Multitasking and Heterogeneous Treatment Effects in Pay-for-Performance in Health Care: Evidence from Rwanda

Tisamarie B. Sherry
Brigham and Women's Hospital
tsherry@partners.org

Sebastian Bauhoff
Center for Global Development
sbauhoff@cgdev.org

Manoj Mohanan
Duke University
manoj.mohanan@duke.edu

March 2016

Abstract: Performance-based contracting is particularly challenging in health care, where multiple agents, information asymmetries and other market failures compound the critical contracting concern of multitasking. As performance-based contracting grows in developing countries, it is critical to better understand not only intended program impacts on rewarded outcomes, but also unintended program impacts such as multitasking and heterogeneous program effects in order to guide program design and scale-up. We use two waves of data from the Rwanda Demographic and Health Surveys collected before and after the quasi-experimental roll-out of Rwanda's national pay-for-performance (P4P) program to analyze impacts on utilization of healthcare services, health outcomes and unintended consequences of P4P. We find that P4P improved some rewarded services, as well as some services that were not directly rewarded, but had no statistically significant impact on health outcomes. We do not find evidence that clearly suggests multitasking. We find that program effects vary by baseline levels of facility quality, with most improvements seen in the medium quality tier.

We are grateful to William Beardslee, Katherine Donato, Richard Frank, Martin Hoff, Mireille Jacobson, Thomas McGuire, Marcos Vera Hernandez, and Alan Zaslavsky, and participants at Harvard HCP Seminar, Harvard Health Economics Workshop, iHEA 2011, U. Washington/ IHME, VCU Health Policy Seminar, the World Bank, and Yale HCP Colloquium for comments and suggestions.

1. Introduction

Contracts that link payments to performance can align the incentives of agents with the objectives of principals (Holmstrom and Milgrom 1991) and are a key feature of compensation schemes in numerous industries, notably health care where they are termed "pay-for-performance" (Miller and Babiarz 2014). The design of a successful P4P scheme faces many challenges: P4P may have unintended impacts if agents engage in multitasking (i.e., focusing effort on highly remunerated tasks at the expense of tasks that are rewarded less generously) or skimp on dimensions of care that are more difficult to observe or measure such as unobservable aspects of quality (Prendergast 1999). Moreover, contracts that reward multiple activities may even fail to increase effort spent on specific rewarded services because of countervailing incentives for other rewarded activities (Sherry 2016). Agents may respond differently to P4P according to their baseline level of performance (Rosenthal et al. 2005, Petersen et al. 2006, Rosenthal and Dudley 2007), further complicating predictions about how such programs will fare when implemented on a broader scale. Understanding how contracted agents respond to incentive contracts – in both intended and unintended ways – is therefore essential to informing the design of P4P programs. Yet the adoption of P4P programs in health care has grown so rapidly that it has outpaced empirical evidence on their impacts on health care quality, health outcomes, and fundamental economic mechanisms underlying performance-based contracting.

In this paper we examine the impact of Rwanda's national P4P scheme on maternal and child health care and family planning services, with a focus on unintended consequences of P4P, including multitasking by providers and heterogeneity in the effects of P4P. Performance-based financing in health care has expanded rapidly in middle and lower-income countries, and in

Africa in particular (Fritsche, Soeters, and Meessen 2014, Miller and Babiarz 2014). Rwanda's P4P program is notable in that it was among the earliest such initiatives to be implemented on a national scale and in such a way as to allow rigorous evaluation, and has influenced the design of numerous other P4P initiatives in lower-income settings – lessons from Rwanda therefore have implications for the management of a number of existing P4P programs in resource-limited countries, as well as the design of future initiatives in these settings.

We leverage the quasi-experimental roll-out of the program across districts in conjunction with household data from two waves of the Rwanda Demographic and Health Surveys (DHS), collected before and after the initial roll-out, to estimate the impact of P4P on the outcomes of interest in a difference-in-differences model. First, we examine the effect of P4P on rewarded measures, and health outcomes and behaviors. Second, we examine the effect of P4P on measures that are recommended as part of Rwanda's national clinical guidelines but were not directly rewarded, to test for multitasking. Finally, we explore heterogeneities in program impact across areas with different levels of baseline quality. These questions are of particular interest in the Rwandan context due to a distinctive feature of Rwanda's P4P incentive scheme: all P4P bonus payments to a given health care facility are scaled by a "quality multiplier" reflecting that facility's overall quality. We test four key hypotheses based both on the recent theoretical literature on P4P and the particular structure of the Rwandan payment formula: (a) P4P leads to an increase in the provision of the most generously rewarded services, as previously reported; (b) P4P leads to an increase in the provision of certain services that are not directly rewarded, but that form part of the quality multiplier and are therefore indirectly incentivized, and/or share commonalities in production with generously rewarded services; (c)

P4P leads to improved health outcomes and behaviors associated with the rewarded services; and (d) improvements in service provision vary according to baseline levels of facility quality.

We find that the P4P program had mixed effects on service provision. Among the 9 services that were directly rewarded by the program and are observable in the DHS data, the program significantly increased the share of women delivering in facilities by almost 10 percentage points (against a baseline share of 30%), and the prevalence of contraception use among women who want to space or limit births by 4 percentage points (baseline 1.6%). These are two of the more highly rewarded services under P4P. There were no significant impacts on less generously rewarded maternal and child health services. We discuss possible explanations, including low real incentives (i.e. rewards net of costs), increasing marginal costs, features of the health care production function, and demand-side factors.

Among the services that were not directly rewarded and that the DHS collects information on, P4P increased urinalysis during prenatal care by 5 percentage points against a baseline of 9% and increased iron supplementation by 9 percentage points (baseline of 30%), but had no significant impact on other unrewarded prenatal care measures or child vitamin A supplementation. This suggests that providers did not skimp on activities that were individually unrewarded but were indirectly incentivized through the quality multiplier. We also find that P4P had no detectable impact on health outcomes and behaviors available in the DHS data, such as the tested prevalence of anemia in mothers and children, and reported breastfeeding. Finally, we find evidence of heterogeneity in the impacts of P4P, with facilities in the middle of the baseline quality distribution generally showing larger improvements in both directly rewarded services, and those that were individually unrewarded but related to the quality multiplier.

Our analysis makes several contributions to the evidence base on P4P. First, we provide evidence that performance-based financing can improve some aspects of health care quality in a resource-limited setting. Using the DHS permits us to assess the impacts of Rwanda's P4P program on a larger scale than previous studies, and on an expanded set of rewarded measures, including those not explored in previous research. Second, we examine impacts of P4P on health outcomes and behaviors, the ultimate policy objective of the program. Third, ours is the first study to provide empirical evidence on unintended consequences of Rwanda's P4P program that are nonetheless critically important to the design of these initiatives, such as program impacts on services that were not directly rewarded. The challenges of multitasking and other unintended consequences of P4P have been investigated in the US and Europe, but to our knowledge ours is the only study to address these important questions in a lower-income country. Ours is also the only study to explore the implications of Rwanda's quality multiplier – a distinctive feature of the incentive structure that has since been replicated in other countries – for multitasking. Finally, we explore differences in the impact of Rwanda's national P4P program according to baseline levels of facility quality. This question is of key policy relevance in informing the design and scale-up of other large-scale performance-based financing initiatives, but has also received little attention in prior studies of P4P in both lower and higher-income settings.

The remainder of the paper proceeds as follows: in Section 2 we discuss intended and potential unintended consequences of P4P programs, and the evidence to date on their impacts. Section 3 describes Rwanda's P4P program and its quasi-experimental roll-out, and motivates the hypotheses we aim to test in this paper. Section 4 provides details on the data. Sections 5 and 6 describe our empirical strategies and present our results, and Section 7 discusses our findings in light of the theoretical predictions.

2. Evidence on the Intended and Unintended Consequences of Pay-for-Performance

Performance-based contracts can have complex effects on both rewarded and unrewarded activities (Holmstrom and Milgrom 1991). In a Holmstrom-Milgrom production function where two activities are substitutes in an agent's cost function (i.e. effort devoted to one activity increases the marginal cost of effort devoted to the other activity), when only one of these activities is rewarded with incentive pay its output is expected to increase at the expense of the unrewarded activity. This multitasking problem arises when only a subset of providers' activities are rewarded – a setup common in P4P programs in health care (Prendergast 1999). The classic intuition of multitasking may break down, however, when P4P rewards multiple targets (Mullen, Frank, and Rosenthal 2010, Sherry 2016) because the incentives to increase some rewarded activities may be weakened by countervailing incentives to increase other rewarded activities, such that the net impact of P4P on any given rewarded service is ambiguous. In the presence of multiple targets, P4P therefore does not necessarily increase the output of a given rewarded service. P4P's impacts are further complicated by joint production or "commonality" (Glazer, McGuire, and Normand 2008, Mullen, Frank, and Rosenthal 2010), another common feature of health care settings in which a common input affects the output of multiple services. Examples that are relevant to the Rwandan setting include the availability of clean syringes as a common input in both childhood immunizations and prenatal tetanus immunizations; or HIV voluntary counseling and testing services as a common input in both HIV treatment and the dispensation of modern contraceptive methods that limit the spread of sexually transmitted infections (e.g. condoms). In the presence of significant joint production, P4P may actually increase the supply of unrewarded services. Since production functions in health care settings are likely to include some properties associated with both multitasking (i.e., effort devoted to one task increasing the

cost of effort spent on other tasks) and joint production (i.e., inputs contributing to more than one output), contrary to the classic view that P4P should increase the output of rewarded services and decrease the output of unrewarded services the impact on both is in fact ambiguous (Sherry 2016). The net effect of P4P is therefore an empirical question, and will depend on relative prices, the degree of multitasking and the existence of commonalities in production (Mullen, Frank, and Rosenthal 2010, Sherry 2016). P4P will be more likely to increase the output of services that (1) earn a higher bonus payment relative to their production costs (McGuire and Pauly 1991); and (2) are jointly produced with other highly rewarded services.

In light of the ambiguity of theoretical predictions, it is not surprising that the evidence on the impact of targeted performance incentives on physician behavior has been mixed. In the US, most studies have found no or at best modest improvements in quality, but also no disruptions in care (Mullen, Frank, and Rosenthal 2010, Rosenthal and Frank 2006, Christianson, Leatherman, and Sutherland 2008). Even in the context of single-payer settings like Ontario, Canada, P4P incentives have led to only modest improvements in some rewarded services and no improvements in others (Li et al. 2011). The UK's P4P program for family practitioners has also been associated with only modest short-term improvements in quality indicators (Campbell et al. 2007, Campbell et al. 2009) despite its comprehensive approach and substantial financial incentives: under the P4P contract, high-performing family practitioners stood to increase their income by up to 25% (Doran et al. 2006).

Only a limited number of studies have explicitly tested for unintended consequences of P4P such as multitasking (Mullen, Frank, and Rosenthal 2010, Glickman et al. 2007). These studies have focused on high-income settings and the findings have been mixed. The UK's

national P4P program, for example, witnessed improvements in some unpaid indicators, but declines in other unpaid measures (Campbell et al. 2009, Steel et al. 2007, Sutton et al. 2010). Heterogeneity in P4P's impacts also remains an important yet understudied area. There are concerns that the incentive structures used by many P4P schemes may discourage improvement by the lowest and/or highest performers (Rosenthal et al. 2005, Petersen et al. 2006, Rosenthal and Dudley 2007). High and low performing facilities are likely to have different marginal costs of production, so their responses to different types of incentives could vary considerably. Understanding the relationship between baseline performance and the impact of P4P is therefore especially important as these initiatives are scaled up. This issue is particularly relevant to Rwanda's P4P program, where it is not only the marginal costs but also the incentive size that varies between high and low performing facilities, as described later.

P4P has a growing presence in middle- and low-income countries, and there is emerging evidence on its impacts in these settings (Borghi et al. 2015, Van de Poel et al. 2015, Miller and Babiarz 2014). Evaluations of P4P programs in such diverse countries as the Philippines, Indonesia, the Democratic Republic of the Congo and Burundi have found improvements in provider knowledge (Peabody et al. 2011), increases in healthcare workers' labor (Olken, Onishi, and Wong 2014), increases in utilization of care and improvements in in some measures of health care quality (Soeters et al. 2011, Bonfrer, Van de Poel, and Van Doorslaer 2014). Still, there exist gaps in our understanding of P4P's impacts in these settings. As Miller and Babiarz (2014) note , there is little evidence on key conceptual issues in the design of P4P programs, such as the choice of which outcomes to reward, the structure of incentive schemes, and unintended consequences of P4P. How P4P programs interact with existing incentive structures and social norms, and their impacts on equity, also remains largely unstudied. Finally, from a

methodological standpoint, in some studies it has not been possible to disentangle the effects of P4P financial incentives from other quality improvement initiatives (Soeters et al. 2011).

Previous studies of Rwanda's national P4P program showed improvements in rewarded services that had higher unit payments and were easier for providers to control such as institutional deliveries (i.e. deliveries in health care facilities) and HIV testing (Basinga et al. 2011, de Walque et al. 2015), and positive impacts on certain health outcomes, specifically children's weight-for-age and height-for-age (Gertler and Vermeersch 2013). This paper expands on existing research by examining the impacts of P4P on a broader set of rewarded measures using a larger and independently collected data set, the Demographic and Health Surveys (DHS). It is also the first study to explicitly test for unintended consequences of P4P such as multitasking and heterogeneity in program impacts by baseline facility quality, as well as the first study to consider the role of the quality multiplier – a distinctive feature of Rwanda's P4P incentive scheme that has since been replicated in numerous other settings – in explaining the patterns of impacts observed. As detailed below, we follow a similar empirical strategy to earlier work on Rwanda's program, leveraging the quasi-experimental roll-out of P4P in a differences-in-differences estimation.

3. Rwanda's National P4P Program

Rwanda's national P4P scheme was launched in May 2006, with the goal of improving the quality of care provided in the country's 38 district hospitals and 420 district health facilities (Basinga et al. 2011). We focus on the program implemented in district health facilities (hereafter "facilities"), which provide the majority of primary care services (Hoff 2010).

3.1 The P4P Program

Prior to 2006, facility budgets were input-based, meaning facilities were allocated funds prospectively by the Ministry of Health according to the anticipated amount of physical and human capital inputs required to serve their catchment populations (Hoff 2010). The P4P program supplemented these budgets with bonus payments to facilities based on their performance on 24 output and quality indicators. The unit payments for rewarded indicators are shown in Table 1. The size of the payments varies considerably across services, from US $0.09 for each first-time prenatal care visit to US $8.93 for HIV testing of each exposed child. In general, institutional deliveries and HIV/AIDS care are the most generously rewarded services under the P4P scheme. P4P bonus payments are disbursed to facilities quarterly and the facilities may allocate the payments at their discretion. In the first two years, the bonuses increased facilities' budgets by 22% on average. Overall, 77% of the bonuses were used to increase compensation, resulting in a 38% increase in staff salaries (Basinga et al. 2011). The remaining 23% was typically spent on infrastructure (i.e. physical attributes of the facilities/buildings themselves) or health care inputs (i.e. medical supplies).

A distinctive feature of Rwanda's P4P program, which has subsequently been adopted in P4P schemes in other middle- and low-income countries (e.g. Burundi) but has not been previously examined, is that bonus payments are scaled by a quality multiplier M based on a facility's overall performance. M may take any value between 0 and 1, where higher values indicate higher overall facility quality. Overall quality is assessed on a broad range of service categories such as infrastructure, personnel, supplies, and adherence to clinical practice guidelines, as listed in Table 1. Quality scores for each category are determined by district health

officials during unannounced quarterly assessments. The overall quality multiplier *M* is a weighted average of the quality scores across all service categories.

The overall bonus payment formula therefore scales the total payment to each facility in a payment period by the value of the quality multiplier *M* assigned to each facility for that period:

$$Bonus = M \times \left(\sum_{n=1}^{24} p_n \, r_n \right) \qquad (1)$$

Where r_n is the number of times rewarded indicator *n* is met and p_n is the corresponding unit payment for rewarded indicator *n*, and *M* is the value of the quality multiplier.

