the relative price increase is small or the relative marginal productivity of effort low, then there

maybe no incentive for the provider to supply more effort to that service despite the absolute increase in price. In this case, effort for that service will remain at the minimum ethical bound.

Evidence reported in Basinga et al. (2011) is consistent with the empirical predictions from this framework (Appendix Table A).[9] First, there are significant effects of incentives on the quality of prenatal care. Increasing quality is completely in the control of providers and the incentives are high-powered, as the quality index is a scaling factor that is applied to all payments for all services. Second, there was a large significant effect of P4P on institutional delivery, which had by far the highest payment rate of $4.59 (Table 1). The payment was so much larger for institutional delivery, compared to other services, that providers reported paying community health workers to find women about to deliver and bring them to the facility. Well-child care also responded strongly to the P4P scheme, even though its unit payment is relatively low at $0.18 per visit. However, the payment rate jumps substantially to $2.03 if the provider identifies a malnourished child and gets them into treatment. Since 64 percent of Rwandan children under age five are malnourished,[10] the expected payment for a high quality growth-monitoring visit is quite high.

On the other hand, P4P did not increase the initiation of prenatal care, the number of women who completed four prenatal care visits or contraceptive resupply visits. These services have a high marginal cost of effort since utilization depends on maternal decisions to seek care and have low monetary payoffs for providers. For example, the payment for prenatal care is $0.09 for women who initiate care. At baseline 95 percent of women had at least one prenatal visit. It would be very costly for providers to find the remaining 5% of women and convince them to get care.

4. HEALTH OUTCOMES

We now ask whether the increased utilization of higher quality of care induced by the P4P incentives resulted in better child health outcomes. This section is organized as follows. We first describe the field experiment that generated the variation in the data that we used for identification. We then present our empirical specification and statistical inference. Next, we describe data collection and measurement. Finally we present and discuss results.

[9] The results from Basinga et al. (2011) can be found in Appendix Table A with a new correction for the standard errors to account for a small number of clusters.
[10] Institut National de la Statistique du Rwanda and ORC Macro (2006).

a. Evaluation Design

The evaluation design took advantage of the phased implementation of the program at the district level over a 23-month period. Rwanda manages its health care system at the district level and P4P is no exception. As a result the government mandated that all facilities in a district must be incorporated into the P4P scheme at the same time. Hence, the evaluation employed a stratified cluster randomized designed where districts were first grouped into pairs with common characteristics and then randomly assigned to treatment comparison groups[11]. Administrative districts with pre-existing P4P schemes were excluded from the experimental design. The remaining districts were grouped into blocks based on similar characteristics for relief, rainfall, and predominant livelihoods as per the 2002 Census.

However, just before implementation, administrative district boundaries were redrawn in the context of a government-wide decentralization effort (MINALOC 2004). The objective of the decentralization was to enhance institutional development and capacity building for responsive local governance., to develop an efficient transparent and accountable fiscal and financial management system at local government and grassroots level.

As a result, some of the experimental areas were combined into new districts with areas that already had the pilot P4P schemes. Because P4P could not be "removed" from health facilities that were already implementing the schemes, and because P4P is managed at the district level, the MoH required that all facilities within those new districts be in the treatment group. This led the evaluation team to switch the assignment of treatment and comparison for eight districts. In the end, the study included 10 districts in the treatment group and 9 in the comparison group. In the analysis, we will address this by treating the evaluation as a quasi-experimental design and estimating difference-in-difference models.

The sample included 166 of Rwanda's 401 primary care facilities, 80 in treatment districts and 86 in comparison districts. The facilities in the treatment group started receiving P4P in 2006, while the facilities in the comparison group continued with traditional input-based financing for an additional 23 months. P4P was introduced in treatment districts over a 5-month period, yielding a minimum program exposure of at least 18 months.

[11] A couple of small districts were grouped together in order to achieve balance in sample size.

Since a primary objective of the evaluation was to isolate the impact of the P4P incentives separately from the effect of increased resources, it was necessary to hold the level of resources constant across treatment and comparison facilities. To accomplish this, comparison facilities' traditional input based budgets were increased by the average amount of P4P payments to treatment facilities on a quarterly basis during the entire 23-month treatment window on a quarterly basis. As a result, treatment and comparison facilities had the same levels of financial resources on average throughout the study. In this design, the differences in outcomes between the two groups at follow-up are attributed to the difference in incentives and not to a difference in available financial resources.

b. Data Collection and Control-Treatment Group Balance

We surveyed all 166 facilities plus a random sample of households in the catchment area of each facility. The surveys were conducted at baseline in 2006 prior to the implementation of P4P in treatment facilities and again approximately 23 months later, before the comparison facilities were incorporated in the program. The surveys were conducted independently from the operation of the P4P program. Payment to facilities was based on administrative records and reports from the facilities and never on the evaluation surveys.

The facility survey collected information on staffing, expenditures, medical equipment, drug availability from the facility administrator, and provider knowledge about the appropriate clinical procedures for quality prenatal care. As part of survey, enumerators also conducted exit interviews with approximately 10 women who visited the facility to collect information on the actual clinical services (quality) provided during their prenatal care visit. The sample of facilities was well balanced at baseline. There were no differences in the baseline means between the treatment and control groups for the 17 baseline characteristics presented in Table 3.

In addition, we specifically note that there is no difference in mean log expenditures at endline in 2008 between treatment and comparison groups. This implies that the identification strategy of compensating the control facilities for incentive payment to treatment facilities did indeed hold facility budgets constant on average across treatment and control groups. This result supports our interpretation that any differences in outcomes are caused by the P4P performance incentives as opposed to resource differences.

The household survey consists of a random sample of 13 households living in each facility's catchment area, for a total sample of 2,158 households. To build the sample, we first sampled 13 census zones from each facility's official list of zones in their catchment area. We then physically listed all households in the sampled zones and randomly selected one household with at least one child under 5 years old from each zone.

Response rates were high as only 2 percent of sampled households refused to participate in the interview. In the follow-up survey, 88 percent of the baseline households were re-interviewed. The rate of attrition from the baseline sample was not statistically different between the treatment and comparison groups (12 percent each). In addition, the household data were well balanced between treatment and comparison groups at baseline. Only 1 of the 30 characteristics reported in Table 4 were significantly different at the 5% level between treatment and control groups at baseline.

c. Child Health Measurement

We measure child health outcomes for two age groups: children 0-11 months and 24-47 months. Children in the 0-11 months range in the treatment group in the 2008 survey would have been exposed to P4P during the full prenatal period and during the full time after birth. Children between 24 and 47 months in the treatment group in the 2008 survey would have been exposed to P4P for 18-23 months during the early stages of life, but not during the prenatal period.