The quality multiplier has several interesting implications for Rwanda's P4P program that we explore in this study. First, divergence between the directly targeted services *n* and the services that are part of the broader quality multiplier *M* can potentially mitigate the standard multitasking problem and even promote broader health care quality improvements beyond the directly targeted services, as facilities have an incentive to raise *M* in order to increase the marginal revenue earned on each directly rewarded service *n*.

Second, because the quality multiplier is based on a very broad array of service categories covering virtually all of a facility's clinical activities (e.g. pharmacy, laboratory services, general administration, curative care, preventive care, prenatal care, delivery, etc.), nearly any service can contribute to increasing *M* and therefore could be considered "indirectly incentivized" as a result of the quality multiplier. The directly rewarded services are likely more salient to providers, but since other clinical activities can also be revenue-generating through a separate channel (the quality multiplier), in this study rather than referring to these non-targeted

services as "unrewarded", we view them as indirectly or partially incentivized. As shown in Section 2, multitasking can occur wherever relative rewards of services differ, and thus remains relevant in this context.

Third, the quality multiplier creates varying incentives for facilities with different levels of overall quality. Initially, higher-performing facilities with quality multipliers closer to 1 face the highest marginal incentive for each unit of directly rewarded service. Lower-performing facilities, on the other hand, face lower marginal incentives for each unit of a directly rewarded service, but face strong incentives to increase their quality multiplier value to eventually earn larger overall P4P bonus payments. We therefore expect facilities with higher baseline performance to achieve larger absolute increases in services that are directly rewarded under P4P relative to facilities with lower baseline performance, and we expect facilities with lower baseline performance to achieve larger increases in individually unrewarded services that still contribute to the quality multiplier *M*. Without knowledge of the facilities' production functions, it is difficult to predict further the relative influence of the incentive payments.

3.2 The Quasi-Experimental Roll-Out of P4P

The Rwandan Ministry of Health instituted a phased and quasi-experimental roll-out of the national P4P program to facilitate evaluation. Prior to the national roll-out, several non-governmental organizations implemented experimental P4P programs in 11 of Rwanda's 30 "Phase 0" districts between 2002 and 2006. Basinga et al. (2011) describe the block-randomization of the remaining 19 districts into treatment and control districts, based on rainfall and Census population density and livelihood measures. Prior to implementation of the planned P4P program, the government redrew district administrative boundaries as part of a national

decentralization program (Ministry of Local Government 2004). The decentralization was unrelated to health outcomes in the districts or the P4P program, but it resulted in several control areas being combined with areas with existing P4P pilots since Phase 0 to form new districts (Gertler and Vermeersch 2013). The Rwandan government required that these new combined districts be designated as "treatment" districts since P4P was managed at the district-level, and therefore health facility financing systems had to be uniform within a given district – this necessitated switching the assignment of several control districts to "treatment". We provide further details on district randomization and decentralization in Appendix A.2. Because the redistricting compromised the original randomization, we refer to the rollout as quasi-experimental and address methodological concerns below.[1]

The final assignment resulted in 12 treatment districts and 7 control districts. Together with the 11 districts that had implemented pilot programs in Phase 0, the 12 treatment districts adopted the national P4P scheme between June and October 2006 in Phase 1. In Phase 2, from April through May 2008, the program was rolled out to the 7 control districts. To isolate the effect of giving facilities performance incentives from the effect of simply increasing their available resources, during Phase 1 the Ministry of Health increased the control facilities' input-based budgets by the average bonus payment to the treatment facilities.[2] In this study, we

[1] The goal of the decentralization program was to grant more autonomy and responsibility for the administration of health and other social services to district governments. This required redrawing district boundaries to ensure that each district had a designated hospital and adequate primary care network. Redistricting resulted in the combination of some control areas with areas with existing P4P pilots – therefore the "treatment group" includes some localities with longer experience with P4P. The specific districts affected by these redistricting changes are unknown – this information is not publicly available.

[2] Basinga et al. (2011) show in Table 5 that there was no significant difference in log total expenditures between intervention or control facilities in 2006 or 1008, indicating that facilities

analyze the impacts of P4P on the 19 originally randomized districts, as described in Basinga et al. (2011), since the Phase 0 initiatives were not randomly assigned and varied in their approaches.[3]

Rwanda's P4P program therefore incorporates elements that are common across many P4P schemes in lower-income countries such as Burundi, Cambodia, Democratic Republic of Congo, and Zambia, but also has several distinctive features: the quality multiplier, the quasi-experimental roll-out to facilitate evaluation, and the payments given to control facilities to allow the separation of income and substitution effects. Based on these distinctive features of Rwanda's payment formula and the broader theoretical literature on P4P, we test the following hypotheses: (a) P4P leads to an increase in the provision of the most generously rewarded services, as previously reported; (b) P4P leads to an increase in the provision of certain services that are not directly rewarded, but that form part of the quality multiplier and/or share commonalities in production with generously rewarded services; (c) P4P leads to improved health outcomes and behaviors associated with the rewarded services; and (d) improvements in service provision vary according to baseline levels of facility quality.

4. Data

The Rwanda Demographic and Health Surveys (DHS) is a large, nationally representative household survey that collects cross-sectional data on health and demography. We use two recent waves of the DHS, DHS-III and DHS-IV, for our analysis. DHS-III surveyed 11,321 women

---

received similar amounts of funding in both the pre- and post-intervention periods and that the incentive effect was therefore isolated.

[3] Basinga et al. (2011) and Gertler and Vermeersch (2013) describe further details of the study protocol. Our results are robust to inclusion of these early adopter Phase 0 districts in our estimations (results available upon request).

(ages 15-49) drawn from 10,272 households, between February and July of 2005. DHS-IV surveyed 7,313 women drawn from 7,377 households, between December 2007 and April 2008[4]. DHS-III was collected prior to the roll-out of P4P to the treatment districts, while DHS-IV was collected 18-22 months after this initial roll-out, but prior to the expansion of P4P to control districts. Each wave collects information from female respondents about current health behaviors, any pregnancies in the preceding 5 years, and the health of children born in the preceding 5 years. Depending on the outcome variable of interest, we delineate time periods using either the birth dates of children (e.g., for institutional delivery) or the interview date (e.g., for current use of contraception). Appendix A.1, Figure A.1 provides a graphical timeline.

The primary sampling units used in DHS-III and DHS-IV are enumeration areas (EAs) from Rwanda's 2002 Census and represent large villages or clusters of villages. In DHS-III, 432 EAs were sampled, while in DHS-IV a subset of 250 of these EAs were sampled. Our analytical sample consists of a balanced panel of EAs (i.e. villages or localities) containing two repeated cross-sections of households – i.e., although the same localities are visited in each wave, sample households within these localities are visited only once in either the pre- or post-intervention periods. We restrict our analysis to households and individuals in the 89 EAs from the 12 treatment districts, and 64 EAs from the 7 control districts that were sampled in both DHS-III and DHS-IV, thus allowing us to control for local-area fixed effects.[5]

[4] While the DHS-III was a standard, full-length DHS, the DHS-IV was an interim survey with a smaller sample size and abbreviated questionnaire.

[5] We match EAs using geographic identifiers contained in both waves of the data. We describe in Appendix A.3 our approach to identifying households that may have been interviewed in both waves of the DHS and therefore double-counted, and we find very few potential duplicate households. Although the geographic identifiers in the two DHS waves match up exactly, the coordinates in the public use files contain random positional error to maintain confidentiality.

As a result of varying reference populations across the two waves of DHS and the different recall periods across measures, the analyses include different sample sizes across different outcome variables. For example, information about institutional deliveries is available for all births within the past five years while maternal anemia is only measured for women who are currently pregnant. We use the maximum available sample for each dependent variable. Table 2 summarizes the baseline characteristics of women and children in the treatment and control areas, prior to the roll-out of P4P. Several of the unadjusted pre-intervention measures differ across the treatment and control groups, but these differences are not statistically significant and neither group performs clearly better than the other in the initial period. Our data are broadly comparable with those used in earlier research (Basinga et al. 2011): female household members have very low education, predominantly live with a partner, and had between 4 and 5 births at the time of the interview. Households have an average of 5-6 members, and slightly more than half have insurance.

The DHS has several advantages over the data employed in previous studies of Rwanda's program. It is a large dataset that provides greater sample size for estimation of a number of the outcomes studied; it is representative of the study areas, better reflecting impacts of P4P on overall population health rather than only facility-level outcomes; and is collected independently of the P4P intervention. Since we exclusively use household reports, our measures are not

---

Urban and rural clusters are displaced by up to 2km and 5km, respectively; an additional 1 percent of rural clusters are randomly displaced by up to 10km. This could lead to misclassification if apparent control EAs are in fact located in treatment areas and vice versa. In a robustness check (available upon request) we exclude EAs whose treatment assignment would be altered by this positional error (i.e. urban EAs from control districts within 2 km of the boundary with a treatment district; urban EAs from treatment districts within 2 km of the boundary of a control district; rural EAs from control districts within 5 km of the boundary with a treatment district; and rural EAs from treatment districts within 5 km of the boundary of a control district). The broad findings remain comparable.

subject to disproportionate over-reporting by providers in the treatment group, nor are they subject to selection bias based on the possibility that patients seeking care at facilities in the pre-intervention period might differ systematically and in unobservable ways from other patients in the facility catchment areas.

5. Methods

5.1 Average Impact of P4P

We estimate the impact of Rwanda's national P4P program on the outcomes of interest using a difference-in-differences (DD) model:

$$y = \beta_0 + \gamma \times POST + \delta \times POST \times TREAT + \beta_1 \times C + \alpha + \theta + \varepsilon \qquad (2)$$

where y is the outcome of interest; *POST* is an indicator variable for the post-intervention period[6, 7]; *TREAT* identifies the treatment districts so that δ is the coefficient of interest; C is a vector of household and individual-level control variables; α is a vector of EA-level fixed effects; and θ is a vector of birth-year fixed effects.[8] Standard errors are clustered at the district

[6] The *POST* indicator equals 1 for outcomes y occurring during or after the roll-out of P4P (i.e. June 2006 or later), and 0 for outcomes occurring before the roll-out. This approach should yield conservative estimates of P4P's impacts, as P4P did not start in all treatment districts at once but was rolled-out between June and October 2006; during the roll-out period, the post-period sample will therefore include some individuals who have not yet been exposed to P4P. Robustness checks that exclude outcomes occurring during the roll-out period yield similar results and are available on request.

[7] Note that the coefficient on *POST* captures both pure income effects in the control group facilities as a result of the payments they received from the Ministry of Health, and other secular changes in the control group facilities. It is not possible to disentangle these other secular changes from the pure income effects.

[8] Our covariate set compares with the most complete specification in Basinga et al. (2011). Birth-year fixed effects are omitted from the models estimating the impact of P4P on contraceptive coverage.

level, i.e. the level of the treatment. Because we only have G=19 clusters, we also use critical values from a *t*-distribution with G-2 degrees of freedom (Cameron, Gelbach, and Miller 2008, Cohen and Dupas 2010). The critical *t*-values for the 1%, 5% and 10% significance levels are 2.90, 2.11 and 1.74, respectively.

The identification assumption is that in the absence of P4P, the evolution of outcomes in the treatment districts would have been comparable to the observed changes in the control districts. The EA-level and birth-year fixed effects eliminate area-specific unobservable factors (such as distance to facilities and local attitudes) and time-variant unobservable factors (e.g., national policies and changes in the environment that may have influenced nutrition and health) that are common to both types of districts. One potential concern is the expansion of community-based health insurance over the study period[9]. As noted by Basinga et al (2011) , the insurance expansion occurred at a national level and did not differ systematically across P4P treatment and control districts; nonetheless we control for household-level health insurance coverage to eliminate confounding due to coincidental overlap of the insurance expansion with the P4P roll-out. An analytical benefit of this expansion is that the insurance schemes are coordinated and operated on the district level: until very recently, and during the period of time included in our study, households enrolled in Rwanda's community-based health insurance program were affiliated with a specific health center, and only from there could obtain referrals to other district health centers or hospitals (Lu et al. 2012). This should serve to reduce potential

[9] The community-based health insurance scheme ("Mutuelles") is an initiative supported by the Rwandan Ministry of Health to improve the affordability of basic health services. Households enroll in the scheme, then are assigned to a designated health center from which they may receive care. Households pay an annual premium and a copay for each visit to a health facility (Lu, Chin et al 2012).

spillovers from individuals residing in control districts seeking care in treatment districts and vice versa (Ministry of Health of Rwanda 2010).[10]

As discussed above, the redistricting compromised the original randomization of districts, introducing the possibility that final treatment assignment might be correlated with unobservable characteristics. We make use of the final treatment assignment for several reasons. First, the difference-in-difference design that we employ in this paper subsumes time-invariant unobservable characteristics. Second, in line with related research using household and facility data, we find that a wide range of covariates are balanced at baseline (Table 2). Finally, our approach facilitates comparisons with earlier studies.[11]

5.2 Heterogeneity by Pre-Intervention Quality

In the second part of our analysis, we examine how the impact of P4P varies according to an area's pre-intervention health care quality. Data on the quality multiplier *M* for each facility is not available, and it is not possible to precisely match DHS data on health service utilization to individual facilities or nearby facilities since the DHS does not identify these facilities in the data, nor does it provide geographic coordinates for facilities due to privacy considerations.

[10] As described in footnote 3, in a robustness check (available upon request) we exclude EAs in close proximity to the boundaries between treatment and control districts (i.e. urban EAs from control districts within 2 km of the boundary with a treatment district; urban EAs from treatment districts within 2 km of the boundary of a control district; rural EAs from control districts within 5 km of the boundary with a treatment district; and rural EAs from treatment districts within 5 km of the boundary of a control district), to reduce these potential spillovers. The broad findings remain comparable. We cannot directly test for changes in travel time and patterns, as this information is not available in the DHS.

[11] One possible analytical strategy - given the reassignment - was to implement the Intent to treat (ITT) estimator. However, as also argued by Gertler and Vermeersch (2013), the decision to reassign districts in Rwanda was made prior to the evaluation and it was not a result of choices made by respondents or participants (facilities) in this study.

Instead, we construct a score Q for each enumeration area (EA), which reflects the average extent of recommended health care coverage in that EA, and we use Q as a proxy for the quality of the nearest facility or facilities.[12]

To calculate Q, for each EA we determine the percentile in the pre-roll-out period for: facility deliveries; pregnant women receiving appropriate tetanus vaccination; pregnant women receiving appropriate malaria prophylaxis; pregnant women completing four prenatal visits; children receiving vitamin A supplementation; children fully immunized; and women living with a partner who are using a modern method of contraception.[13] These are representative measures available in DHS data from each of the service categories that we examine in this analysis: delivery and prenatal care, child health and family planning. The pre-intervention coverage score Q for a given EA is the average of these pre-intervention percentiles.

EAs are then assigned to three tiers (low, medium, and high) according to their coverage score *Q*, with the bottom third of EAs in the low tier, the middle third in the middle tier, and top third of EAs in the high tier. We then repeat our difference-in-differences analysis using the following model:

$$y = \beta_0 + \sum_{i=1}^{3} \gamma_i \times POST \times \bar{Q}_i + \sum_{i=1}^{3} \delta_i \times POST \times TREAT \times \bar{Q}_i + \beta_1 \times C + \alpha + \theta + \varepsilon \qquad (3)$$

[12] During the period of time included in our study, households enrolled in Rwanda's community-based health insurance program were affiliated with a specific health center (Lu, Chin et al 2012) – thus, in the absence of facility-level quality date, the average quality of health care in a given locality is likely to be a reasonable proxy for the quality of care offered at the nearest facilities.

[13] Modern contraceptive methods are defined by the DHS to include: the oral contraceptive pill, intrauterine device, hormonal injections, diaphragm, condoms, female condoms, female sterilization, male sterilization, contraceptive implants, contraceptive foam/jelly, and lactational amenorrhea. (NISR 2005)

In this specification, $\bar{Q}_i$ is an indicator variable equal to 1 if the EA in which the respondent lives belongs to the *i*-th tier and 0 otherwise. All other variables are the same as in the main specification in equation 2. The impact of P4P on *y* in tier *i* is therefore captured by the coefficient δ_i, and $\bar{Q}_i$ is subsumed in α, the vector of EA-level fixed effects.

The estimated coefficients δ_i for each quality tier require a different interpretation than the single coefficient δ for the average effects. Notably, the P4P roll-out was designed to isolate the substitution effect of P4P by providing control facilities with the average quarterly payment received by the treatment facilities. While all control facilities received the same average payment, the actual payments to individual treatment facilities varied with their performance. As a consequence, a comparison of treatment and control facilities with similar baseline quality may no longer isolate the incentive effect. For example, if treatment facilities in the lower quality tier received lower payments than facilities in higher tiers, then they will also have received lower payments than the corresponding control facilities. The coefficients δ_i for each quality tier should therefore be interpreted as the combined income and incentive effects.

5.3 Dependent Variables

We analyze the impact of P4P on services related to family planning, maternal and child health care, and health outcomes and behaviors. We examine both directly rewarded services, and services that were not individually rewarded but were indirectly incentivized through the quality multiplier. Figure 1 summarizes these outcome variables. The list of directly rewarded services that we focus on is a subset of the 24 measures rewarded in the P4P program that are also reported in the DHS. Most of our outcome measures map very closely to the services rewarded under P4P. As a proxy measure for curative care visits, we use visits to a health care

facility for treatment of an infectious illness in a child under the age of 5. We use reported use of modern contraceptive methods as a proxy measure for family planning visits.[14] In addition to overall contraceptive coverage, we also consider the subset of the women who report that they want to limit or space births and may therefore have higher demand for contraception. As a proxy for contraceptive resupply, we employ the prevalence of modern contraceptive methods that require regular resupplies. Appendix Table A.2 provides further details on the variable construction.

Services examined that were not directly rewarded include appropriate vitamin A supplementation for children (i.e. receipt of a vitamin A supplement in the past 6 months by children age 1-5) and seven binary prenatal care quality indicators that are available in the data: whether prenatal care was received from a trained provider (i.e. a licensed medical professional such as a doctor, nurse or midwife); and whether during pregnancy the respondent was weighed, had a blood pressure measurement, had a urinalysis, had blood drawn, received an iron supplement, or was explained the possible complications of pregnancy.

Health outcomes and behaviors include favorable and adverse events. We include whether the child was breastfed for 6 months or more as a favorable event. Adverse outcomes include the prevalence of anemia in children and pregnant women; the prevalence of vision

[14] A number of studies have demonstrated that family planning visits and counseling are associated with the increased use of modern contraceptive methods, thereby supporting our use of the reported prevalence of modern contraception use as a proxy measure for family planning visits (Ahmed and Mosley 2002, Barber 2007, Borges, OlaOlorun et al. 2015, Yadav and Dhillon 2015).

difficulties during pregnancy (any vision difficulties and adjusted night blindness)[15] ; and the incidence of children's infectious illness in the preceding two weeks.[16]

Following Kling, Liebman, and Katz (2007) we also calculate summary indices, scaled 0-100, to address concerns about multiple comparisons. Since the denominators differ across outcomes, we group outcomes that are substantively related into four indices. Each index is the simple average of the binary components. Specifically, the index for "directly rewarded prenatal care" comprises correct tetanus vaccination, malaria prophylaxis in pregnancy and four or more prenatal visits. For "other prenatal care" we group the seven aforementioned binary indicators of services that were not directly rewarded under P4P. The index for "directly rewarded contraception services" includes the use of modern contraception and the use of a method requiring resupply. The vision index contains the measures of any and night vision problems.

[15] Vision changes occur frequently in pregnant women, with one cohort study finding a prevalence of 25% (Pizzarello 2003). In our own study sample, 12-14% of women report vision difficulties in their last pregnancy, with 2-4% reporting night vision difficulties specifically (see Table 2). Vision changes during pregnancy at any time of the day or night may result from gestational diabetes or pre-eclampsia; several studies have used vision changes as a marker of pre-eclampsia, though often in combination with other symptoms of pre-eclampsia that are not available in the DHS (e.g. swelling, hypertension) (Lakshmi et al. 2013, Agrawal & Walia 2014, Liambila & Kuria 2014). However because vision changes in pregnancy may also be normal, this symptom is fairly non-specific (Roos et al. 2012). We therefore also measure adjusted night blindness, which is a more specific and well-established symptom of severe vitamin A deficiency during pregnancy (WHO, 1998) – in fact, the WHO uses the prevalence of night blindness among pregnant women as an indicator of community levels of vitamin A deficiency, and uses this indicator to guide vitamin A supplementation policies (WHO 2011). Worldwide more than 6 million pregnant women develop night blindness annually, with the estimated prevalence ranging from 1% to 16% of pregnant women in low-income countries (Jayasekera et al. 1991, West 2002, Christian 2002). Following the DHS, we report the adjusted night blindness as the prevalence of women reporting vision changes only at night – women who also report vision changes during the day are excluded because these are not consistent with vitamin A deficiency. (NISR 2006)

[16] The prevalence of anemia is measured by blood tests of study children; in contrast, the presence of infectious illness in the past 2 weeks is reported by the child's mother.

As Table 2 shows, observable characteristics are balanced across treatment and control districts at baseline. We include the following key covariates as controls in regressions: the total number of births per woman, the wealth quintile of households, and household insurance status. Depending on the outcome we also include other covariates, e.g., for analyses of child outcome variables we also control for the child's age and sex. The covariate sets are described further in Appendix A.4.

6. Results

6.1 Average Effects

Table 3 shows the results for the summary indices, while Figure 1 summarizes the average effects of the P4P program, from top to bottom, on the disaggregated directly rewarded services, services that were not directly rewarded, and health outcomes and behaviors. These effects are also reported in the top (Panel A) of Tables 4, 5 and 6, respectively.[17]

Beginning with the directly rewarded services, the program significantly increased the share of women delivering in facilities by 9.8 percentage points against a pre-intervention baseline of 30% (Table 4, Panel A). We do not find any significant impact of P4P on the index of rewarded prenatal care services (Table 3), nor on the individual components of the index (i.e. the share of pregnant women completing 4 prenatal care visits, receiving appropriate tetanus vaccination or malaria prophylaxis) (Table 4, Panel A), and we do not find any significant impact on childhood immunizations or the share of children with a recent infectious illness who were treated at a public facility. P4P also had no significant impact on the index of contraceptive

[17] In Tables 3-6 we only show the main coefficients of interest. The full set of results, including coefficients on all included covariates, are available in Appendix B, Tables B.1-B.6.

use or the overall rate of contraceptive use. Among the subset of women with the highest need for contraception (i.e. those without fertility problems who expressed a desire to space or limit births), however, P4P increased the prevalence of modern contraceptive use by 3.9 percentage points (baseline 1.6%). Delivery and contraception services have the highest unit rewards (US $4.59 and $1.83) of the services in our data. While these findings could indicate a dose-response relationship, the rewards to providers net of production costs remain unclear. Moreover, we find no significant impact on childhood immunizations or the overall rate of contraceptive use, two services that are also rewarded at US $1.83.

Our results are broadly consistent with previous research: Basinga et al. (2011) report on a total of eight rewarded measures, four of which are comparable to those in our analysis: institutional delivery, tetanus vaccination, attending at least 4 prenatal care visits, and having a child fully immunized by age 1. Appendix A.5 and Table A.1 display our results next to those of Basinga et al. (2011) to facilitate comparison. The results coincide for three of the comparable measures. For example, they find an 8.1 percentage point increase in the rate of institutional deliveries (compared to our point estimate of 9.9) using an unrelated dataset, and, like us, they find no statistically significant effects for the completion of four antenatal visits and childhood immunizations. This replication result also provides further reassurance that our (few) statistically significant findings are unlikely a result of the multiple comparisons.

The impact of P4P on services that are not directly rewarded is presented in column 3 of Table 3, and Table 5. There was a statistically significant improvement in the index of unrewarded prenatal care measures by 3.7 points. Looking at the individual components of the index, we find that the P4P program had statistically significant, positive impacts on urinalysis

(5.1 percentage points) and iron supplementation (9.3 percentage points) during prenatal care. We find no statistically significant effect on the other aspects of prenatal care quality or child vitamin A supplementation.

As Tables 3 & 6 show, P4P had no significant impact on any of the health outcomes or behaviors studied: breastfeeding for at least 6 months; the vision index or its individual components (i.e. night vision or any vision difficulties during pregnancy); anemia in currently pregnant women or in children; and infectious illnesses in the past 2 weeks in children under the age of 5.

An incidental finding in these analyses concerns household insurance status: having health insurance is associated with large and significant improvements in several rewarded and unrewarded services. This result should not be interpreted as evidence of a causal relationship between insurance and health service utilization or health outcomes. Aside from insurance, the coefficients on the *POST* indicator show substantial secular improvements between 2006 and 2008 for several measures. Some of these improvements could be attributable to the near-quadrupling of total annual health expenditures between 2004 and 2009 (US $130M to $531M in 2012 dollars) and an influx of external funds and government programs, including for family planning (World Health Organization 2012, Basinga et al. 2011). The improvement in the control areas could also be due to the increase in the input-based budgets of the control facilities.

As a robustness check, we also estimated 'raw' difference-in-difference models that do not include the vector of covariates. These results (included in Appendix C) are very close to those from regressions reported in the paper, which control for additional covariates. Most coefficients have similar size and statistical significance. Only the outcome "modern

contraception" is statistically insignificant in the raw specifications; however the coefficient and standard error have a similar size and the coefficient was marginally significant in our preferred specification.

6.2 Heterogeneity by Pre-Intervention Quality

The lower panels (B) in Tables 3-6 present the results of the P4P program interacted with the low, medium, and high quality tiers, following the DD specification in equation 3. Recall that the comparison by baseline quality may no longer isolate only the incentive effect of P4P, as all control facilities receive the average payment made to all treatment facilities but specific treatment facilities may receive varying payments.

For several directly rewarded health services there are significant variations in program impacts by baseline quality: improvements in institutional deliveries are concentrated in the high tier, while significant improvements in the contraception measures (and the contraception index) and prenatal malaria prophylaxis are concentrated in the medium tier. The only service with significant improvements in the low tier areas is the use of modern contraception by higher-need women. Curiously, the program also led to a *reduction* in the number of prenatal visits in the low tier (-7.4 percentage points). While there are no unambiguous trends across the tiers, our findings suggest that the top baseline quality tier areas achieved improvements in the most generously rewarded service studied (i.e. institutional deliveries), but the smallest number of rewarded services overall. Overall, medium-tier facilities appear to have achieved improvements across a broader array of directly rewarded services.

Among services that are not directly rewarded, we see a similar pattern. For one of the unrewarded prenatal care services studied - iron supplementation - improvements are concentrated in the high tier (12.6 percentage points). Conversely, in the high quality tier there is a weakly significant reduction (-5.8 percentage points) in the percent of pregnant women being informed of potential complications of pregnancy. The medium tier areas show improvements in three unrewarded quality indicators: iron supplementation (14.2 percentage points), blood pressure measurement (12.0 percentage points), and urinalysis (7.0 percentage points). Improvements in the index of unrewarded services are also concentrated in the medium tier. There are no significant improvements or deteriorations in unrewarded services in the low quality tier. Improvements in services that are not directly rewarded are therefore also concentrated in the medium tier facilities.

The impact of P4P on health outcomes and behaviors does not vary significantly by an area's baseline quality, with the exception of the reported incidence of infectious illnesses in children (diarrhea, cough or fever), which increased in the low- and medium-quality areas (10.6 and 17.0 percentage points, respectively).

7. Discussion

Performance-based contracting can have complex effects that are often overlooked during the planning and analysis of P4P interventions. When programs reward multiple targets and services are jointly produced, the likely impacts on rewarded and unrewarded measures are ambiguous *ex ante* and depend on the relative net rewards, the degree of multitasking, commonalities in production and other features of the production function.

In this study, we use the quasi-experimental roll-out of Rwanda's national P4P program to empirically assess the program's effects on directly rewarded services, services that are not directly rewarded but are indirectly incentivized through the quality multiplier, and health outcomes. We find improvements in some (but not all) of the services studied, and no detectable effect on health outcomes and behaviors that we observe in the DHS data. We also observe no deterioration in services or outcomes under P4P.

Among directly rewarded services, we find that P4P increased institutional deliveries and contraceptive coverage for women with the highest need – which are the two most highly rewarded services observable in the DHS data – but had no statistically significant effects on services with lower rewards. Consistent with our theoretical predictions, these results indicate that the size of the reward does influence provider behavior – indeed, during focus groups Rwandan providers admitted that since institutional deliveries were so generously rewarded under P4P, they paid community health workers to recruit pregnant women to deliver in health care facilities (Basinga 2009).

In an exception to this trend, however, we see no significant improvements in two services that are also relatively generously rewarded: childhood immunizations and modern contraception. This may reflect the importance of marginal costs of production in determining the net reward that providers face which, in turn, could depend on the provider's costs as well as demand-side factors. For example, women who report wanting to limit or space births may be more motivated to seek out contraception, thereby decreasing the marginal cost of achieving contraceptive coverage in this group compared to the general female population. Childhood immunizations require multiple visits, so are also costly to improve. The findings indicate that

while the size of P4P rewards matters, other factors such as production costs play an important role in influencing providers. Our findings on institutional deliveries and childhood immunizations are similar to those of Basinga et al. (2011).

Turning to the services that were *not* directly rewarded, we find no evidence that their output decreased as a result of P4P – on the contrary, P4P had a positive impact on some indirectly rewarded aspects of prenatal care. There are several possible explanations for these findings. First, while these services are not individually rewarded, they are related to components of the quality multiplier *M* and therefore are indirectly incentivized through the P4P scheme. Facilities seeking to improve their quality multiplier score to earn higher marginal revenue from directly rewarded services therefore might have invested in improving health services that are not directly rewarded. By creating incentives to invest in the indirectly rewarded services, the quality multiplier might have mitigated multitasking. Second, the increase in these services could reflect positive spillovers due to commonalities in production with directly rewarded measures. In particular, family planning services are likely to be jointly produced with highly rewarded HIV/AIDS services given the common emphasis on safe sexual practices: Rwanda's P4P scheme pays facilities US $2.68 for every HIV-positive woman using a modern contraceptive method, and Basinga (2009) notes that HIV services and family planning services were integrated at many health care facilities. The observed increase in indirectly rewarded prenatal care measures such as urinalysis and iron supplementation is another example of positive spillovers that could arise from joint production with directly rewarded prenatal care measures.

We do not find evidence that health outcomes or behaviors associated with rewarded services improved in response to P4P. One possible explanation is that activities that responded

positively to incentives contribute only weakly to health improvements, so that even with significant increases in the volume of these activities, improvements in outcomes do not follow automatically. Another possibility is that small sample sizes for the health outcomes limit our statistical power to detect effects – although at least for breastfeeding and night vision problems, our estimates are sufficiently precise to rule out substantively meaningful impacts. In contrast, Gertler and Vermeersch (2013) find positive impacts on a different set of health outcomes – child height-for-age and weight-for-age – suggesting that these outcomes may be more sensitive to improvements in the quality of prenatal and early childhood care than those examined in this paper.