We consider two measures of health: (i) height-for-age z-score and (ii) weight-for-age z-score. A child's height results from her genetic potential, adjusted for insufficient nutrient intake and inability to absorb nutrition because of illness. Therefore, height is a summary measure of health and nutrition since conception. By contrast, weight is an indicator of current nutrition and illness status and does not represent factors that accumulate over the lifetime. Better prenatal care, which includes nutritional advice to mothers and the diagnosis and treatment of maternal illness, could in principle increase both infant height and weight. Better child preventive care, which includes vaccination and growth monitoring, and child curative care, limits the duration and severity of illness and thereby has the potential to affect height and weight.

We measured child height and weight using standard international procedures and portable scales and stadiometers, which were recalibrated on a twice-weekly basis in the field. As part of the quality control procedures, all children were measured twice during the visit. We

first report height and weight in centimeters and kilograms respectively. Then, we standardize these measurements into height-for-age and weight-for-age z-scores in accordance with World Health Organization guidelines. The z-scores measure the number of standard deviations from age-sex standardized height of a healthy reference population.

d. Estimation

Given the reassignment of districts between the treatment and comparison groups before the start of the study, and the limited number of districts that could be assigned to the treatment and comparison groups, we view our study as quasi-experimental. While the sample is balanced at baseline on outcomes and characteristics, it is possible that the reassignment of districts was correlated with something unobservable to us and related to health outcomes. However, redistricting took place within the context of a decentralization agenda that was led by the Ministry of Local Government, and we find no evidence that it was driven by or related to health outcomes (MINALOC 2004). Therefore, we think it is likely that any relevant unobservable factors were likely to be invariant over the time period of the intervention.

We used difference-in-differences (DiD) methods to estimate the impacts of P4P on outcomes. DiD compare the change in outcomes in the treatment group to the change in outcomes in the comparison group.[12] The method controls for observed and unobserved time-invariant characteristics as well as for time-varying factors that are common to the treatment and comparison groups. As we discussed above in section 4.b, the final assignment to the treatment and comparison groups was orthogonal to pre-intervention observable variables, leading us to believe that there is likely no correlation between this assignment and unobservables that drive program effects.

All of the individual outcomes relate to pregnancies, however many women do not give birth in both waves of the survey. Hence, we treat the 2006 and 2008 household surveys,

[12] An alternative, sometimes used in the literature, is the intent to treat estimator that compares the originally assigned treatments to controls. In this case, however, we would have misassigned 40% of the observations and would be grossly underpowered. Also, all of the examples we could find use the ITT in cases where the study entered the field intending to implement the original design and where behavioral choices by the study participants compromised the study design. In our case, the design was changed before we entered the field and was not compromised by the study participants. Hence, while our difference in difference estimator requires stronger assumptions, we believe that it is appropriate in terms of identification and is valid based on the balance tests and knowledge of the institutional designs that drove the change in design. In our view, the difference in difference choice maximizes potential power without sacrificing internal validity.

described below, as repeated cross-sections and estimate the following regression specification of the difference-in-difference model for individual outcomes:

$$Y_{ijt} = \alpha_j + \gamma_{2008} + \beta \cdot P4P_j \cdot I_{2008} + \sum_k \lambda_k X_{kijt} + \varepsilon_{ijt} \quad (6)$$

where Y_{ijt} is the health of child *i* living in facility *j*'s catchment area in year *t*; $P4P_j$ is a dummy variable that takes value 1 if facility *j* belongs to Phase I (i.e. started receiving P4P in 2006) and 0 otherwise; α_j is a facility fixed effect; γ_{2008} is a fixed effect for 2008; I_{2008} is a dummy variable that takes value 1 if the year of observation is 2008 and 0 otherwise; the X_{kijt} are individual characteristics; and ε_{ijt} is a zero mean error term. We estimate each regression both with and without individual characteristics.

A limitation of our design is the small number of districts for assignment. Since the unit of assignment to treatment and comparison was the district and not the facility, there may be inter-cluster correlation in the error terms. The asymptotic justification for inference with cluster-robust standard errors assumes that the number of clusters goes to infinity. Yet in our application there are too few clusters for this assumption to hold. Therefore we base our statistical inference on randomization inference hypothesis tests that use WILD bootstrapping Monte Carlo methods allowing for intercluster correlation as recommended in Cameron et al (2008). These tests return a p-value for the hypothesis rather than a standard error. Hence we report the estimated coefficient and the p-value for the test of significance.

e. Results

We consider two age groups: (1) children 0-11 months at endline, and (2) children 24-47 months at endline. We estimated 2 versions of equation (6): one without controls and a second with controls. The controls include the child's age and sex, maternal height, mother's age, whether the mother has completed primary school, whether the father lives in the household, whether the family is a member of a Mutuelle (health insurance fund), the total number of household members, the number of household members under the age of 6, whether the

household owns land, and dummy variables for quartiles of the household asset value.[13] The child's age was entered as a series of dummy variables that represent one-month increments.

The estimated effects of P4P on child health outcomes are reported in Table 5. Among the 0-11 month old children who benefited from P4P since conception, we find large and significant positive effects on weight-for-age z-score. Infants in the treatment areas gained 0.53 of a standard deviation in weight as a result of P4P. Among 24-47 months old, who benefited from the program for 23 months starting between age 1 and 24 months after birth, the program led to a gain of over 0.25 of a standard deviation in height for age. Note that the size of the effect on height for age of the younger group is 0.22 without controls and 0.16 with controls. However, the sample size for the younger group is less than half of the sample size for the older group.

While effects on health outcomes are large for both age groups, we see effects on weight for the younger group and on height for the older group. This difference can be explained by the role of breast-feeding on weight gain and height growth. Micronutrients that are not present in large amounts in breast milk are critical for improvements in height (Dewey and Adu-Afarwuah, 2008). Hence, during the period of exclusive breast-feeding children tend to gain weight rather than height. Reductions in illnesses that impede a child's absorption of the nutrition in breast milk tend to lead to gains in weight. Once supplemental foods that contain more micronutrients are introduced into the child's diet, reductions in illnesses that improve a child's ability to absorb nutrition are manifest in height. According to the 2005 Demographic and Health Survey, 88% of children less than 6 months old are exclusively breastfed. And 31 percent of children age 6 to 9 months did not receive supplementary foods (Institut National de la Statistique du Rwanda and ORC Macro 2006).

[13] The household asset value was constructed on the basis of the value of household assets including owned houses, household durable goods, farm animals, farm equipment and micro-enterprise equipment.