Finally, our analysis reveals variations in P4P's impacts by baseline levels of performance. Consistent with our prediction that higher-performing facilities face the largest incentive to increase highly rewarded services, we find that facilities in the high quality tier achieved the largest increases in institutional deliveries, one of the most generously rewarded services. However these facilities fell short in improving other directly rewarded services, despite facing the highest marginal revenue for these services. There are several possible explanations for this finding. First, there could be a higher marginal cost of improvement in these high-performing areas if facilities must attract new patients rather than changing clinical activities for existing patients. Higher quality facilities may also be closer to their production possibility frontiers rendering further increases in service provision very costly, or requiring lumpy, significant investments such as increased staffing or infrastructure. A final explanation relates to the quality multiplier and multitasking. Because P4P bonus payments are scaled by the multiplier, higher-performing facilities have the strongest incentive to improve the most generously rewarded services, which under Rwanda's P4P scheme are HIV/AIDS services, so

higher-performing areas may have focused on these measures (which we do not observe in the DHS data) at the expense of the less generously rewarded measures observed in our data.

Consistent with our prediction that lower-performing facilities face the highest incentive to increase the value of the quality multiplier, we find that facilities in the medium quality tier achieved the largest increases in both directly rewarded prenatal care services and individually unrewarded prenatal care services that are a component of the quality multiplier. We do not, however, find similar improvements in the low-quality tier, nor do we see improvements in most directly rewarded services. It is possible that the lowest-performing facilities did respond to P4P incentives to strengthen aspects of the quality multiplier that are not captured in our outcome measures (i.e. pharmacy management, general administration, TB services, etc.). Alternatively, these facilities may have lacked the infrastructure, capital or human resources required to make even basic improvements in health service delivery. Other possible explanations to consider are that services that were not directly rewarded may have been less salient to facilities despite their inclusion in the quality multiplier, or that the lowest-quality facilities may have been less forward-looking and hence did not make short-term investments in the quality multiplier that might have increased their longer-term revenue. These factors may change over time as facilities learn about the program's parameters and the returns to investment. Finally, recall that in the heterogeneity analysis, we are no longer able to isolate the incentive effect independently of the income effect so that the estimated program effects reflect a combination of both, whereas facilities in the control group received a fixed payment. Thus, relative to low-tier facilities in the control group, low-tier facilities in the treatment group received lower average quarterly payments and the negative income effect may offset any positive incentive effects.

While the Rwandan P4P program provides a unique opportunity to investigate empirically the complex effects of P4P programs in a rigorous way, there are several challenges to evaluating such real-life policies. First, it is possible that concurrent public health initiatives outside of P4P might have influenced the program's effects – indeed, the study period coincided with both a national immunization campaign and increased HIV/AIDS funding and programming, which is closely related to family planning. The positive coefficients on the *POST* indicator suggest the importance of such longitudinal, secular changes. To our knowledge, these concurrent initiatives did not coincide with the roll-out of P4P but they might have indirectly influenced responses to P4P if providers chose not to allocate extra effort to immunization and HIV/AIDS campaigns above and beyond these programs, or conversely if these initiatives sensitized patients and providers to the importance of these public health challenges. Second, the Rwandan program was implemented through block-randomization, which may have resulted in remaining unobserved differences across the treatment and control areas. Consistent with research using alternative datasets, we do not find statistically significant differences in observable characteristics between the treatment and control districts in the DHS data. A third limitation is that because we lack facility-level data, in the heterogeneity analysis we must use village/locality-level health care coverage as a proxy for the baseline quality of the nearest facility/facilities, and there may be divergence between these measures. Given that at the time of the study households enrolled in Rwanda's community-based health insurance program were assigned to a specific health center within the same local area (Lu et al. 2012, Kayonga 2007), this should increase the likelihood that individuals in a given locality access care at the same facility and thus achieve greater concordance between facility-level quality and local area-level health service coverage and quality. Finally, the reassignment of several districts from control to

treatment status that was necessitated by the redistricting process may have biased our results. Because information on the specific districts affected has not been made publicly available, it is not possible to perform robustness checks that omit these districts, and it is difficult to predict exactly how the redistricting might influence our results. As also argued by Gertler and Vermeersch (2013), however, we note that the decision to reassign districts in Rwanda was made prior to the evaluation and was not a result of choices made by respondents or participants (facilities) in this study.

Rwanda's P4P initiative, as the first such program implemented and evaluated at scale in a developing country, has inspired the introduction of a large number of similar programs in Africa and elsewhere. There is a great need for program impact data to inform the design and scale-up of these initiatives, and Rwanda offers several lessons. From a policy perspective, our findings suggest that Rwanda's national P4P program has contributed to some improvements in health care quality, particularly in the areas of institutional deliveries, contraceptive coverage and prenatal care. Our findings also point to several areas in which Rwanda's P4P program might be strengthened, with lessons for provider payment schemes not only in developing countries but in higher-resource settings as well. First, rewarding health care services that are jointly produced with other desirable services may achieve broader improvements. Rewarding only a subset of services that are jointly produced is also more cost-effective than rewarding each of these services individually, because of positive production spillovers. Second, the use of a quality multiplier or scaling factor based on broader measures can potentially mitigate multitasking concerns. Finally, our findings highlight significant variations in the effectiveness of P4P in areas with different levels of baseline quality, with important implications for the design and scale-up of these initiatives. In Rwanda's case, either a modified payment formula or other

complementary approaches are needed to encourage higher-performing facilities to increase the output of rewarded services. Similarly, lower-performing facilities may require additional investment or support beyond P4P to achieve meaningful quality improvements – otherwise P4P risks widening disparities in health care quality between these facilities and their higher-performing peers. Any future efforts to offer incremental performance rewards based on inputs requires an understanding of the underlying production function and cost structures, which remains especially challenging in health care and even more so in a developing country. There is also a need for more research and evidence on alternative methods of rewarding performance, such as outcome-contingent contracts.

REFERENCES

Agrawal, S, and GK. Walia. 2014. "Prevalence and risk factors for symptoms suggestive of pre-eclampsia in Indian women." *Journal of Women's Health:Issues & Care* 3 (6).

Ahmed, Saifuddin, and W. Henry Mosley. 2002. "Simultaneity in the use of maternal-child health care and contraceptives: evidence from developing countries." *Demography* 39 (1):75-93.

Barber, Sarah L. 2007. "Family Planning Advice and Postpartum Contraceptive Use Among Low-Income Women in Mexico." *International Family Planning Perspectives* 33 (1):6-12.

Basinga, Paulin. 2009. "Impact of performance-based financing on the quantity and quality of maternal health services in Rwanda. (Doctoral dissertation) Retrieved from Proquest, UMI number 3353893." PhD, International Health, Tulane University.

Basinga, Paulin, PJ Gertler, A Binagwaho, AL Soucat, J Sturdy, and CM 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* 377 (9775):1421–1428.

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

Borges, Ana Luiza Vilela, Funmilola OlaOlorun, Elizabeth Fujimori, Luiza Akiko Komura Hoga, and Amy Ong Tsui. 2015. "Contraceptive use following spontaneous and induced abortion and its association with family planning services in primary health care: results from a Brazilian longitudinal study." *Reproductive Health* 12:94.

Borghi, Josephine, Richard Little, Peter Binyaruka, Edith Patouillard, and August Kuwawenaruwa. 2015. "In Tanzania, The Many Costs Of Pay-For-Performance Leave Open To Debate Whether The Strategy Is Cost-Effective." *Health affairs* 34 (3):406-414.

Cameron, AC, JB Gelbach, and DL Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90:414–427.

Campbell, SM, D Reeves, E Kontopantelis, E Middleton, B Sibbald, and M. Roland. 2007. "Quality of primary care in England with the introduction of pay for performance." *New England Journal of Medicine* 57 (2):181-190.

Campbell, SM, D Reeves, E Kontopantelis, B Sibbald, and M. Roland. 2009. "Effects of pay for performance on the quality of primary care in England." *The New England Journal of Medicine* 361:368–78.

Christian, Parul. 2002. "Recommendations for Indicators: Night Blindness during Pregnancy—A Simple Tool to Assess Vitamin A Deficiency in a Population." *The Journal of Nutrition* 132 (9):2884S-2888S.

Christianson, JB, S Leatherman, and K Sutherland. 2008. "Lessons from evaluations of purchaser pay-for-performance programs: a review of the evidence." *Medical Care Research and Review* 65:5S–35S.

Cohen, Jessica, and Pascaline Dupas. 2010. "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." *Quarterly Journal of Economics* 125:1–45.

de Walque, Damien, Paul J. Gertler, Sergio Bautista-Arredondo, Ada Kwan, Christel Vermeersch, Jean de Dieu Bizimana, Agnès Binagwaho, and Jeanine Condo. 2015. "Using provider performance incentives to increase HIV testing and counseling services in Rwanda." *Journal of health economics* 40:1-9.

Doran, T, C Fullwood, H Gravelle, D Reeves, E Kontopantelis, U Hiroeh, and Roland M. 2006. "Pay-for-performance programs in family practices in the United Kingdom. ." *New England Journal of Medicine* 355 (4):375-384.

Fritsche, György Bèla, Robert Soeters, and Bruno Meessen. 2014. *Performance-Based Financing Toolkit.* . Washington, DC: World Bank.

Gertler, Paul, and Christel Vermeersch. 2013. Using Performance Incentives to Improve Medical Care Productivity and Health Outcomes. In *NBER Working Paper 19046*. Cambridge, MA: National Bureau of Economic Research.

Glazer, Jacob, Thomas McGuire, and Sharon-Lise T. Normand. 2008. "Mitigating the problem of unmeasured outcomes in quality reports. ." *The B.E. Journal of Economic Analysis & Policy* 8 (2):1935-1682.

Glickman, SW, FS Ou, ER DeLong, MT Roe, BL Lytle, J Mulgund, JS Rumsfeld, Gibler W Brian, Ohman E Magnus, Kevin A Schulman, and Peterson ED. 2007. "Pay for performance, quality of care, and outcomes in acute myocardial infarction." *JAMA: Journal of the American Medical Association* 297 (21):2373-2380.

Hoff, Martin. 2010. Improving the incentive mechanisms in Rwanda's health care system. Cambridge, MA: Harvard University.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7 (special issue):24-52.

Jayasekera, J. P., T. M. Atukorala, and H. R. Seneviratne. 1991. "Vitamin A status of pregnant women in five districts of Sri Lanka." *Asia Oceania J Obstet Gynaecol* 17 (3):217-24.

Kayonga, Caroline. 2007. Towards Universal Health Coverage in Rwanda: Summary Notes from Briefing at Brookings Global Economy and Development. Brookings Institution.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75 (1):83-119.

Lakshmi, P. V. M., Navkiran Kaur Virdi, Atul Sharma, Jaya Prasad Tripathy, Kirk R. Smith, Michael N. Bates, and Rajesh Kumar. 2013. "Household air pollution and stillbirths in India: Analysis of the DLHS-II National Survey." *Environmental Research* 121:17-22.

Li, J, J Hurley, DeCicca P, and Buckley G. 2011. Physician response to pay-for-performance: evidence from a natural experiment. . In *NBER Working Paper 16909*. Cambridge MA.

Liambila, Wilson N., and Shiphrah N. Kuria. 2014. "Birth attendance and magnitude of obstetric complications in Western Kenya: a retrospective case–control study." *BMC Pregnancy and Childbirth* 14 (1):1-15.

Lu, Chunling, Brian Chin, Jiwon Lee Lewandowski, Paulin Basinga, Lisa R. Hirschhorn, Kenneth Hill, Megan Murray, and Agnes Binagwaho. 2012. "Towards Universal Health Coverage: An Evaluation of Rwanda Mutuelles in Its First Eight Years." *PLoS ONE* 7 (6):e39282.

McGuire, Thomas G., and Mark V. Pauly. 1991. "Physician response to fee changes with multiple payers." *Journal of health economics* 10 (4):385-410.

Miller, G., and K. S. Babiarz. 2014. "Pay-for-Performance Incentives in Low- and Middle-Income Country Health Programs A2 - Culyer, Anthony J." In *Encyclopedia of Health Economics*, 457-466. San Diego: Elsevier.

Ministry of Health of Rwanda. 2010. Rwanda Community-Based Health Insurance Policy. Kigali, Rwanda.

Ministry of Local Government, Community Development and Social Affairs (MINALOC). 2004. Rwanda Five-Year Decentralization Implementation Program. Kigali: Ministry of Local Government, Community Development and Social Affairs (MINALOC), Republic

of Rwanda Original edition, Republic of Rwanda. Ministry of Local Government, Community Development and Social Affairs.

Mullen, Kathleen J., Richard G. Frank, and Meredith B. Rosenthal. 2010. "Can you get what you pay for? Pay-for-performance and the quality of healthcare providers." *The Rand Journal of Economics* 41 (1):64-91.

National Institute of Statistics of Rwanda (NISR), and Macro International Inc. 2005. Rwanda Demographic and Health Survey edited by Inc. Macro International. Calverton, United States: .

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

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* 30 (4):773-781. doi: 10.1377/hlthaff.2009.0782.

Petersen, Laura A., LeChauncy D. Woodard, Tracy Urech, Christina Daw, and Supicha Sookanan. 2006. "Does Pay-for-Performance Improve the Quality of Health Care?" *Annals of Internal Medicine* 145 (4):265-272. doi: 10.7326/0003-4819-145-4-200608150-00006.

Pizzarello, L D. . 2003. "Refractive changes in pregnancy." *Graefe's Archive for Clinical and Experimental Opthalmology* 241 (6):484-488.

Prendergast, Canice. 1999. "The provision of incentives in firms." *Journal of economic literature* 37 (1):7-63.

Rosenthal, Meredith B, and A Dudley. 2007. "Pay-for-Performance: Will the Latest Payment Trend Improve Care?" *JAMA: Journal of the American Medical Association* 297.

Rosenthal, Meredith B, and Richard G. Frank. 2006. "What is the empirical basis for paying for quality in health care?" *Medical Care Research and Review* 63:135–157.

Rosenthal, Meredith B, Richard G. Frank, Zheng Li, and A. M. Epstein. 2005. "Early experience with pay-for-performance: from concept to practice." *JAMA: Journal of the American Medical Association* 294:1788–1793.

Sherry, Tisamarie B. 2016. "A Note on the Comparative Statics of Pay-for-Performance in Health Care." *Health Economics* 25 (5).

Soeters, Robert, Peter Bob Peerenboom, Pacifique Mushagalusa, and Célestin Kimanuka. 2011. "Performance-Based Financing Experiment Improved Health Care In The Democratic Republic Of Congo." *Health Affairs* 30 (8):1518-1527. doi: 10.1377/hlthaff.2009.0019.

Steel, N, S Maisey, A Clark, R Fleetcroft, and A. Howe. 2007. "Quality of clinical primary care and targeted incentive payments: an observational study." *British Journal of General Practice* 57:449–454.

Sutton, M, R Elder, B Guthrie, and G. Watt. 2010. "Record rewards: The effects of targeted quality incentives on the recording of risk factors by primary care providers." *Health Economics* 19:1-13.

Van de Poel, Ellen, Gabriela Flores, Por Ir, and Owen O'Donnell. 2015. "Impact of Performance-Based Financing in a Low-Resource Setting: A Decade of Experience in Cambodia." *Health Economics*:n/a-n/a. doi: 10.1002/hec.3219.

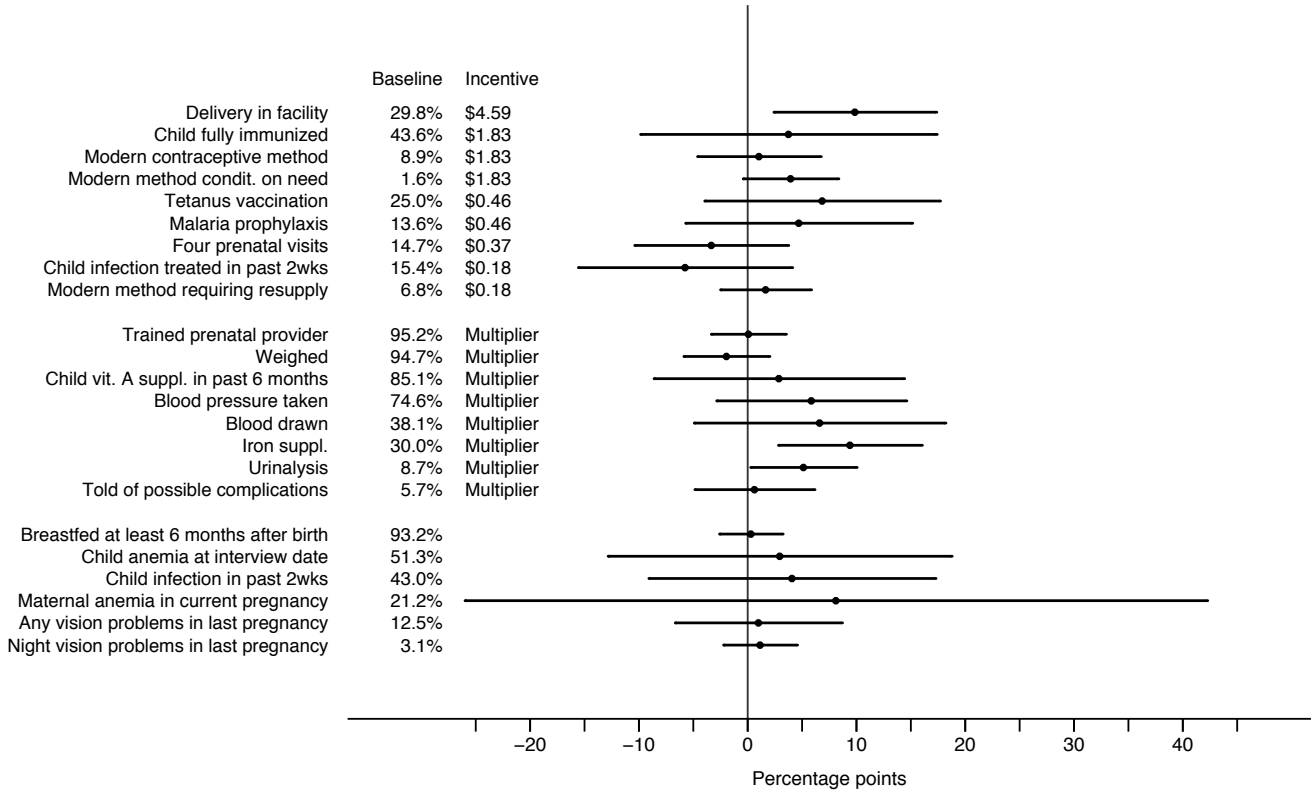
West, K. P., Jr. 2002. "Extent of vitamin A deficiency among preschool children and women of reproductive age." *J Nutr* 132 (9 Suppl):2857s-2866s.

World Health Organization. 2011. Guideline: Vitamin A supplementation in pregnant women. . Geneva: World Health Organization.

World Health Organization. 2012. WHO National Health Accounts Database. Geneva: World Health Organization.

Yadav, Diwakar, and Preeti Dhillon. 2015. "Assessing the Impact of Family Planning Advice on Unmet Need and Contraceptive Use among Currently Married Women in Uttar Pradesh, India." *PLoS ONE* 10 (3):e0118584.

Figure 1: Estimated treatment effects for rewarded services, services in the multiplier, and health outcomes and behaviors



Coefficients for post*treat from models in Tables 3-5, with 95% CIs based on a t(G-2) distribution. Positive estimates represent improvement for services in the top and middle panel. Vision problems can be a symptom of pre-eclampsia/gestational diabetes. Night blindness is a specific symptom of vitamin A deficiency.

Table 1: Rewarded indicators and components of quality multiplier M

Rewarded output and quality indicators	Payment/unit (US $)	Available in DHS
Primary Care		
Emergency referrals during curative treatment	1.83	
Curative care visits	0.18	Y
Family Planning		
First-time family planning visits	1.83	Y†
1-month contraceptive resupply	0.18	Y†
Maternal Health		
Deliveries in the facility	4.59	Y
Emergency transfers to hospital for obstetric care during delivery	4.59	
At-risk pregnancies referred to hospital for delivery during prenatal care	1.83	
Women who received 2nd dose of malaria prophylaxis during prenatal care	0.46	Y
Women who received appropriate tetanus vaccination during prenatal care	0.46	Y
Women who completed 4 prenatal care visits	0.37	Y
First prenatal care visits	0.09	
Child Health		
Malnourished children referred for treatment during preventive care visit	1.83	
Children who completed vaccination on time	1.83	Y
Child (0-59 months) preventive care visits	0.18	
HIV/AIDS		
PMTCT: exposed children tested	8.93	
New pediatric clients put on ARVs	6.70	
Prevention of mother-to-child-transmission (PMTCT): partner tested	4.58	
PMTCT: women under treatment with ARVs during labor	4.58	
New adult clients put on ARVs	4.58	
HIV+ clients tested for CD4 count	4.58	
HIV+ women who use modern method of family planning	2.68	
HIV+ clients tested for TB	2.68	
Voluntary counseling and testing	0.89	
HIV+ clients treated with cotrimoxazole each month	0.44	

Components of the quality multiplier M	Weight in multiplier	Share of structural factors (rather than process factors)
Curative Care	0.170	0.23
Delivery	0.130	0.40
Prenatal care	0.126	0.12
Family Planning	0.114	0.22
HIV Services	0.090	1.00
Immunization	0.070	0.40
Pharmacy Management	0.060	1.00
General Administration	0.052	1.00
Financial Management	0.050	1.00
Preventive Care	0.052	0.15
Laboratory Services	0.030	1.00
Cleanliness	0.028	1.00
TB Services	0.028	0.28

† First-time family planning visit includes prescription of modern contraceptive method, medical history and physical exam; DHS records "current use of modern method" which is used as proxy here. Contraceptive resupply measured as current use of a modern method requiring regular resupplies.

Table 2: Summary statistics for pre-intervention period

Outcomes (binary)	N	Mean Control	Mean Treat	p-value
Rewarded services				
Delivery in facility	4,914	0.30	0.29	0.64
Tetanus vaccination	2,723	0.27	0.24	0.22
Malaria prophylaxis	2,749	0.13	0.13	0.95
Four prenatal visits	2,748	0.13	0.15	0.42
Child fully immunized	3,988	0.47	0.41	0.40
Child infection treated in past 2wks	1,108	0.13	0.17	0.19
Modern contraceptive method	1,814	0.10	0.08	0.33
Modern method condit. on need	684	0.01	0.02	0.70
Modern method requiring resupply	1,814	0.08	0.06	0.12
Measures in quality multiplier				
Trained prenatal provider	2,755	0.96	0.94	0.19
Weighed (prenatal care)	2,628	0.94	0.95	0.47
Blood pressure taken (prenatal care)	2,627	0.76	0.73	0.41
Blood drawn (prenatal care)	2,626	0.40	0.36	0.49
Iron suppl. (prenatal care)	2,728	0.34	0.27	0.22
Urinalysis (prenatal care)	2,613	0.08	0.09	0.49
Told of possible complic. (prenatal care)	2,624	0.06	0.06	0.98
Child vit. A suppl. in past 6 months	1,965	0.85	0.85	1.00
Health outcomes and behaviors				
Breastfed at least 6 months after birth	4,496	0.90	0.89	0.79
Any vision problems in last pregnancy	2,739	0.14	0.12	0.56
Night vision problems in last pregnancy	2,739	0.04	0.02	0.13
Maternal anemia in current pregnancy	198	0.20	0.22	0.83
Child anemia at interview date	1,268	0.53	0.50	0.64
Child infection in past 2wks	2,577	0.43	0.43	0.95
Covariates				
Wealth Quintile	4,914	1.91	1.73	0.30
Insured	4,914	0.53	0.57	0.56
Mother's education	4,914	3.57	3.46	0.65
Mother's age	4,914	31.23	31.20	0.91
Married/cohabitating	4,914	0.88	0.87	0.52
Hhold size	4,914	5.69	5.62	0.52
N children $\leq$ 5 years	4,914	1.87	1.85	0.70
N births	4,914	4.32	4.47	0.30
EA baseline quality Q: 1(low) to 3(high)	153	2.00	1.99	0.94
Indices				
Rewarded prenatal index	2,704	0.18	0.17	0.63
Contraception index	1,814	0.09	0.07	0.19
Multiplier prenatal index	2,573	0.51	0.49	0.46
Vision index	2,739	0.09	0.07	0.36

Covariate values based on estimation sample for facility deliveries. See Appendix for details on sample definitions and variable constructions. Vision problems are symptomatic of pre-eclampsia/gestational diabetes. Night blindness is a specific symptom of vitamin A deficiency. p-values clustered on district level and based on t-distribution with G-2 d.f.

Table 3: Effect of P4P on indices (range 0-1)

	Rewarded			
	(1)	(2)	(3)	(4)
	Rewarded prenatal index	Contra-ception index	Multiplier prenatal index	Vision index
Pre-period mean (%)	17.72	7.86	50.34	7.78
A. Average effect (results for covariates omitted)				
Treat*Post	3.34	1.34	3.79***	1.06
	(3.41)	(2.23)	(1.30)	(2.41)
Post	0.81	15.49***	-0.20	-1.88
	(3.28)	(1.65)	(1.53)	(1.99)
B. Effect by baseline quality Q				
Treat*Post Q-low	-1.39	-0.79	2.13	0.18
	(4.77)	(4.27)	(2.48)	(2.93)
Treat*Post Q-medium	6.30	8.03**	6.15***	2.40
	(5.92)	(2.96)	(1.59)	(2.68)
Treat*Post Q-high	3.55	-2.23	3.17	1.93
	(4.78)	(3.43)	(2.29)	(3.26)
Post Q-low	2.70	17.69***	2.77	1.10
	(4.73)	(2.47)	(1.85)	(1.80)
Post Q-medium	1.89	9.02***	-1.65	-4.72*
	(5.23)	(2.27)	(1.45)	(2.44)
Post Q-high	-1.63	18.68***	-1.98	-2.73
	(3.83)	(2.40)	(1.81)	(2.74)
Wealth Quintile	0.77*	1.68***	0.80**	-0.02
	(0.39)	(0.52)	(0.28)	(0.26)
Insured	1.54	1.58	-0.12	0.89
	(1.02)	(1.16)	(0.46)	(1.10)
Mother's education	0.14	0.95***	0.28***	-0.03
	(0.16)	(0.17)	(0.08)	(0.18)
Age at birth	-0.19**		0.06	0.13
	(0.08)		(0.07)	(0.10)
Mother's age		-0.57***		
		(0.13)		
Additional covars	Yes	Yes	Yes	Yes
Birth year FE	Yes	No	Yes	Yes
Area FE	Yes	Yes	Yes	Yes
N†	3,718	3,709	3,574	3,781
R2 (quartile models)	0.20	0.10	0.18	0.01
p(equal Treat*Post*Q)‡	0.51	0.05	0.36	0.80

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Panel B uses same models as panel A, replacing treat and post*treat with quartile interactions. Indices constructed as unweighted average of binary outcomes. Sample for contraception index restricted to married/cohabitating women. † Details on analytical samples and covariate sets in Data section and Appendix. ‡ F-test for equality of Treat*Post*Q coefficients.

Table 4: Effect of P4P on rewarded services (percentage point changes)

		Prenatal Care			Child health		Modern contraception		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Delivery in facility	Tetanus vaccination	Malaria prophylaxis	Four prenatal visits	Child fully immunized	Child infection treated in past 2wks	Use of modern method	Modern method condit. on need	Method requiring resupply
Incentive in USD	4.59	0.46	0.46	0.37	1.83	0.18	1.83	1.83	0.18
Pre-period mean (%)	29.79	25.03	13.58	14.67	43.63	15.43	8.88	1.61	6.84
A. Average effect (results for covariates omitted)									
Treat*Post	9.85**	6.84	4.69	-3.35	3.75	-5.76	1.03	3.94*	1.64
	(3.53)	(5.12)	(4.93)	(3.34)	(6.44)	(4.65)	(2.68)	(2.06)	(1.96)
Post	-1.62	-0.11	-1.11	4.72	8.81	-1.23	16.27***	1.83	14.72***
	(3.78)	(4.66)	(4.79)	(3.87)	(5.38)	(6.51)	(1.98)	(1.33)	(1.42)
B. Effect by baseline quality Q									
Treat*Post Q-low	6.69	0.42	2.37	-7.42*	-2.12	-7.17	0.87	8.85**	-2.45
	(5.26)	(4.42)	(9.16)	(3.97)	(8.31)	(6.38)	(4.30)	(3.56)	(4.39)
Treat*Post Q-medium	8.10	11.46	11.32*	-5.94	4.30	-0.08	5.90*	5.18*	10.16***
	(4.93)	(9.77)	(5.38)	(6.30)	(7.36)	(6.55)	(3.32)	(2.89)	(2.89)
Treat*Post Q-high	12.04**	9.41	-2.13	0.56	5.73	-10.85	-2.79	-0.97	-1.66
	(5.51)	(6.48)	(7.21)	(5.19)	(11.36)	(8.76)	(3.80)	(3.69)	(3.56)
Post Q-low	-1.31	8.45*	-4.06	4.41	11.11*	2.26	17.60***	-1.32	17.77***
	(4.59)	(4.66)	(8.67)	(4.51)	(6.15)	(7.25)	(2.30)	(2.21)	(2.74)
Post Q-medium	4.26	-3.63	-1.41	11.29*	15.11**	-3.86	10.67***	-0.37	7.36***
	(5.31)	(8.74)	(3.96)	(5.64)	(5.67)	(5.85)	(2.54)	(0.66)	(2.22)
Post Q-high	-6.32	-5.78	2.36	0.36	1.73	-1.84	19.56***	6.49***	17.80***
	(4.78)	(5.07)	(5.39)	(4.48)	(9.97)	(9.96)	(2.32)	(2.13)	(2.66)
Wealth Quintile	2.76***	1.13*	0.32	1.04*	0.74	1.40	1.79***	0.33	1.57***
	(0.44)	(0.65)	(0.73)	(0.51)	(0.52)	(0.85)	(0.56)	(0.49)	(0.49)
Insured	6.43***	1.85	-0.21	2.26	3.74**	10.43***	1.77	1.34	1.39
	(1.29)	(1.69)	(1.78)	(1.36)	(1.41)	(1.46)	(1.30)	(1.23)	(1.08)
Mother's education	1.52***	-0.32	0.77**	0.09	0.42	0.46	1.20***	0.61***	0.69***
	(0.22)	(0.21)	(0.30)	(0.29)	(0.30)	(0.50)	(0.18)	(0.21)	(0.20)
Age at birth	-0.11	-0.68***	-0.12	0.30*	0.31				
	(0.14)	(0.21)	(0.18)	(0.16)	(0.23)				
Mother's age						-0.11	-0.55***	-0.05	-0.58***
						(0.20)	(0.13)	(0.11)	(0.13)
Additional covars	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
Area FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N†	5,657	3,753	3,786	3,791	4,588	2,276	3,709	1,173	3,709
R2 (quartile models)	0.17	0.11	0.27	0.03	0.08	0.04	0.11	0.06	0.09
p(equal Treat*Post*Q)‡	0.72	0.27	0.42	0.46	0.72	0.61	0.15	0.11	0.02

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Panel B uses same models as panel A, replacing treat and post*treat with quartile interactions. Samples in cols 7-9 restricted to married/cohabitating women. Models on child health also control for child's age and gender. Child infection treated at government facility. † Details on analytical samples and covariate sets in Data section and Appendix. ‡ F-test for equality of Treat*Post*Q coefficients.