5. PRODUCTIVITY

We now turn to a deeper investigation of how the performance incentives affected provider productivity in terms of quality of prenatal care. We ask 3 questions: (1) how much of the observed effect of incentives on quality was driven by increased provider effort versus improved provider knowledge of the appropriate clinical procedures, (2) did better skilled providers take more advantage of the incentives and increase quality more than lower skilled providers, and (3) how much did the incentives improve efficiency as measured by the difference between knowledge and practice of the appropriate clinical guidelines.

a. Measurement

In rural Rwanda, nurses are the primary care givers and they work in clinics that are fully equipped to be able to provide care based on the standards dictated by Rwandan clinical practice guidelines. Specifically, clinics are staffed with about 6 nurses and 4 medical technicians or midwives (Table 3 Panel A). The availability of equipment and drugs needed to provide quality care is reasonable (Table 3 Panel B). The structural quality indices are the share of drugs and equipment available at the facility among those that the Ministry of Health guidelines define as necessary in order to deliver each type of care (Ministère de la Santé du Rwanda 1993, 1997, 2003 and 2009). At baseline clinics had on average 96 percent of the drugs and equipment necessary to provide prenatal care services. Hence, improvements in prenatal care quality could not have come through more drugs, supplies and equipment, but rather through better applying their knowledge of appropriate clinical procedures.

We measure a medical care provider's skill and capability by their knowledge of appropriate clinic protocol for a prenatal care visit. Our measure of provider knowledge is the share of the appropriate prenatal care clinical procedures specified in the official Rwandan Clinical Practice Guidelines (CPG) for prenatal care (Ministry of Health Contractual Approach Unit, Rwanda 2006, and Ministère de la Santé du Rwanda 2008).[14] The 24 specific clinical services cover previous pregnancy history, medical history, current pregnancy status, physical exams, diagnostic laboratory tests, and case-risk management (Table 6).

[14] CPGs are a recommended set of clinical procedures conducted during the prenatal care visit that maximize the probability of good health outcomes based on the clinical literature and expert opinion. First developed in the United States (Field and Lohr 1990), the Rwandan guidelines are based on the US version adjusted to local resource constraints.

We used a clinical "vignette," which is a standardized hypothetical patient case with a specific medical history to collect the data on knowledge. We presented the vignette to a randomly selected health worker who regularly provides prenatal care and asked the health worker to describe the clinical protocol that she would apply. We then used the answers to compute the share of official CPG clinical content items that the provider mentioned without prompting from the interviewer.[15]

Providers, however, do not necessarily deliver clinical services up to their level of knowledge. We define quality of care as the actual clinical services delivered to patients. We collected information on actual clinical services provided during prenatal care from exit interviews of patients leaving facilities and from the household surveys. We measured the quality of prenatal care by computing the share of actual clinical content items delivered during a prenatal care consultation to the items that should compose a typical prenatal consultation as recommended in the Rwandan CPGs.[16]

The quality index and the knowledge index cover the same items and therefore are comparable. At baseline, medical providers knew about two-thirds of the prescribed clinical protocol for prenatal care (Table 3 Panel D). While providers know two-thirds of the appropriate protocols, they only delivered about 45 percent of the CPG prescribed protocols (Table 3 Panel E). This implies a gap of 18 percentage points between knowledge and practice on average, which translated into providers delivering about two-thirds of clinical services they know they should deliver. Hence there is substantial distance from the provider practice to their production possibility frontier. This distance can be interpreted as a measure of inefficiency.

b. Knowledge versus Effort

We have argued that the incentives induced providers to put more effort into delivering higher quality services. An alternative explanation might be that P4P affected provider knowledge through which changed provider practice.[17] This could have happened because

[15] This measure of competency was used in Das and Hammer (2004), Kak et al (2001) and Peabody et al (2004).
[16] This measure has been used in Barber (2006), Barber et al (2007), Peabody et al. (1997), and Das et al (2007).
[17] A third possibility is that the clinic invested in improved structural quality (i.e. drugs and equipment). However, the identification strategy held constant the level of resources between treatment and control facilities plus mean structural quality at baseline was already at 96% of drugs and equipment necessary to provide quality prenatal care as prescribed by the Rwandan Clinical Practice Guidelines (Table 3). Hence, there was very little room for improvement on this margin and analysis of the data confirm no impact of incentives on prenatal structural quality.

facilities might have changed personnel, substituting better-trained workers for lower-trained ones or because there could have been an increase in knowledge of the existing personnel. Recall that as part of the supervision process, district hospital supervisors discuss the results of their quality assessments with facility personnel, providing recommendations to improve care where needed. However, these visits were applied to both treatment and comparison facilities, so the opportunities to learn provided by these visits were no different between treatment and comparison groups. It is possible, however, that treatment and control facilities differed in the amounts of learning gained from the feedback provided, because treatment facilities knew that the quality assessments would directly influence their payment rates; hence treatment facilities may have been more attentive to advice from the district hospital teams.

To ascertain the importance of this potential causal pathway, we first estimate the impact of the P4P program on provider knowledge of the prenatal care protocol using a specification similar to the difference in difference specification as in equation (6):

$$K_{jy} = \propto_j + \gamma_{2008} + \beta \cdot P4P_j \cdot I_{2008} + \varepsilon_{jt} \quad (7)$$

where K_{jt} is provider j's knowledge of prenatal care clinical procedures in year t and the other variables are defined as in (6). The unit of observation in this case, however, is the facility and not the patient.

The results, reported in the first column of Table 7 show some weak evidence that health worker knowledge may have improved as a result of P4P. We find a large point estimate of a 0.40 standard deviation increase in knowledge, but it is not statistically significant. The lack of significance may due to small sample sizes rather than no effect.

In order to assess whether the association between P4P and quality is driven by knowledge as opposed to effort, we estimate a difference-in-difference model similar to the one specified in equation (6) for quality controlling for knowledge:

$$Q_{ijt} = \propto_j + \gamma_{2008} + \beta \cdot P4P_j \cdot I_{2008} + \delta \cdot K_{jt} + \sum_k \lambda_k X_{it} + \varepsilon_{ijt} \quad (8)$$

where Q_{ijt} is the quality of prenatal care that provider j gave to individual i in year t and the rest of the variables are defined as above. In this case, the unit of observation is the patient as the dependent variable is the quality of prenatal care received by the patient.

In our base excluding knowledge in Quality of Care Model 1 in Table 7, we estimate that P4P is associated with a 0.16 standard deviation increase in quality and is significant. Additionally controlling for knowledge in model 2 does not appreciably alter the estimated impact of P4P. Therefore, knowledge is not likely to be the main path through which P4P improved practice; rather, P4P almost surely increased quality through increased provider effort.

c. Knowledge Complementarities

Model 2 in Table 7 also allows us to examine the direct effect of knowledge on quality. In fact, one of the most common interventions to improve quality of care is by training medical care providers in proper clinical procedures. However, we find no evidence of a direct effect of increased knowledge on quality of care.

While changes in knowledge do not seem to directly impact quality, knowledge may be complementary to P4P in the sense that P4P may be more effective when providers are relatively high-skilled. More knowledgeable providers maybe able to more easily exploit the P4P incentives for financial gain by increasing quality of care than less knowledgeable providers.

To test this hypothesis, we amend the specification in (8) to include an interaction between treatment and whether the provider was in the top half of the baseline knowledge distribution:

$$Q_{ijt} = \propto_j + \gamma_{2008} + \beta \cdot P4P_j \cdot I_{2008} + \delta \cdot \overline{K}_{j,2006} \cdot P4P_j \cdot I_{2008} + \rho \cdot K_{jt} + \varepsilon_{ijt} \qquad (9)$$

Where $\overline{K}_{j,2006}$ indicates whether the provider was in the top half of the knowledge distribution at baseline and all of the other variables are defined as above.

In this specification, reported in Model 3 of Table 7, we do find differences in the program's impact on prenatal care practice between facilities that had health workers with above-the-median levels of knowledge, as opposed to below the median. Specifically, we estimate that incentives led to an insignificant effect of 0.07 standard deviations for lower skilled providers and a highly significant 0.21 increase for higher skilled providers.

d. Efficiency

Another interpretation of how P4P works is based on the idea that providers are not delivering services up their full ability (knowledge) and that the difference between ability and practice is a measure of efficiency. If we consider a provider's knowledge as their production possibilities frontier, then one can interpret the gap between knowledge and practice holding budget constant as a measure of inefficiency.

There is evidence of substantial inefficiency as provider delivery of clinical prenatal services is substantially lower than their knowledge of appropriate clinical procedures. Recall that providers on average know 63 percent of appropriate procedures, but deliver only 45 percent, leaving an 18-percentage point difference between knowledge and practice.

We depict the efficiency gap in figure 1 where knowledge is on the horizontal axis as the share of prenatal CPG recommended clinical services that the provider knows and quality is on the vertical axis as the share of prenatal CPG recommended clinical services actually provided. The 45^{o} line is the production possibility frontier (PPF) where providers deliver clinical quality care to the best of their knowledge. If providers deliver quality of care below their level of knowledge, then they would be preforming inside the PPF. The vertical distance between the frontier and the performance point is a measure of inefficiency.

We also included in Figure 1 the actual performance curves of the providers in our data set. The curves are bivariate nonparametric regressions of quality against knowledge separately for treatment and comparison groups at endline. Notice that both lines are well inside the PPF implying substantial levels of inefficiency at all skill levels. In addition, while the performance curves are upwards sloping, they are flatter than the PPF. This implies that while knowledge improves performance, inefficiency increases with knowledge. Finally, the performance curve for the treatment group is above and steeper sloped than the curve for the comparison group. This implies that P4P reduced the inefficiency and reduces it more for more skilled providers.

We estimate the impact of P4P on inefficiency using the specification in equation (8):

$$K_{jt} - Q_{ijt} = \propto_j + \gamma_{2008} + \beta \cdot P4P_j \cdot I_{2008} + \delta \cdot \overline{K}_{j,2006} \cdot P4P_j \cdot I_{2008} + \rho \cdot K_{jt} + \varepsilon_{ijt} \quad (9)$$

where the dependent variable is now inefficiency measured as the share of CPG clinical services the provider knows minus the share of CPG clinical services delivered to a patient.

The results are reported in Table 8. In the first specification we exclude knowledge and the interaction with knowledge from the model. In this case, we find that P4P reduces inefficiency by 3.5 percentage points or about 20 percent on average. When we control for provider knowledge in Model 2, the effect of P4P on inefficiency increases slightly to 4 percentage points. In this specification higher knowledge is actually associated with greater inefficiency. In other words, while increases in provider knowledge improve the quality of care, the improvement in quality is less than the improvement in knowledge. Finally, in Model 3 we include the interaction between treatment and more knowledgeable providers, we estimate that P4P has a much larger effect on inefficiency for more knowledgeable providers. We find no reduction in inefficiency for providers below the knowledge median, but we find a 6-percentage point improvement among providers above the knowledge median.

6. CONCLUSION

In this paper we provide evidence on the effect of incentives on provider productivity and on health outcomes in Rwanda, where we nested a prospective evaluation into the national rollout of a P4P scheme. In order to identify the incentive effect separately from the increase in resources, the budgets of the comparison group were compensated so that treatment and comparison facilities had the same budgets on average, but a portion of the treatment facilities' budgets were determined based on their performance whereas the comparison facilities' resources were not.

We show that the incentives improved access to higher quality care that resulted in substantial improvements in child health outcomes. In addition, we find that provider incentives led to a 20 percent improvement in efficiency. These findings lend strong support for the use of provider performance incentives to improve health outcomes. Finally, we find evidence of complementarity between the P4P incentive and the knowledge (skill) of health care providers. This suggests that effects of P4P incentives would be higher if completed with interventions that improve provider skill, such as training, and the training would have a great impact in settings with performance incentives.

REFERENCES

Alshamsan, Riyadh, Azeem Majeed, Mark Ashworth, Josip Car, and Christopher Millett (2010) "Impact of pay for performance on inequalities in health care: systematic review," *Journal of Health Services Research and Policy, Vol.*15, pp.178—184; doi:10.1258/jhsrp.2010.009113

Barber, Sarah (2006), "Does the Quality of Prenatal Care Matter in Promoting Skilled Institutional Delivery? A Study in Rural Mexico," *Maternal and Child Health Journal*, Vol. 10, pp. 419–425.

Barber Sarah, Stefano Bertozzi and Paul Gertler (2007), "Variations in prenatal care quality for the rural poor in Mexico", *Health Affairs*, Vol. 26(3), pp.w310-23.

Basinga, Paulin, Paul Gertler, Agnes Binagwaho, Agnes Soucat, Jennifer Sturdy and Christel Vermeersch (2011), "Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation," *The Lancet*, Vol 377, pp. 1421-28.

Cameron, Colin, with Jonah Gelbach and Douglas Miller (2008)"Bootstrap-Based Improvements for Inference with Clustered Errors", *Review of Economics and Statistics*, Vol. 90, 414-427.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan and F. Halsey Rogers (2006), "Missing in Action: Teacher and Health Worker Absence in Developing Countries," *The Journal of Economic Perspectives*, Vol. 20, No. 1, pp. 91-116.

Das, Jishnu and Paul Gertler (2007) "Variations In Practice Quality In Five Low-Income Countries," *Health Affairs*, Vol. 26: pp.296-309

Das, Jishnu and Jeffrey Hammer (2004), "Which Doctor? Combining Vignettes and Item Response to Measure Clinical Competence," *Journal of Development Economics*, Vol. 78, pp. 348-383.