Table 5: Effect of P4P on services in multiplier (percentage point changes)

	Prenatal Care							
	(1) Trained prenatal provider	(2) Weighed	(3) Blood pressure taken	(4) Blood drawn	(5) Iron suppl.	(6) Urinalysis	(7) Told of possible complica-tions	(8) Child vit. A suppl. in past 6 months
Incentive in USD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pre-period mean (%)	95.20	94.66	74.65	38.06	30.01	8.68	5.71	85.14
A. Average effect (results for covariates omitted)								
Treat*Post	0.06	-1.96	5.85	6.61	9.40***	5.13**	0.62	2.87
	(1.62)	(1.85)	(4.12)	(5.46)	(3.12)	(2.30)	(2.60)	(5.44)
Post	-1.53	4.98**	0.52	-3.84	-1.14	-2.16	0.44	-2.34
	(1.41)	(1.86)	(4.07)	(4.65)	(4.29)	(2.80)	(3.13)	(5.84)
B. Effect by baseline quality Q								
Treat*Post Q-low	-2.46	-3.57	3.97	1.59	1.60	3.63	4.07	-2.12
	(2.04)	(3.85)	(7.06)	(9.64)	(6.64)	(3.39)	(3.87)	(9.11)
Treat*Post Q-medium	0.86	-0.42	12.69**	8.95	14.64*	7.27*	2.89	10.14
	(2.32)	(2.00)	(5.21)	(8.42)	(7.50)	(3.59)	(4.50)	(7.15)
Treat*Post Q-high	1.48	-1.05	-0.12	10.74	12.70*	4.68	-5.81*	0.37
	(1.94)	(2.31)	(4.89)	(6.29)	(6.80)	(4.71)	(2.97)	(7.95)
Post Q-low	-1.04	8.37***	8.06	1.99	5.68	-3.59	0.68	-1.69
	(1.46)	(2.44)	(4.64)	(7.42)	(6.63)	(2.74)	(4.33)	(8.07)
Post Q-medium	-1.17	3.02	-2.01	-6.84	-4.84	-3.77	-0.28	-4.80
	(1.69)	(2.04)	(5.49)	(7.33)	(6.08)	(3.11)	(3.95)	(6.88)
Post Q-high	-2.27	3.08	-4.68	-7.43	-4.89	0.39	0.90	-0.53
	(1.71)	(1.90)	(3.55)	(5.89)	(6.06)	(3.71)	(2.77)	(7.39)
Wealth Quintile	0.42	0.15	1.77**	0.85	1.88***	0.82	-0.02	0.77
	(0.29)	(0.30)	(0.69)	(0.87)	(0.60)	(0.51)	(0.24)	(0.68)
Insured	1.25	0.23	-0.69	-0.07	0.30	-0.16	-0.30	3.38**
	(0.74)	(0.65)	(1.80)	(1.43)	(1.60)	(1.14)	(1.05)	(1.57)
Mother's education	0.22	0.24***	0.94***	-0.19	0.41*	0.44	0.27*	0.10
	(0.18)	(0.08)	(0.25)	(0.28)	(0.22)	(0.25)	(0.14)	(0.26)
Age at birth	-0.19*	0.14	0.18	-0.21	0.04	0.05	0.00	
	(0.10)	(0.09)	(0.15)	(0.16)	(0.18)	(0.12)	(0.10)	
Mother's age								0.01
								(0.15)
Additional covars	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Area FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N†	3,803	3,651	3,650	3,648	3,768	3,626	3,644	3,856
R2 (quartile models)	0.02	0.02	0.06	0.25	0.03	0.05	0.01	0.01
p(equal Treat*Post*Q)‡	0.28	0.68	0.28	0.72	0.44	0.70	0.13	0.13

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Panel B uses same models as panel A, replacing treat and post*treat with quartile interactions. Model on child's vitamin A also control for child's age and gender. † Details on analytical samples and covariate sets in Data section and Appendix. ‡ F-test for equality of Treat*Post*Q coefficients.

Table 6: Effect of P4P on health outcomes and behaviors (percentage point changes)

		Adverse events				
	(1)	(2)	(3)	(4)	(5)	(6)
	Breastfed at least 6 months after birth	Any vision problems in last pregnancy	Night vision problems in last pregnancy	Maternal anemia in current pregnancy	Child anemia at interview date	Child infection in past 2wks
Pre-period mean (%)	93.25	12.51	3.05	21.21	51.26	43.00
A. Average effect (results for covariates omitted)						
Treat*Post	0.29	0.98	1.14	8.11	2.94	4.07
	(1.37)	(3.62)	(1.59)	(16.16)	(7.48)	(6.24)
Post	1.89	-1.72	-2.04	-7.32	-9.68	-0.35
	(1.64)	(2.59)	(1.74)	(13.66)	(7.59)	(9.28)
B. Effect by baseline quality Q						
Treat*Post Q-low	-0.60	-1.10	1.46	25.75	8.57	10.28**
	(2.26)	(4.59)	(2.32)	(15.04)	(12.62)	(3.68)
Treat*Post Q-medium	3.27	4.33	0.48	-18.78	3.07	16.83*
	(2.71)	(3.78)	(2.38)	(27.68)	(8.12)	(8.36)
Treat*Post Q-high	-2.65	1.27	2.59	-4.42	-3.19	-15.46
	(3.80)	(4.46)	(2.51)	(10.35)	(9.14)	(11.44)
Post Q-low	2.17	3.03	-0.82	-18.48	-20.34**	-3.93
	(2.12)	(2.29)	(1.70)	(13.89)	(7.22)	(7.16)
Post Q-medium	1.45	-6.12*	-3.32	14.19	-5.38	-7.71
	(2.69)	(3.10)	(2.50)	(25.33)	(8.88)	(9.11)
Post Q-high	2.07	-3.10	-2.35	2.49	-3.26	10.59
	(2.99)	(3.47)	(2.30)	(5.57)	(11.13)	(13.22)
Wealth Quintile	-0.30	0.05	-0.09	-0.01	-0.90*	-0.57
	(0.33)	(0.38)	(0.19)	(2.13)	(0.47)	(0.90)
Insured	-0.55	0.74	1.04	1.35	-5.59**	0.93
	(0.90)	(1.60)	(0.74)	(4.31)	(1.98)	(1.60)
Mother's education	0.02	0.04	-0.10	0.23	-0.04	-0.35
	(0.12)	(0.26)	(0.13)	(1.27)	(0.21)	(0.24)
Age at birth	-0.21**	0.19	0.07			
	(0.09)	(0.15)	(0.08)			
Mother's age				1.64***	-0.42*	0.06
				(0.54)	(0.21)	(0.19)
Additional covars	Yes	Yes	Yes	Yes	Yes	Yes
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes
Area FE	Yes	Yes	Yes	Yes	Yes	Yes
N†	5,188	3,781	3,781	492	3,734	5,175
R2 (quartile models)	0.07	0.01	0.01	0.03	0.08	0.03
p(equal Treat*Post*Q)‡	0.48	0.41	0.67	0.00	0.74	0.04

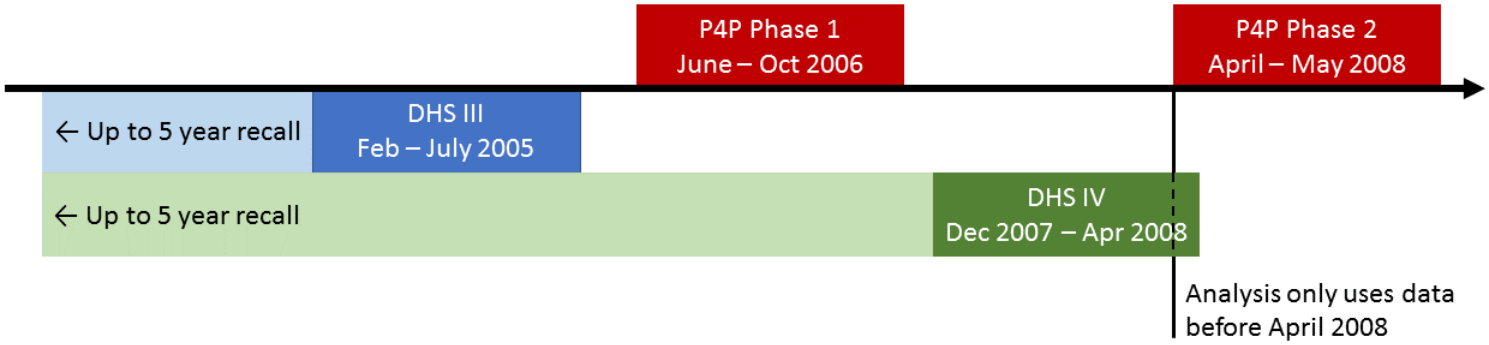
* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Panel B uses same models as panel A, replacing treat and post*treat with quartile interactions. Models on child's anemia and infection also control for child's age and gender. † Details on analytical samples and covariate sets in Data section and Appendix. ‡ F-test for equality of Treat*Post*Q coefficients.

A Paper Appendix

A.1 Timeline

Figure A.1 shows the relation of the P4P rollout phases and the DHS surveys. The analysis data ends in March 2008, before the P4P Phase 2 begins.

Figure A.1: Timeline



A.2 Treatment assignment

In Phase 1 of the national P4P experimental roll-out, between June and October 2006, the 19 districts that had not previously participated in a P4P pilot were organized into 8 groups. Districts in each group had similar rainfall and population density, and their inhabitants had similar sources of livelihoods. Half the districts in each group were randomly assigned to treatment and the remaining half to control status, resulting in 10 treatment and 9 control districts. Prior to the roll-out of P4P, the government of Rwanda implemented a decentralization initiative that involved redrawing district boundaries. Consequently, 2 control districts were combined with Phase 0 districts that had existing P4P pilots and were thus reassigned to treatment status, resulting in 12 treatment districts and 7 control districts in Phase 1. Further details are described in Basinga et al. (2011).

A.3 Combining the DHS waves

Since the two DHS waves were collected in the same EAs, the respondents may overlap in the two surveys, so that births before the earlier wave may be double-counted. We identified households within an EA that were potentially interviewed in both waves by using the respondent's birthdate, and her children's birthdates, sex and twin status (for children born before January 2005, before the first wave). We conservatively assume that all respondents who cannot be uniquely identified across waves may be duplicates. We delete from the 2005 estimation sample the less than 1 percent of mothers and children who may have been re-interviewed in the 2007 wave.

A.4 Covariate sets

The models in Tables 3-5 vary slightly in their covariate sets, depending on the outcome variable. The full results with all covariates are available in Appendix B. There are three main differences:

- Some specifications control for the mother's age at the interview date: contraception, child infection in the past 2 weeks (and infection treated), vitamin A supplementation in the past 6 months, and anemia. These models do not control for the recall period.
- Specifications for outcomes that potentially occurred a long time before the interview date (e.g. prenatal services) control for the mother's age at birth. These specifications also control for the recall period (in years).
- The specifications on child health (immunization, vitamin A supplementation, infection, infection treated, and anemia) also control for the child's age and gender.

A.5 Comparison with estimates from Basinga et al (2011)

In Table A.1 we compare findings from our analysis with those reported by Basinga et al. (2011). Basinga et al. report on a total of eight rewarded measures (their Table 6).

Four outcomes are sufficiently similar across the two analyses. Both analyses find a statistically significant positive effect on institutional deliveries and non-significant effects on prenatal visits and immunizations. Basinga et al. also report a statistically significant ($p<0.10$) positive effect on tetanus vaccination during pregnancy. Our estimated effect is also positive but not statistically significant.

Table A.1: Comparison of Average Treatment Effects with estimates reported in Basinga et al. (2011)

Outcome variable	N	Coeff*100	Stat. significance[†]
A. Findings from this analysis			
Delivery in facility	5,657	9.85	$p<0.05$
Tetanus vaccination	3,753	6.84	NS
Four prenatal visits	3,791	-3.35	NS
Child fully immunized by age 1	4,588	3.75	NS
B. Findings from Basinga et al.			
Institutional delivery	2,108	8.1	$p<0.05$
Tetanus vaccine during prenatal visit	2,856	5.1	$p<0.10$
Four or more prenatal care visits	2,223	0.8	NS
12 to 23 months fully immunized	872	-5.5	NS

[†] NS not significant at $p<0.10$ or lower. Data in Panel A from Table 3 of this paper. Data in Panel B from Basinga et al (2011), Table 6.

Table A.2: Sample and variable definitions for outcome measures

Outcome	Sample and variable definitions
Directly rewarded services	
Delivery in facility	All births in the 5 years preceding each survey wave. 1 if birth occurred in a public or private health facility; 0 if not.
Tetanus vaccination	Most recent birth in the 5 years preceding each survey wave. 1 if at least two tetanus vaccine injections were given during the pregnancy; 0 if fewer than 2 injections were given.
Malaria prophylaxis	Most recent birth in the 5 years preceding each survey wave. 1 if any antimalarial drugs were taken during the pregnancy; 0 if not.
Four prenatal care visits	Most recent birth in the 5 years preceding each survey wave. 1 if 4 or more prenatal care visits were completed during the pregnancy; 0 if fewer than 4 visits.
Child fully immunized	Children age 12-59 months. 1 if fully immunized by age 1 (i.e. received BCG, measles vaccine, 4 doses of oral polio vaccine & 3 doses of DPT vaccine); 0 if not.
Child infection treated	All study children who had an infection (i.e. diarrheal disease, fever or cough) in the 2 weeks preceding the survey. 1 if treated for infection at a health facility; 0 if not.
Modern contraceptive method	All women married or cohabitating with a partner. 1 if currently using modern contraception; 0 if not.
Modern method condit. on need	All fecund women married or cohabitating with a partner, who are not currently pregnant, breastfeeding or amenorrheic, and who report that they wish to limit or space births. 1 if currently using modern contraception; 0 if not.
Modern method req. resupply	All women married or cohabitating with a partner. 1 if currently using a modern contraceptive method requiring regular re-supply (i.e. pill, injectable, condom or female condom, spermicide); 0 if not.
Services in the multiplier	
Trained prenatal provider	Most recent birth in the 5 years preceding each survey wave. 1 if prenatal provider was doctor or nurse/midwife; 0 if not.
Prenatal care (7 measures)	Most recent birth in the 5 years preceding each survey wave. 7 prenatal measures: weighed in pregnancy, blood pressure measured, urinalysis performed, blood drawn, given iron supplement, told of possible complications in pregnancy. Each individual measure equals 1 if the service/procedure was received; 0 if not.
Child vitamin A suppl. in past 6 months	Children age 12-59 months. 1 if received vitamin A supplement in the past 6 months; 0 if no supplement.
Health outcomes and behaviors	
Breastfed at least 6 months after birth	All children who are either alive and older than 6 months, or dead but died after 6 months of age. 1 if child was breastfed for at least 6 months.
Any vision problems	Daytime vision difficulties can be symptomatic of advanced gestational diabetes or pre-eclampsia; nighttime vision difficulties result from vitamin A deficiency. Most recent birth in the 5 years preceding each survey wave. 1 if experienced difficulties with daytime or nighttime vision during pregnancy; 0 if no vision difficulties.
Night vision problems	Nighttime vision difficulties result from vitamin A deficiency. Most recent birth in the 5 years preceding each survey wave. 1 if experienced difficulties with nighttime vision only during pregnancy; 0 if experienced daytime vision difficulties or no vision difficulties.
Maternal anemia	Random sample of 50% of pregnant women surveyed in DHS III; all pregnant women surveyed in DHS IV. 1 if classified as mild, moderate or severe anemia; 0 if classified as no anemia. See DHS documentation for clinical protocols.
Child anemia	Random sample of 50% of the children surveyed in DHS III; all children surveyed in DHS IV. 1 if classified as mild, moderate or severe anemia; 0 if classified as no anemia. See DHS documentation for clinical protocols.
Child infection	All study children who had an infection (i.e. diarrheal disease, fever or cough) in the 2 weeks preceding the survey. 1 if treated for infection at a health facility; 0 if not.

B Online Appendix: Full estimation results

Table B.1: Effect of P4P on indices (range 0-1)

	Rewarded			
	(1)	(2)	(3)	(4)
	Rewarded prenatal index	Contra-ception index	Multiplier prenatal index	Vision index
A. Average effect				
Treat*Post	3.34	1.34	3.79***	1.06
	(3.41)	(2.23)	(1.30)	(2.41)
Post	0.81	15.49***	-0.20	-1.88
	(3.28)	(1.65)	(1.53)	(1.99)
Wealth Quintile	0.72*	1.69***	0.80**	0.03
	(0.39)	(0.53)	(0.29)	(0.26)
Insured	1.60	1.51	-0.07	0.88
	(1.03)	(1.13)	(0.46)	(1.12)
Mother's education	0.14	0.95***	0.28***	-0.03
	(0.16)	(0.17)	(0.08)	(0.19)
Age at birth	-0.19**		0.06	0.13
	(0.08)		(0.07)	(0.10)
Mother's age		-0.56***		
		(0.13)		
Married/cohabitating	1.16	0.00	-0.13	0.63
	(1.33)	(.)	(0.84)	(0.79)
Hhold size	-0.21	1.27***	0.09	0.10
	(0.31)	(0.42)	(0.21)	(0.28)
N children $\leq$ 5 years	-1.69**	1.78**	-1.13***	-0.66
	(0.75)	(0.64)	(0.32)	(0.55)
N births	-1.48***	1.28***	0.00	0.16
	(0.25)	(0.39)	(0.22)	(0.42)
Recall length (years)	2.17***		3.13***	-0.63
	(0.51)		(0.34)	(0.46)
Birth year FE	Yes	No	Yes	Yes
District FE	Yes	Yes	Yes	Yes
N†	3,718	3,709	3,574	3,781
R2	0.20	0.10	0.19	0.01
Pre-period mean (%)	17.72	7.86	50.34	7.78

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Indices constructed as unweighted average of binary outcomes. Sample for contraception index restricted to married/cohabitating women. † Details on analytical samples and covariate sets in Data section and Appendix.

Table B.2: Effect of P4P on indices (range 0-1)

	Rewarded			
	(1)	(2)	(3)	(4)
	Rewarded prenatal index	Contra-ception index	Multiplier prenatal index	Vision index
B. Effect by baseline quality Q (results for covariates omitted)				
Treat*Post Q-low	-1.39	-0.79	2.13	0.18
	(4.77)	(4.27)	(2.48)	(2.93)
Treat*Post Q-medium	6.30	8.03**	6.15***	2.40
	(5.92)	(2.96)	(1.59)	(2.68)
Treat*Post Q-high	3.55	-2.23	3.17	1.93
	(4.78)	(3.43)	(2.29)	(3.26)
Post Q-low	2.70	17.69***	2.77	1.10
	(4.73)	(2.47)	(1.85)	(1.80)
Post Q-medium	1.89	9.02***	-1.65	-4.72*
	(5.23)	(2.27)	(1.45)	(2.44)
Post Q-high	-1.63	18.68***	-1.98	-2.73
	(3.83)	(2.40)	(1.81)	(2.74)
Wealth Quintile	0.77*	1.68***	0.80**	-0.02
	(0.39)	(0.52)	(0.28)	(0.26)
Insured	1.54	1.58	-0.12	0.89
	(1.02)	(1.16)	(0.46)	(1.10)
Mother's education	0.14	0.95***	0.28***	-0.03
	(0.16)	(0.17)	(0.08)	(0.18)
Age at birth	-0.19**		0.06	0.13
	(0.08)		(0.07)	(0.10)
Mother's age		-0.57***		
		(0.13)		
Married/cohabitating	1.22	0.00	-0.18	0.52
	(1.33)	(.)	(0.84)	(0.80)
Hhold size	-0.21	1.26***	0.08	0.10
	(0.31)	(0.42)	(0.21)	(0.28)
N children $\leq$ 5 years	-1.68**	1.79**	-1.12***	-0.64
	(0.75)	(0.64)	(0.32)	(0.56)
N births	-1.46***	1.30***	-0.00	0.15
	(0.24)	(0.39)	(0.22)	(0.43)
Recall length (years)	2.16***		3.10***	-0.65
	(0.51)		(0.33)	(0.47)
Birth year FE	Yes	No	Yes	Yes
District FE	Yes	Yes	Yes	Yes
N[†]	3,718	3,709	3,574	3,781
R2	0.20	0.10	0.18	0.01
p(equal Treat*Post*Q)[‡]	0.51	0.05	0.36	0.80

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Indices constructed as unweighted average of binary outcomes. Sample for contraception index restricted to married/cohabitating women. [†] Details on analytical samples and covariate sets in Data section and Appendix. [‡] F-test for equality of Treat*Post*Q coefficients.

Table B.3: Effect of P4P on rewarded services (percentage point changes)

		Prenatal Care			Child health		Modern contraception		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Delivery in facility	Tetanus vaccination	Malaria prophylaxis	Four prenatal visits	Child fully immunized	Child infection treated in past 2wks	Use of modern method	Modern method condit. on need	Method requiring resupply
A. Average effect									
Treat*Post	9.85**	6.84	4.69	-3.35	3.75	-5.76	1.03	3.94*	1.64
	(3.53)	(5.12)	(4.93)	(3.34)	(6.44)	(4.65)	(2.68)	(2.06)	(1.96)
Post	-1.62	-0.11	-1.11	4.72	8.81	-1.23	16.27***	1.83	14.72***
	(3.78)	(4.66)	(4.79)	(3.87)	(5.38)	(6.51)	(1.98)	(1.33)	(1.42)
Wealth Quintile	2.71***	1.18*	0.21	0.98*	0.72	1.37	1.81***	0.36	1.58***
	(0.44)	(0.66)	(0.70)	(0.51)	(0.52)	(0.86)	(0.57)	(0.48)	(0.50)
Insured	6.49***	1.91	-0.07	2.28	3.79**	10.51***	1.66	1.32	1.37
	(1.30)	(1.69)	(1.82)	(1.36)	(1.40)	(1.49)	(1.28)	(1.25)	(1.03)
Mother's education	1.52***	-0.33	0.78**	0.09	0.43	0.45	1.20***	0.62**	0.70***
	(0.22)	(0.21)	(0.30)	(0.29)	(0.30)	(0.49)	(0.18)	(0.21)	(0.20)
Age at birth	-0.11	-0.69***	-0.11	0.30*	0.31				
	(0.14)	(0.21)	(0.18)	(0.16)	(0.23)				
Mother's age						-0.12	-0.55***	-0.05	-0.58***
						(0.20)	(0.13)	(0.10)	(0.13)
Married/cohabitating	5.34**	-1.64	0.48	4.12**	8.18***	-2.36	0.00	0.00	0.00
	(1.86)	(2.35)	(2.26)	(1.86)	(2.47)	(1.93)	(.)	(.)	(.)
Hhold size	1.29**	-0.93	-0.12	0.41	0.13	-0.46	1.47***	-0.65*	1.06**
	(0.52)	(0.57)	(0.44)	(0.51)	(0.48)	(0.79)	(0.49)	(0.31)	(0.38)
N children $\leq$ 5 years	-3.39***	-2.10***	-0.94	-2.02*	0.23	-0.22	1.92**	2.85***	1.64**
	(1.07)	(0.66)	(1.48)	(1.16)	(1.31)	(1.35)	(0.67)	(0.91)	(0.68)
N births	-3.22***	-3.44***	0.20	-1.40**	-1.13*	0.74	1.21***	0.45	1.35***
	(0.64)	(0.71)	(0.55)	(0.54)	(0.60)	(0.94)	(0.40)	(0.38)	(0.42)
Male child					0.27	0.34			
					(1.17)	(1.62)			
Child's age					-8.04***	-1.94			
					(1.81)	(1.97)			
Recall length (years)	2.76***	1.17	4.83***	0.44	-0.69				
	(0.41)	(1.01)	(0.69)	(0.83)	(1.82)				
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N†	5,657	3,753	3,786	3,791	4,588	2,276	3,709	1,173	3,709
R2	0.17	0.11	0.27	0.03	0.08	0.04	0.10	0.05	0.08
Pre-period mean (%)	29.79	25.03	13.58	14.67	43.63	15.43	8.88	1.61	6.84

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Samples in cols 7-9 restricted to married/cohabitating women. Models on child health also control for child's age and gender. Child infection treated at government facility. † Details on analytical samples and covariate sets in Data section and Appendix.

Table B.4: Effect of P4P on rewarded services (percentage point changes)

		Prenatal Care			Child health		Modern contraception		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Delivery in facility	Tetanus vaccination	Malaria prophylaxis	Four prenatal visits	Child fully immunized	Child infection treated in past 2wks	Use of modern method	Modern method condit. on need	Method requiring resupply
B. Effect by baseline quality Q (results for covariates omitted)									
Treat*Post Q-low	6.69	0.42	2.37	-7.42*	-2.12	-7.17	0.87	8.85**	-2.45
	(5.26)	(4.42)	(9.16)	(3.97)	(8.31)	(6.38)	(4.30)	(3.56)	(4.39)
Treat*Post Q-medium	8.10	11.46	11.32*	-5.94	4.30	-0.08	5.90*	5.18*	10.16***
	(4.93)	(9.77)	(5.38)	(6.30)	(7.36)	(6.55)	(3.32)	(2.89)	(2.89)
Treat*Post Q-high	12.04**	9.41	-2.13	0.56	5.73	-10.85	-2.79	-0.97	-1.66
	(5.51)	(6.48)	(7.21)	(5.19)	(11.36)	(8.76)	(3.80)	(3.69)	(3.56)
Post Q-low	-1.31	8.45*	-4.06	4.41	11.11*	2.26	17.60***	-1.32	17.77***
	(4.59)	(4.66)	(8.67)	(4.51)	(6.15)	(7.25)	(2.30)	(2.21)	(2.74)
Post Q-medium	4.26	-3.63	-1.41	11.29*	15.11**	-3.86	10.67***	-0.37	7.36***
	(5.31)	(8.74)	(3.96)	(5.64)	(5.67)	(5.85)	(2.54)	(0.66)	(2.22)
Post Q-high	-6.32	-5.78	2.36	0.36	1.73	-1.84	19.56***	6.49***	17.80***
	(4.78)	(5.07)	(5.39)	(4.48)	(9.97)	(9.96)	(2.32)	(2.13)	(2.66)
Wealth Quintile	2.76***	1.13*	0.32	1.04*	0.74	1.40	1.79***	0.33	1.57***
	(0.44)	(0.65)	(0.73)	(0.51)	(0.52)	(0.85)	(0.56)	(0.49)	(0.49)
Insured	6.43***	1.85	-0.21	2.26	3.74**	10.43***	1.77	1.34	1.39
	(1.29)	(1.69)	(1.78)	(1.36)	(1.41)	(1.46)	(1.30)	(1.23)	(1.08)
Mother's education	1.52***	-0.32	0.77**	0.09	0.42	0.46	1.20***	0.61***	0.69***
	(0.22)	(0.21)	(0.30)	(0.29)	(0.30)	(0.50)	(0.18)	(0.21)	(0.20)
Age at birth	-0.11	-0.68***	-0.12	0.30*	0.31				
	(0.14)	(0.21)	(0.18)	(0.16)	(0.23)				
Mother's age						-0.11	-0.55***	-0.05	-0.58***
						(0.20)	(0.13)	(0.11)	(0.13)
Married/cohabitating	5.43***	-1.81	0.64	4.27**	8.17***	-2.37	0.00	0.00	0.00
	(1.84)	(2.35)	(2.27)	(1.86)	(2.44)	(2.01)	(.)	(.)	(.)
Hhold size	1.27**	-0.94	-0.13	0.41	0.12	-0.45	1.46***	-0.66**	1.06**
	(0.52)	(0.58)	(0.44)	(0.52)	(0.49)	(0.79)	(0.49)	(0.31)	(0.38)
N children $\leq$ 5 years	-3.41***	-2.05***	-0.92	-2.04*	0.24	-0.30	1.92**	2.84***	1.65**
	(1.07)	(0.67)	(1.45)	(1.13)	(1.31)	(1.33)	(0.67)	(0.91)	(0.68)
N births	-3.21***	-3.46***	0.25	-1.41**	-1.12*	0.72	1.23***	0.44	1.37***
	(0.65)	(0.71)	(0.54)	(0.56)	(0.59)	(0.93)	(0.40)	(0.39)	(0.41)
Male child					0.21	0.35			
					(1.17)	(1.62)			
Child's age					-8.02***	-2.00			
					(1.83)	(2.05)			
Recall length (years)	2.73***	1.10	4.87***	0.43	-0.74				
	(0.41)	(1.01)	(0.70)	(0.82)	(1.83)				
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N[†]	5,657	3,753	3,786	3,791	4,588	2,276	3,709	1,173	3,709
R2	0.17	0.11	0.27	0.03	0.08	0.04	0.11	0.06	0.09
p(equal Treat*Post*Q)[‡]	0.72	0.27	0.42	0.46	0.72	0.61	0.15	0.11	0.02

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Samples in cols 7-9 restricted to married/cohabitating women. Models on child health also control for child's age and gender. Child infection treated at government facility. [†] Details on analytical samples and covariate sets in Data section and Appendix. [‡] F-test for equality of Treat*Post*Q coefficients.

Table B.5: Effect of P4P on services in multiplier (percentage point changes)

	Prenatal Care							
	(1) Trained prenatal provider	(2) Weighed	(3) Blood pressure taken	(4) Blood drawn	(5) Iron suppl.	(6) Urinalysis	(7) Told of possible complica-tions	(8) Child vit. A suppl. in past 6 months
A. Average effect								
Treat*Post	0.06	-1.96	5.85	6.61	9.40***	5.13**	0.62	2.87
	(1.62)	(1.85)	(4.12)	(5.46)	(3.12)	(2.30)	(2.60)	(5.44)
Post	-1.53	4.98**	0.52	-3.84	-1.14	-2.16	0.44	-2.34
	(1.41)	(1.86)	(4.07)	(4.65)	(4.29)	(2.80)	(3.13)	(5.84)
Wealth Quintile	0.42	0.18	1.73**	0.90	1.92***	0.82	-0.06	0.72
	(0.29)	(0.30)	(0.71)	(0.86)	(0.63)	(0.52)	(0.24)	(0.69)
Insured	1.26	0.24	-0.49	-0.07	0.34	-0.17	-0.21	3.50**
	(0.75)	(0.66)	(1.85)	(1.45)	(1.63)	(1.16)	(1.04)	(1.63)
Mother's education	0.22	0.24***	0.95***	-0.20	0.41*	0.44*	0.27*	0.09
	(0.18)	(0.08)	(0.25)	(0.27)	(0.21)	(0.25)	(0.14)	(0.26)
Age at birth	-0.19*	0.14	0.16	-0.21	0.04	0.06	-0.01	
	(0.10)	(0.09)	(0.15)	(0.16)	(0.18)	(0.12)	(0.09)	
Mother's age								0.01
								(0.15)
Married/cohabitating	4.10***	0.19	1.87	0.03	0.32	-2.36	0.35	-0.13
	(1.30)	(0.98)	(2.43)	(2.39)	(2.11)	(1.83)	(1.02)	(1.87)
Hhold size	0.25	0.19	0.44	0.42	-0.55	0.68	-0.17	-0.13
	(0.24)	(0.22)	(0.47)	(0.41)	(0.69)	(0.47)	(0.23)	(0.68)
N children $\leq$ 5 years	-0.24	-1.10*	-1.14	-0.39	-1.93	-2.74***	-0.74	0.41
	(0.47)	(0.54)	(1.23)	(1.11)	(1.24)	(0.91)	(0.81)	(1.13)
N births	-0.04	-0.35	-0.25	-0.28	0.65	-0.43	0.84**	-0.02
	(0.27)	(0.32)	(0.48)	(0.55)	(0.46)	(0.44)	(0.37)	(0.43)
Male child								2.37**
								(1.07)
Child's age								-0.40
								(1.41)
Recall length (years)	0.33	1.14*	5.59***	9.61***	3.65***	2.23**	0.11	
	(0.41)	(0.62)	(1.08)	(0.84)	(0.76)	(0.85)	(0.61)	
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N[†]	3,803	3,651	3,650	3,648	3,768	3,626	3,644	3,856
R2	0.02	0.02	0.07	0.25	0.03	0.04	0.01	0.01
Pre-period mean (%)	95.20	94.66	74.65	38.06	30.01	8.68	5.71	85.14

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Model on child's vitamin A also control for child's age and gender. [†] Details on analytical samples and covariate sets in Data section and Appendix.