Das, Jishnu and Jeffrey Hammer (2007), "Money for Nothing: The Dire Straits of Medical Practice in Delhi, India," *Journal of Development Economics*, Vol. 83(1), pp. 1-36.

Das, Jishnu, Jeffrey Hammer, and Kenneth Leonard. (2008) "The Quality of Medical Advice in Low-Income Countries." *Journal of Economic Perspectives*, 22(2): 93–114.

Dewey, Kathryn and Seth Adu-Afarwuah (2008) "Systematic Review of the Efficacy and Effectiveness of Complementary Feeding Interventions in Developing Countries," *Maternal and Child Nutrition*, Vol. 4, pp.24-85.

Donabeidian, A. (1988), "The Quality of Care. How Can it be Assessed?" *The Journal of the American Medical Association*, Vol. 260(12), pp. 1743-1748.

Field, Marilyn and Kathleen Lohr (1990) Clinical Practice Guidelines: Directions for a New Program, Institute of Medicine, National Academy Press, Washington D.C.

Flodgren, Gerd, Martin Eccles, Sasha Shepperd, Anthony Scott, Elena Parmelli, and Fiona Beyer (2011), "An Overview of Reviews Evaluating the Effectiveness of Financial Incentives in Changing Healthcare Professional Behaviours and Patient Outcomes," *Cochrane*

Database of Systematic Reviews, Issue 7. Art. No.: CD009255. DOI: 10.1002/14651858.CD009255.

Fritsche, György, Louis Rusa, Rigobert mpendwanzi, Agnes Soucat, Claude Sekabaranga, Bruno Meesen (2010), "The National Rollout of Performance-Based Financing for Health Services in Rwanda: How It Was Done," World Bank Working Paper.

Glewwe,Paul, Nauman Ilias, and Michael Kremer (2010) "Teacher Incentives" *American Economic Journal: Applied Economics* 2: 1–25.

Health, Development and Performance (HDP) (2008), "Rapport d' Enquête de Contre-Vérification par la Communauté dans les Districts de Nyamasheke, Nyaruguru et Rulindo," Kigali, Rwanda, online: http://www.pbfrwanda.org.rw/index.php?option=com_docman&task=cat_view&gid=24&Itemid=29&limitstart=35.

Institut National de la Statistique du Rwanda (INSR) and ORC Macro (2006), *Rwanda Demographic and Health Survey 2005.* Calverton, Maryland, U.S.A.: INSR and ORC Macro.

Kak, Neeraj, Bart Burkhalter and Merri-Ann Cooper (2001), "Measuring the Competence of Health Care Providers," *QAP Issues Paper*, Vol. 2(1), pp. 1-28.

Kalk, Andreas, Jean Kagubare Mayindo, Laurent Musango, Gerard Foulon. (2005), "Paying for Health in two Rwandan Provinces: Financial Flows and Flaws," *Tropical Medicine and International Health*, Vol. 10(9), pp. 872-878.

Leonard, K.L. and M.C. Masatu, "Using the Hawthorne Effect to examine the gap between a doctors best possible practice and actual practice," *Journal of Development Economics*. 93 (2): 226-243 (2010a)

Leonard, K.L. and M.C. Masatu, "Professionalism and the Know-Do Gap: Exploring Intrinsic Motivation among Health Workers in Tanzania," *Health Economics* 19 (12): 1461-1477 (2010b).

Levine, Ruth and Rena Eichler, Eds. (2009), *Performance Incentives for Global Health*, Brookings Institution Press, Washington, DC.

Logie, Dorothy, Michael Rowson and Felix Ndagije (2008), "Innovations in Rwanda's health system: looking to the future," *The Lancet*, Vol. 372, pp. 256-61.

Miller, Grant, Renfu Luo. Linxiu Zhang, Sean Sylvia, Yaojiang Shi, Patrica Foo, Qiran Zhao, Reynaldo Martorell, Alexis Medina and Scott Rozelle (2012), "Effectiveness of Provider Incentives for Anaemia Reducation in Rual Child: A Cluster Randomized Trial" *British Medical Journal*, Vol. 345.

Ministère de la Santé du Rwanda (1993), *Standards de Prestation des Services au Centre de Santé. Soins Préventifs en SMI/PF /Nutrition*. Volume 1, Première Edition.

Ministère de la Santé du Rwanda (1997), *Normes du District Sanitaire au Rwanda*, Kigali, Rwanda.

Ministère de la Santé du Rwanda (2003) *Normes du District de Santé au Rwanda*, Kigali, Rwanda.

Ministère de la Santé du Rwanda (2006), *Fiche Technique du Programme Elargi de Vaccination*, Kigali, Rwanda.

Ministère de la Santé du Rwanda (2008), *Module de Référence de Formation Continue en Planification familiale. A l'Usage des Formateurs, Superviseurs et Prestataires au Niveau des Formations Sanitaires*, Mars, Kigali, Rwanda.

Ministry of Health. Contractual Approach Unit, Republic of Rwanda. (2006) *Guide for Performance Based Financing. Training module for actors involved in the implementation of the PBF program*. Kigali: Rwanda.

Ministry of Local Government, Community Development and Social Affairs (MINALOC), Republic of Rwanda (2004), *Five Year Decentralization Implementation Programme* 2004-2008, available at http://www.minaloc.gov.rw.

Muralidharan, Karthik and Venkatesh Sundararaman (2011), "Teacher Performance Pay: Empirical Evidence from India," *Journal of Political Economy*, Vol. 119(1), pp. 39-77.

Olken, Benjamin, Junko Onishi, and Susan Wong (2011), "Should Aid Reward Performance? Evidence from a field experiment on health and education in Indonesia," Working Paper, MIT Department of Economics.

Paris, Valérie, Marion Devaux and Lihan Wei (2010), "Health Systems Institutional Characteristics : A Survey of 29 OECD Countries," OECD Health Working Papers, No. 50, OECD Publishing.

Peabody John and Paul Gertler (1997), "Are clinical criteria just proxies for socioeconomic status? A study of low birth weight in Jamaica," *J.Epidemiol.Community Health*, Vol. 51(1), pp. 90-95.

Peabody John, J. Luck, P. Glassman, S. Jain, J. Hansen, M. Spell, et al. (2004), "Measuring the quality of physician practice by using clinical vignettes: a prospective validation study," *Annuals Internal Medicine*, Vol. 141(10), pp. 771-780.

Peabody, John, Riti Shimkhada, Stella Quimbo, Jhiedon Florentino, Marife Bacate, Charles E. McCulloch, and Orville Solon (2011), "Financial Incentives and measurement Improved Physicians' Quality of Care In the Philippines," *Health Affairs*, Vol. 30(4), pp.773-781.

Reinikka, Ritva and Jakob Svensson (2010), "Working for God? Evidence From A Change in Financing of Nonprofit Health Care Providers in Uganda," *Journal of the European Economic Association*, Vol. 8(6), pp. 1159-78.