Table B.6: Effect of P4P on services in multiplier (percentage point changes)

	Prenatal Care							
	(1) Trained prenatal provider	(2) Weighed	(3) Blood pressure taken	(4) Blood drawn	(5) Iron suppl.	(6) Urinalysis	(7) Told of possible complica-tions	(8) Child vit. A suppl. in past 6 months
B. Effect by baseline quality Q (results for covariates omitted)								
Treat*Post Q-low	-2.46	-3.57	3.97	1.59	1.60	3.63	4.07	-2.12
	(2.04)	(3.85)	(7.06)	(9.64)	(6.64)	(3.39)	(3.87)	(9.11)
Treat*Post Q-medium	0.86	-0.42	12.69**	8.95	14.64*	7.27*	2.89	10.14
	(2.32)	(2.00)	(5.21)	(8.42)	(7.50)	(3.59)	(4.50)	(7.15)
Treat*Post Q-high	1.48	-1.05	-0.12	10.74	12.70*	4.68	-5.81*	0.37
	(1.94)	(2.31)	(4.89)	(6.29)	(6.80)	(4.71)	(2.97)	(7.95)
Post Q-low	-1.04	8.37***	8.06	1.99	5.68	-3.59	0.68	-1.69
	(1.46)	(2.44)	(4.64)	(7.42)	(6.63)	(2.74)	(4.33)	(8.07)
Post Q-medium	-1.17	3.02	-2.01	-6.84	-4.84	-3.77	-0.28	-4.80
	(1.69)	(2.04)	(5.49)	(7.33)	(6.08)	(3.11)	(3.95)	(6.88)
Post Q-high	-2.27	3.08	-4.68	-7.43	-4.89	0.39	0.90	-0.53
	(1.71)	(1.90)	(3.55)	(5.89)	(6.06)	(3.71)	(2.77)	(7.39)
Wealth Quintile	0.42	0.15	1.77**	0.85	1.88***	0.82	-0.02	0.77
	(0.29)	(0.30)	(0.69)	(0.87)	(0.60)	(0.51)	(0.24)	(0.68)
Insured	1.25	0.23	-0.69	-0.07	0.30	-0.16	-0.30	3.38**
	(0.74)	(0.65)	(1.80)	(1.43)	(1.60)	(1.14)	(1.05)	(1.57)
Mother's education	0.22	0.24***	0.94***	-0.19	0.41*	0.44	0.27*	0.10
	(0.18)	(0.08)	(0.25)	(0.28)	(0.22)	(0.25)	(0.14)	(0.26)
Age at birth	-0.19*	0.14	0.18	-0.21	0.04	0.05	0.00	
	(0.10)	(0.09)	(0.15)	(0.16)	(0.18)	(0.12)	(0.10)	
Mother's age								0.01
								(0.15)
Married/cohabitating	4.14***	0.08	1.69	-0.10	0.24	-2.30	0.29	-0.19
	(1.31)	(0.99)	(2.57)	(2.38)	(2.11)	(1.84)	(1.04)	(1.89)
Hhold size	0.25	0.19	0.41	0.42	-0.55	0.69	-0.19	-0.12
	(0.24)	(0.22)	(0.48)	(0.40)	(0.69)	(0.47)	(0.23)	(0.69)
N children $\leq$ 5 years	-0.23	-1.08*	-1.12	-0.35	-1.88	-2.72***	-0.76	0.48
	(0.47)	(0.55)	(1.25)	(1.11)	(1.27)	(0.90)	(0.81)	(1.08)
N births	-0.03	-0.36	-0.27	-0.29	0.64	-0.41	0.84**	-0.02
	(0.27)	(0.33)	(0.48)	(0.55)	(0.46)	(0.45)	(0.38)	(0.44)
Male child								2.31**
								(1.08)
Child's age								-0.42
								(1.42)
Recall length (years)	0.34	1.11*	5.45***	9.58***	3.62***	2.29**	0.06	
	(0.41)	(0.60)	(1.08)	(0.84)	(0.73)	(0.85)	(0.62)	
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N[†]	3,803	3,651	3,650	3,648	3,768	3,626	3,644	3,856
R2	0.02	0.02	0.06	0.25	0.03	0.05	0.01	0.01
p(equal Treat*Post*Q)[‡]	0.28	0.68	0.28	0.72	0.44	0.70	0.13	0.13

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Model on child's vitamin A also control for child's age and gender. [†] Details on analytical samples and covariate sets in Data section and Appendix. [‡] F-test for equality of Treat*Post*Q coefficients.

Table B.7: Effect of P4P on health outcomes and behaviors (percentage point changes)

		Adverse events				
	(1)	(2)	(3)	(4)	(5)	(6)
	Breastfed at least 6 months after birth	Any vision problems in last pregnancy	Night vision problems in last pregnancy	Maternal anemia in current pregnancy	Child anemia at interview date	Child infection in past 2wks
A. Average effect						
Treat*Post	0.29	0.98	1.14	8.11	2.94	4.07
	(1.37)	(3.62)	(1.59)	(16.16)	(7.48)	(6.24)
Post	1.89	-1.72	-2.04	-7.32	-9.68	-0.35
	(1.64)	(2.59)	(1.74)	(13.66)	(7.59)	(9.28)
Wealth Quintile	-0.32	0.10	-0.05	0.03	-1.09*	-0.69
	(0.34)	(0.39)	(0.19)	(1.94)	(0.54)	(0.89)
Insured	-0.51	0.76	1.01	0.59	-5.64**	1.03
	(0.90)	(1.64)	(0.75)	(4.50)	(1.99)	(1.59)
Mother's education	0.03	0.04	-0.10	0.35	-0.06	-0.34
	(0.12)	(0.27)	(0.13)	(1.29)	(0.21)	(0.24)
Age at birth	-0.21**	0.19	0.06			
	(0.09)	(0.15)	(0.08)			
Mother's age				1.47***	-0.42*	0.06
				(0.42)	(0.21)	(0.19)
Married/cohabitating	2.55**	0.77	0.49	6.58	-3.34	-2.45
	(1.20)	(1.19)	(0.82)	(10.36)	(2.14)	(2.20)
Hhold size	-0.44	0.35	-0.15	-2.86	-0.52	0.33
	(0.56)	(0.36)	(0.25)	(2.27)	(0.55)	(0.60)
N children $\leq$ 5 years	4.42***	-1.25	-0.07	6.53*	-0.07	-0.92
	(0.76)	(0.82)	(0.47)	(3.33)	(1.60)	(1.64)
N births	0.46	0.04	0.27	-0.87	1.38*	-0.88
	(0.42)	(0.62)	(0.28)	(1.67)	(0.71)	(0.64)
Male child					1.50	-0.88
					(1.35)	(1.19)
Child's age					-7.38**	-6.76**
					(2.70)	(2.55)
Recall length (years)	-4.34***	-1.35	0.09			
	(0.52)	(0.78)	(0.33)			
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	Yes	Yes
N[†]	5,188	3,781	3,781	492	3,734	5,175
R2	0.07	0.02	0.01	0.02	0.09	0.03
Pre-period mean (%)	93.25	12.51	3.05	21.21	51.26	43.00

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Models on child's anemia and infection also control for child's age and gender. [†] Details on analytical samples and covariate sets in Data section and Appendix.

Table B.8: Effect of P4P on health outcomes and behaviors (percentage point changes)

		Adverse events				
	(1) Breastfed at least 6 months after birth	(2) Any vision problems in last pregnancy	(3) Night vision problems in last pregnancy	(4) Maternal anemia in current pregnancy	(5) Child anemia at interview date	(6) Child infection in past 2wks
B. Effect by baseline quality Q (results for covariates omitted)						
Treat*Post Q-low	-0.60	-1.10	1.46	25.75	8.57	10.28**
	(2.26)	(4.59)	(2.32)	(15.04)	(12.62)	(3.68)
Treat*Post Q-medium	3.27	4.33	0.48	-18.78	3.07	16.83*
	(2.71)	(3.78)	(2.38)	(27.68)	(8.12)	(8.36)
Treat*Post Q-high	-2.65	1.27	2.59	-4.42	-3.19	-15.46
	(3.80)	(4.46)	(2.51)	(10.35)	(9.14)	(11.44)
Post Q-low	2.17	3.03	-0.82	-18.48	-20.34**	-3.93
	(2.12)	(2.29)	(1.70)	(13.89)	(7.22)	(7.16)
Post Q-medium	1.45	-6.12*	-3.32	14.19	-5.38	-7.71
	(2.69)	(3.10)	(2.50)	(25.33)	(8.88)	(9.11)
Post Q-high	2.07	-3.10	-2.35	2.49	-3.26	10.59
	(2.99)	(3.47)	(2.30)	(5.57)	(11.13)	(13.22)
Wealth Quintile	-0.30	0.05	-0.09	-0.01	-0.90*	-0.57
	(0.33)	(0.38)	(0.19)	(2.13)	(0.47)	(0.90)
Insured	-0.55	0.74	1.04	1.35	-5.59**	0.93
	(0.90)	(1.60)	(0.74)	(4.31)	(1.98)	(1.60)
Mother's education	0.02	0.04	-0.10	0.23	-0.04	-0.35
	(0.12)	(0.26)	(0.13)	(1.27)	(0.21)	(0.24)
Age at birth	-0.21**	0.19	0.07			
	(0.09)	(0.15)	(0.08)			
Mother's age				1.64***	-0.42*	0.06
				(0.54)	(0.21)	(0.19)
Married/cohabitating	2.56**	0.62	0.43	7.05	-3.10	-2.62
	(1.21)	(1.18)	(0.84)	(11.03)	(2.12)	(2.20)
Hhold size	-0.44	0.35	-0.14	-2.94	-0.48	0.33
	(0.56)	(0.37)	(0.25)	(2.34)	(0.54)	(0.61)
N children $\leq$ 5 years	4.42***	-1.21	-0.07	6.92*	-0.12	-0.97
	(0.76)	(0.82)	(0.48)	(3.36)	(1.59)	(1.62)
N births	0.48	0.03	0.26	-1.30	1.34*	-0.86
	(0.42)	(0.62)	(0.28)	(1.89)	(0.69)	(0.63)
Male child					1.52	-0.78
					(1.39)	(1.20)
Child's age					-7.47**	-6.92**
					(2.70)	(2.61)
Recall length (years)	-4.36***	-1.38*	0.08			
	(0.52)	(0.79)	(0.33)			
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	Yes	Yes
N†	5,188	3,781	3,781	492	3,734	5,175
R2	0.07	0.01	0.01	0.03	0.08	0.03
p(equal Treat*Post*Q)‡	0.48	0.41	0.67	0.00	0.74	0.04

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. Models on child's anemia and infection also control for child's age and gender. † Details on analytical samples and covariate sets in Data section and Appendix. ‡ F-test for equality of Treat*Post*Q coefficients.

C Online Appendix: Raw difference-in-difference results (no covariates)

Table C.1: Effect of P4P on indices (range 0-1)

	Rewarded			
	(1)	(2)	(3)	(4)
	Rewarded prenatal index	Contra-ception index	Multiplier prenatal index	Vision index
C. Regressions without covariates				
Treat*Post	3.50	2.22	3.65**	1.11
	(3.51)	(2.32)	(1.35)	(2.48)
Post	0.19	14.85	-0.29	-1.91
	(3.36)	(1.49)	(1.54)	(2.09)
Constant	11.79***	8.03	59.96***	11.17***
	(3.23)	(0.63)	(2.24)	(3.85)
Birth year FE	Yes	No	Yes	Yes
District FE	Yes	Yes	Yes	Yes
N†	3,718	3,709	3,574	3,781
R2	0.15	0.05	0.13	0.01
Pre-period mean (%)	17.72	7.86	50.34	7.78

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition.
† Details on reference populations in Data section and Appendix.

Table C.2: Effect of P4P on rewarded services (percentage point changes)

		Prenatal Care			Child health		Modern contraception		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Delivery in facility	Tetanus vaccination	Malaria prophylaxis	Four prenatal visits	Child fully immunized	Child infection treated in past 2wks	Use of modern method	Modern method condit. on need	Method requiring resupply
C. Regressions without covariates									
Treat*Post	9.95***	7.27	4.40	-3.01	4.05	-4.98	2.01	4.14	2.42
	(3.16)	(5.32)	(5.01)	(3.18)	(6.55)	(4.61)	(2.82)	(2.14)	(1.99)
Post	-3.62	-1.07	-1.35	4.06	11.33*	-4.15	15.64	2.84	14.05
	(3.49)	(4.97)	(4.79)	(3.75)	(5.53)	(4.55)	(1.90)	(1.43)	(1.18)
Constant	18.17***	19.43**	6.77	8.41*	37.04***	6.06***	9.14	1.61	6.92
	(2.12)	(7.54)	(4.09)	(4.35)	(3.32)	(2.05)	(0.76)	(0.47)	(0.56)
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N†	5,657	3,753	3,786	3,791	4,588	2,276	3,709	1,173	3,709
R2	0.07	0.01	0.25	0.02	0.05	0.01	0.05	0.02	0.05
Pre-period mean (%)	29.79	25.03	13.58	14.67	43.63	15.43	8.88	1.61	6.84

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. † Details on reference populations in Data section and Appendix.

Table C.3: Effect of P4P on services in multiplier (percentage point changes)

	Prenatal Care							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Trained prenatal provider	Weighed	Blood pressure taken	Blood drawn	Iron suppl.	Urinalysis	Told of possible complications	Child vit. A suppl. in past 6 months
C. Regressions without covariates								
Treat*Post	-0.01	-1.92	5.64	6.19	9.22**	4.99**	0.44	3.16
	(1.67)	(1.90)	(4.34)	(5.46)	(3.22)	(2.21)	(2.59)	(5.74)
Post	-1.94	4.80**	0.06	-3.65	-1.35	-2.39	0.48	-2.62
	(1.43)	(1.91)	(4.28)	(4.79)	(4.22)	(2.69)	(3.03)	(4.38)
Constant	92.80***	89.72***	84.00***	81.47***	35.04***	24.19***	7.36	85.46***
	(3.91)	(3.51)	(2.83)	(4.06)	(7.71)	(7.07)	(4.34)	(2.44)
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N†	3,803	3,651	3,650	3,648	3,768	3,626	3,644	3,856
R2	0.00	0.01	0.03	0.20	0.01	0.02	0.00	0.00
Pre-period mean (%)	95.20	94.66	74.65	38.06	30.01	8.68	5.71	85.14

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. † Details on reference populations in Data section and Appendix.

Table C.4: Effect of P4P on health outcomes and behaviors (percentage point changes)

		Adverse events				
	(1)	(2)	(3)	(4)	(5)	(6)
	Breastfed at least 6 months after birth	Any vision problems in last pregnancy	Night vision problems in last pregnancy	Maternal anemia in current pregnancy	Child anemia at interview date	Child infection in past 2wks
C. Regressions without covariates						
Treat*Post	0.55	1.06	1.16	6.65	2.49	4.65
	(1.46)	(3.75)	(1.59)	(14.82)	(7.97)	(6.05)
Post	1.29	-1.83	-1.99	-2.24	-29.93***	-17.84***
	(1.79)	(2.72)	(1.79)	(14.07)	(5.91)	(5.19)
Constant	97.15***	15.52***	6.83*	11.43**	42.73***	34.22***
	(1.02)	(5.14)	(3.28)	(5.08)	(4.21)	(3.07)
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	Yes	Yes
N†	5,188	3,781	3,781	492	3,734	5,175
R2	0.02	0.01	0.00	0.02	0.08	0.03
Pre-period mean (%)	93.25	12.51	3.05	21.21	51.26	43.00

* p<0.10, ** p<0.05, *** p<0.01 OLS, s.e. clustered within districts and p-values based on t-distribution with G-2 d.f.. Scaled by 100 for exposition. † Details on reference populations in Data section and Appendix.