Scott A, P Sivey, D Ait Ouakrim, L Willenberg, L Naccarella, J Furler, and D Young (2011) "The effect of financial incentives on the quality of health care provided by primary care physicians," *Cochrane Database of Systematic Reviews,* Issue 9, Art. No. CD008451. DOI: 10.1002/14651858.CD008451.pub2.

Soeters, Robert, Laurent Musango and Bruno Meessen (2005). *Comparison of two output based schemes in Butare and Cyangugu provinces with two control provinces in Rwanda*, Global Partnership on Output-Based Aid (GPOBA), The World Bank, Washington, DC.

Soeters, Robert, Christian Habineza and Peter Bob Peerenboom (2006) "Performance based financing and changing the district health system: experience from Rwanda," *Bulletin of the World Health Organization*, November, Vol. 8(11).

United Nations (2010), *The Millennium Development Goals Report*. New York: United Nations.

United Nations (2011), *The Millennium Development Goals Peport*. New York: United Nations.

Van Herck, Pieter, Delphine de Smedt, Lieven Annemans, Roy Remmen, Meredith B Rosenthal and Walter Serneus, "Systematic Review: Effects, Design Choices, and Context of Pay-for-Performance in Health Care," BMC Health Services Research, Vol. 10, pp. 247-260.

Witter Sophie, Atle Fretheim, Flora L Kessy, Karin Lindahl (2012), *Paying for performance to improve the delivery of health interventions in low- and middle-income countries*, *Cochrane Database of Systematic Reviews* 2012, Issue 2. Art. No.: CD007899. DOI: 10.1002/14651858.CD007899.pub2.

World Bank (2008), *World Development Indicators Database, September*, Washington D.C.

World Bank (2010a), "Budgeting for Effectiveness in Rwanda – From Reconstruction to Reform," World Bank Working Paper, No. 205, Washington, DC.

World Bank (2010b), *Rwanda A Country Status Report on Health and Poverty*, Washington, DC.

Table 1: Rwanda P4P Performance Indicators and Payment Rates

		Amount paid per unit (US$)
Visit and Outreach Indicators		
1	Curative care visits	0.18
2	First prenatal care visits	0.09
3	Women who completed 4 prenatal care visits	0.37
4	First time family planning visits (new contraceptive users)	1.83
5	One-month contraceptive resupply visits	0.18
6	Deliveries in the facility	4.59
7	Child (0-59 months) growth monitoring/preventive care visits	0.18
Content of Care Indicators		
8	Children who completed vaccinations on time	0.92
9	Appropriate tetanus vaccine during prenatal care[+]	0.46
10	2nd dose of malaria prophylaxis during prenatal care	0.46
11	Appropriate referral to hospital for delivery[++]	1.83
12	Appropriate Emergency transfers to hospital during delivery[++]	4.59
13	Malnourished child referrals to hospital during preventive care[++]	1.83
14	Other emergency referrals during curative treatment[++]	1.83

[+] Appropriate is defined to any woman who obtains her second, third, fourth or fifth tetanus shot.

[++] Referrals must be confirmed by hospital that patient was treated and referral was necessary.

Source: Ministry of Health. Republic of Rwanda. Guide for Performance Based Financing. Training module for actors involved in the implementation of the PBF program. 2006.

Table 2: Services and Weights Used to Construct the Quality Index for P4P Formula

	Service	Weight	Share of weight allocated to structural measures[+]	Share of weight allocated to process measures[++]	Means of assessment
1	Curative care	0.170	0.23	0.77	Medical record review
2	Delivery	0.130	0.40	0.60	Medical record review
3	Prenatal care	0.126	0.12	0.88	Direct observation
3	Family planning	0.114	0.22	0.78	Medical record review
4	Immunization	0.070	0.40	0.60	Direct observation
6	Growth monitoring	0.052	0.15	0.85	Direct observation
7	HIV services	0.090	1.00	0.00	Direct observation
8	Tuberculosis service	0.028	0.28	0.72	Direct observation
9	Laboratory	0.030	1.00	0.00	Direct observation
10	Facility cleanliness	0.028	1.00	0.00	Direct observation
11	Pharmacy management	0.060	1.00	0.00	Direct observation
12	General administration	0.052	1.00	0.00	Direct observation
13	Financial management	0.050	1.00	0.00	Direct observation
	Total	1.000			

[+] Structural measures are the extent to which the facility has the equipment, drugs, medical supplies and personnel necessary to deliver the listed service.

[++] Process measures capture the clinical content of care provided for the listed service.

Source: Ministry of Health, Republic of Rwanda (2006) "Guide for Performance Based Financing and Training Module for Implementation of the PBF Program." Kigali, Rwanda

Table 3: Facility Characteristics

	Treatment (N=80)		Comparison (N=86)		Difference in Means	P-value
	Mean	St.Dev.	Mean	St.Dev.		
A. Staffing (2006)						
Medical Doctors	0.05	*(0.23)*	0.05	*(0.27)*	0.00	0.94
Nurses	6.31	*(6.90)*	5.48	*(3.30)*	0.83	0.41
Other Clinical Staff	4.13	*(3.09)*	4.47	*(4.05)*	-0.34	0.55
Non-clinical Staff	5.25	*(3.56)*	5.33	*(5.09)*	-0.08	0.90
B. Structural Quality Indices (2006)						
Curative Care	0.80	*(0.07)*	0.81	*(0.07)*	-0.01	0.58
Delivery	0.78	*(0.11)*	0.79	*(0.10)*	0.00	0.84
Prenatal Care	0.96	*(0.15)*	0.97	*(0.11)*	-0.01	0.29
Immunization	0.94	*(0.17)*	0.94	*(0.15)*	0.00	0.90
Laboratory	0.49	*(0.32)*	0.43	*(0.32)*	0.06	0.40
C. Expenditures						
Log Total Expenditures (2006)	15.81	*(1.04)*	15.61	*(1.01)*	0.20	0.42
Log Total Expenditures (2008)	16.91	*(0.71)*	16.99	*(1.08)*	-0.08	0.57
Personnel Budget Share (2006)	0.46	*(0.23)*	0.49	*(0.26)*	-0.03	0.56
Medical Supplies Budget Share (2006)	0.22	*(0.19)*	0.20	*(0.19)*	0.01	0.71
Non-medical Budget Share (2006)	0.32	*(0.25)*	0.30	*(0.22)*	0.02	0.72
D. Prenatal Care Clinical Knowledge (2006)						
Share prenatal care protocol known	0.63	*(0.10)*	0.65	*(0.09)*	-0.02	0.33
Knowledge of protocol (z-score)	0.02	*(0.81)*	0.15	*(0.77)*	-0.13	0.33
E. Prenatal Quality of Care (2006)						
Tetanus vaccine (=1)	0.71	----	0.67	----	0.04	0.33
Share of protocol provided	0.45	*(0.39)*	0.46	*(0.43)*	-0.01	0.66
Clinical protocol provided z-score	-0.13	*(1.49)*	-0.10	*(1.63)*	-0.02	0.76

Notes: P-values are for two-sided tests of the null hypothesis that the difference in means is zero and were calculated using WILD bootstrap with 999 draws. Except for prenatal care quality, the unit of observation is the facility. For prenatal care quality, the unit of observation is the patient visit and the results are based in 1584 observations.

Table 4: Household and Individual Baseline (2006) Characteristics

	Treatment		Comparison		Diff	P-value[+]
	Mean	St.Dev.	Mean	St.Dev.		
A. Household characteristics						
Health insurance (=1)	0.54	---	0.51	---	0.04	0.58
Number of household members	4.92	*(4.40)*	5.00	*(5.16)*	-0.07	0.73
Household-Facility distance (in Km)	3.31	*(6.89)*	3.32	*(8.20)*	-0.02	0.97
Ownership of land (=1)	0.91	---	0.87	---	0.04	0.23
Value of assets	11.20	*(26.74)*	12.59	*(29.44)*	-1.39	0.35
B. Maternal Characteristics and Utilization						
Age < 20 years (=1)	0.03	---	0.02	---	0.01	0.32
Age > 35 years (=1)	0.29	---	0.31	---	-0.02	0.57
Primary education or more (=1)	0.10	---	0.11	---	-0.02	0.47
Living with Partner (=1)	0.94	---	0.91	---	0.04	0.21
Number of pregnancies (Parity)	4.32	*(4.58)*	4.33	*(5.24)*	-0.01	0.97
Any prenatal care (=1)	0.95	---	0.96	---	-0.01	0.77
Made 4 or more prenatal care visits (=1)	0.18	---	0.11	---	0.07	0.03
Number of prenatal care visits made	2.76	*(1.58)*	2.62	*(1.80)*	0.14	0.18
1st Prenatal care visit in 1st trimester (=1)	0.11	---	0.09	---	0.02	0.55
Institutional Delivery (=1)	0.35	---	0.36	---	-0.01	0.80
Use modern contraceptive (=1)	0.09	---	0.13	---	-0.03	0.16
D. Children's Demographic Characteristics						
Age (months)	25.86	*(11.01)*	26.64	*(8.38)*	-0.78	0.06
Female (=1)	0.50	---	0.50	---	0.00	0.87
Maternal height (cms)	157.84	*(14.36)*	158.15	*(16.45)*	-0.31	0.63
Mother's age (years)	31.07	*(12.26)*	31.28	*(13.48)*	-0.22	0.68
Mother completed primary school (=1)	0.09	---	0.12	---	-0.03	0.15
Father present (=1)	0.91	---	0.88	---	0.03	0.33
E. Medical Care Utilization in Last 4 Weeks by Children Age 0-23 Months						
Preventive visit (=1)	0.213	---	0.238	---	-0.025	0.556
Curative visit conditional on illness (=1)	0.305	---	0.266	---	0.039	0.530
F. Medical Care Utilization in Last 4 Weeks by Children Age 24-47 Months						
Preventive visit (=1)	0.084	---	0.140	---	-0.056	0.116
Curative visit conditional on illness (=1)	0.201	---	0.283	---	-0.082	0.124
G. Health outcomes of Children Age 0-11 Months						
Standardized height for age (z-score)	-0.03	*(2.50)*	-0.07	*(2.46)*	0.04	0.87
Standardized weight for age (z-score)	-0.31	*(1.91)*	-0.12	*(2.00)*	-0.19	0.32
H. Health Outcomes of Children Age 24-47 Months						
Standardized height for age (z-score)	-1.95	*(1.79)*	-1.80	*(1.94)*	-0.15	0.22
Standardized weight for age (z-score)	-0.75	*(1.32)*	-0.71	*(1.44)*	-0.04	0.66

Notes: P-values are for two-sided tests of the null hypothesis that the difference in means is zero and were calculated using WILD bootstrap with 999 draws.

Table 5: Impact of P4P on Child Health Outcomes

	Control Mean (2008)	No Controls		With Controls		N
		β	P-Value	β	P-Value	
Children Age 0-11 months						
Height for age Z-Score	-0.20	0.22	0.29	0.16	0.38	800
Weight for age Z-Score	-0.18	0.54	0.02	0.53	0.03	800
Children Age 24-47 Months						
Height for age Z-Score	-1.80	0.23	0.03	0.25	0.00	1957
Weight for age Z-Score	0.69	0.01	0.39	0.03	0.34	1957

Notes: P-Values are for one-sided tests of the null hypothesis that $\beta = 0$ and are calculated based on a WILD bootstrap with 999 draws. Controls include the child's age and sex, maternal height, mother's age, whether the mother has completed primary school, whether the father lives in the household, whether the family is a member of a Mutuelle, the total number of household members, the number of household members under the age of 6, whether the household owns land, and dummy variables for quartiles of the household asset value. Age was entered as a series of dummy variables that represent one-month increments.

Table 6: Rwandan CPG Prenatal Care Items Collected

1.	**MEDICAL HISTORY**
	• High blood pressure
	• Sexually transmitted infections including HIV
	• Tetanus immunizations
	• Pap smear test
	• Tobacco use
	• Alcohol use
2.	**PRIOR PREGNANCIES INFORMATION**
	• Number of previous pregnancies
	• Number of previous miscarriages and stillbirths
3.	**CURRENT PREGNANCY STATUS**
	• Last menstrual date
	• Health problems or concerns during pregnancy
	• Bleeding
	• Weight loss, nausea, or vomiting
	• Medications
4.	**PHYSICAL EXAMINATION**
	• Height
	• Weight
	• Blood pressure
	• Examine abdomen
	• Checked for swelling or water retention
5.	**LABORATORY EXAMINATIONS**
	• Take a blood sample for anemia test
	• Take a urine sample for gestational diabetes test
	• Test for current STI including HIV
6.	**PREVENTION AND CASE MANAGEMENT**
	• Tetanus toxoid injection
	• Iron/vitamin pills
	• Plan delivery

Table 7: Impact of P4P on Knowledge and Quality of Care

	Knowledge Z-Score (Standardized)		Quality of Care Z-Score (Standardized)					
			Model 1		Model 2		Model 3	
	β	P-Value	β	P-Value	β	P-Value	β	P-Value
P4P	0.40	0.12	0.16	0.04	0.13	0.01	0.07	0.14
Knowledge Z-Score					0.03	0.21	0.02	0.36
P4P * Knowledge in Top 50%							0.142	0.06
N Observations	294		3,709		3,709		3,709	

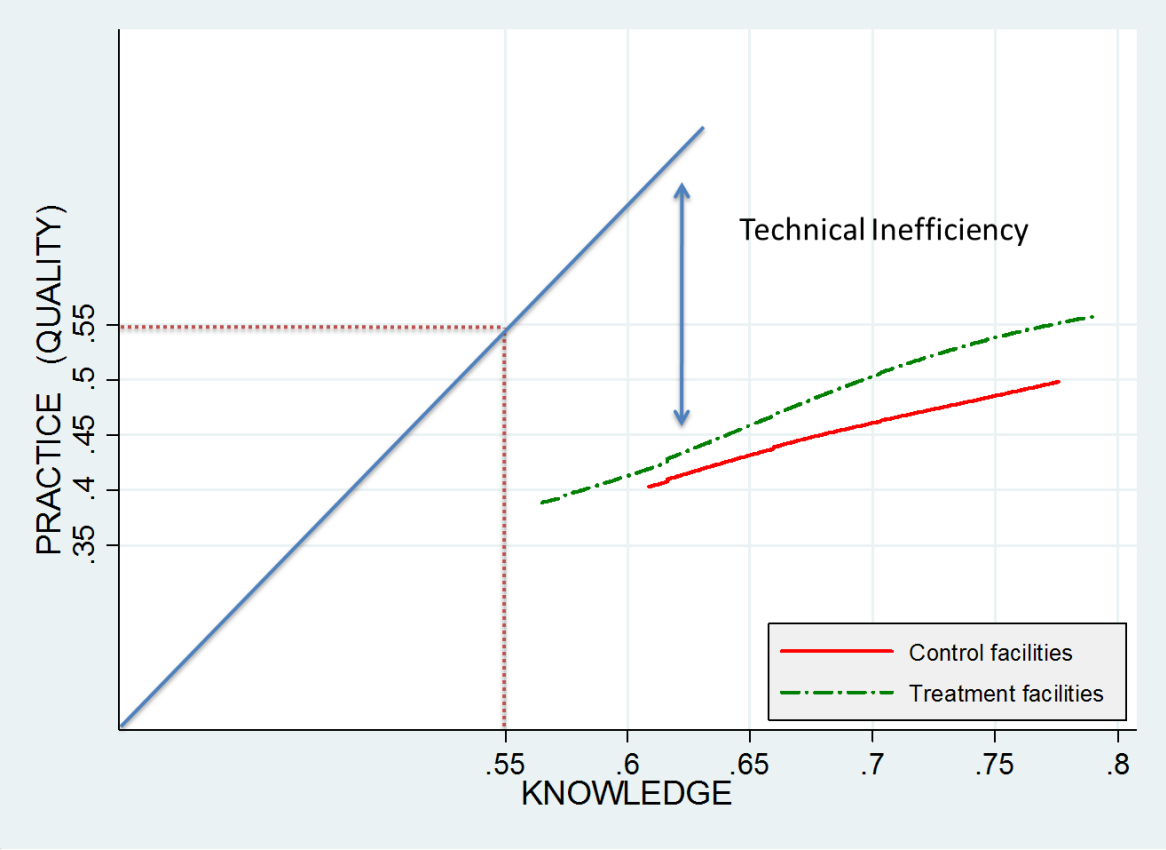
Notes: P-Values are for one-sided tests of the null hypothesis that $\beta = 0$ and are calculated based on a WILD bootstrap with 999 draws. Knowledge share is a continuous variable of the share of the CPG items that the provider knew and is bounded between zero and 1.

Table 8: Impact of P4P on Inefficiency (Knowledge Share – Quality Share)

	Model 1		Model 2		Model 3	
	β	P-Value	β	P-Value	β	P-Value
P4P (=1)	-0.035	0.00	-0.04	0.03	-0.02	0.24
Knowledge Share			0.16	0.00	0.21	0.00
P4P * Knowledge in Top 50%					-0.06	0.01
N Observations	3709		3709		3709	

Notes: P-Values are for one-sided tests of the null hypothesis that $\beta = 0$ and are calculated based on a WILD bootstrap with 999 draws. Knowledge Share is the share of CPG protocol items correctly identified by the provider during the administration of the vignette. Quality Share is the percentage of CPG protocol items that were delivered during prenatal care, as reported in patient exit interviews and in household surveys.

Figure 1: The Knowledge-Quality Efficiency Gap for Prenatal Care (2008)



Notes: The horizontal axis is Knowledge expressed as the percentage of CPG protocol items correctly identified by the provider during the administration of the vignette. The horizontal axis is the percentage of CPG protocol items that were delivered during prenatal care, as reported in patient exit interviews and in household surveys.

Appendix Table A: Impact of P4P on Health Care Utilization and Quality of Care

	Control Mean (2008)	No Controls		With Controls		N
		β	P-Value	β	P-Value	
Maternal Health Care Utilization						
Any Prenatal Care (=1)	0.98	0.00	0.46	0.00	0.45	2309
4 or more Prenatal Care Visits (=1)	0.25	0.01	0.44	0.01	0.44	2223
Institutional Delivery (=1)	0.50	0.07	0.06	0.08	0.04	2108
Use Modern Contraception (=1)	0.35	0.02	0.32	0.02	0.27	3154
Utilization by Children Age 0-23 Months						
Preventive Care Visit (=1)	0.48	0.13	0.02	0.12	0.03	1971
Curative Care Visit (=1)	0.34	-0.00	0.50	-0.02	0.30	986
Utilization by Children Age 24-47 Months						
Preventive Care Visit (=1)	.24	0.11	0.00	0.11	0.00	2902
Curative Care Visit (=1)	0.42	0.10	0.13	0.08	0.18	1229
Quality of Prenatal Care						
Tetanus Vaccine (=1)	0.66	0.05	0.06	0.05	0.07	2856
Clinical Protocol Z-Score	0.00	0.16	0.06	0.16	0.04	3826

Notes: P-Values are for one-sided tests of the null hypothesis that $\beta = 0$ and are calculated based on a WILD bootstrap with 999 draws. Controls for child utilization include the child's age and sex, maternal height, mother's age, whether the mother has completed primary school, whether the father lives in the household, whether the family has health insurance, the total number of household members, the number of household members under the age of 6, whether the household owns land, and dummy variables for quartiles of the household asset value. Age was entered as a series of dummy variables that represent one-month increments. Controls for the maternal utilization and quality regressions include whether the woman is younger than 20 years, is older than 35, has at least primary education, is currently married, in union or living with the partner. Controls also include the number of household members, the number of prior pregnancies, and the distance from the household to the health facility whether the household had health insurance, whether the household owns any land, and quartiles for the household asset value.