CENTER
FOR
GLOBAL
DEVELOPMENT

Pay-for-Performance Contracts in the Lab and the Real World: Evidence from Nigeria

Sebastian Bauhoff and Eeshani Kandpal

Abstract

A two-stage experiment disentangles the effect of various aspects of pay-for-performance contracts. The first is a lab-in-the-field experiment where 1,359 health workers are primed with a checklist of salient clinical tasks, then randomized within 690 clinics to receive no incentives, rewards, or penalties for treating hypothetical patients. Both rewards and penalties improve performance by 20 percent and generate spillovers on unincentivized tasks, but small incentives capture most gains. In the second stage, lab impacts translate into the real world: lab PFP exposure improves by 20 percent the care provided to real-world patients even after the lab experiment.

KEYWORDS
Pay-for-performance; Health workers; Lab-in-the-field experiment

JEL CODES
C93; I11; I15; J41; J45; M52; O15

WORKING PAPER 677 · JANUARY 2024

Pay-for-Performance Contracts in the Lab and the Real World: Evidence from Nigeria

Sebastian Bauhoff
Department of Global Health and Population, Harvard T.H. Chan School of Public Health (sbauhoff@hsph.harvard.edu)

Eeshani Kandpal
Center for Global Development (ekandpal@cgdev.org)

We are grateful to Samuel Adebayo and Florence Oke at the Nigerian National Bureau of Statistics, Emmanuel Meribole at the Federal Ministry of Health, Opeyemi Fadeyibi, Benjamin Loevinsohn, Ayodeji Oluwole Odutolu, and Elina Pradhan for their feedback and collaboration, and to Kevin McGee and Lena Nguyen for their support of experiment design and implementation. We thank Kathleen Beegle, Igna Bonfrer, Jishnu Das, Jonathan de Quidt, Damien de Walque, Quy-Toan Do, Karen Eggleston, David Evans, Jed Friedman, James Habyarimana, Sylvan Herskowitz, Madhulika Khanna, Jason Kerwin, Yusuke Kuwayama, Kenneth Leonard, Jeffrey Livingston, Margaret McConnell, David McKenzie, Owen Ozier, Antonn Park, Bob Rijkers, Ellen Van de Poel, Tom Van Ourti, Adam Wagstaff, Andrew Zeitlin, seminar audiences at Harvard SPH, IFPRI, Georgetown, Bentley, and participants of the ASHEcon, MeasureDev, iHEA, and NEUDC conferences for their comments. Diwakar Kishore and Mark Conlon provided excellent research assistance. Financial support was provided by the World Bank's Health Results Innovation Trust Fund. The experiment was approved by the Health Media Lab Institutional Review Board (IRB #00001211) and registered with the AER RCT Registry as study AEARCTR-0002482. The findings are the opinions of the authors and do not represent the opinions of the World Bank, its Executive Directors, or the governments they represent. Errors and omissions are our own.

Sebastian Bauhoff and Eeshani Kandpal. 2024. “Pay-for-Performance Contracts in the Lab and the Real World: Evidence from Nigeria.” CGD Working Paper 677. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/pay-performance-contracts-lab-and-real-world-evidence-nigeria

CENTER FOR GLOBAL DEVELOPMENT
2055 L Street, NW Fifth Floor
Washington, DC 20036
202.416.4000

1 Abbey Gardens
Great College Street
London
SW1P 3SE

www.cgdev.org

1 Introduction

Pay for performance (PFP) is a common contracting approach in settings that have principal-agent problems, such as health service delivery (Miller and Babiarz, 2013; Lazear, 2000; Prendergast, 1999). PFP contracts typically provide agents with a checklist of selected actions and a financial incentive assigned to each action, thus engaging agents through two economic channels. First, the checklist provides a prime by explicitly communicating what actions the principal values and prioritizes. Agents may respond to this prime alone, for example, because it reduces their uncertainty about how to allocate their effort or because they are intrinsically motivated (Arrow, 1963). Second, agents may respond to the financial incentive presented through the contract (de Quidt, 2018; DellaVigna and Pope, 2018; Rothstein, 2015; Duflo et al., 2012; Hossain and List, 2012; Kahneman et al., 1991; Kahneman and Tversky, 1984). PFP contracts need to consider the roles of these two channels and their potential interactions because, for instance, incentives may amplify the effect of the prime. Moreover, in practice PFP contracts are often incomplete and time-bound, which may lead to spillovers onto excluded actions and behavior offsetting any gains once the contracts end (Celhay et al., 2019; Sherry, 2016; Miller and Babiarz, 2013).

We conduct a two-stage experiment, laid out in Figure 1, to examine the effect of adding varying sizes of incentives to the prime; spillovers on actions that are unprimed and primed-only; and whether PFP effects persist in real-world tasks after the contract has ended. For the first stage, we conduct a lab-in-the-field experiment that randomizes 1,359 maternity health workers in 690 health facilities to one of three study arms: information (i.e., priming without incentives), rewards, and penalties.[1] In all arms, we ask participants to review medical records of five hypothetical patients presenting for labor and delivery, and identify the actions that would be clinically necessary. All participants received a token participation fee and a checklist of seven common clinical actions. The checklist does not contain all actions that are clinically necessary for all patients.[2] For five of the seven actions on the checklist, we provide financial incentives to participants in the rewards and

[1] A design feature, described in detail in Section 3 and Appendix B.2, precluded the inclusion of a pure control arm in the lab experiment stage.

[2] We assess adherence using a standard World Health Organization protocol of essential procedures for childbirth, which is closely aligned with the participants' day-to-day work. We measure performance as the share of clinically necessary actions that participants correctly identify, of the universe of 15 potentially applicable clinical actions.

penalty arms. In the subsequent second stage, we randomly select a subset of 344 participants from the first-stage experiment and 67 pure control health workers from the same facilities and assess the care they provide to real-world antenatal care patients, which shares many common clinical actions with labor and delivery.

We report on four key findings from the first-stage lab experiment and the second stage observation of real-world care. First, adding incentives to the prime improves worker performance in the first-stage lab experiment, regardless of whether they are framed as rewards or penalties.[3] Participants in the reward and penalty arms performed similarly and about 20 percent better than those in the information arm. The incentive effect is notable particularly in a setting such as health care, where workers may be intrinsically motivated (Mohanan et al., 2021; McGuire, 2000; Arrow, 1963) and respond to interventions communicating the importance or value of effort, even without incentives (Gauri et al., 2021; Brock et al., 2018; Leonard and Masatu, 2017; Ashraf et al., 2014; Kolstad, 2013). Moreover, we find no impact of loss framing relative to reward framing.[4] Second, a small incentive generates most of the performance gains in the first stage experiment, perhaps because incentives amplify the prime, at least in the short run. Third, we find evidence for positive spillovers of incentives on unincentivized actions. The rewards and penalty arms perform 14 percent better on actions that are not on the checklist (unprimed and unincentivized) relative to participants in the information arm of the lab experiment. This could arise because of complementarities in production (Sherry, 2016; Mullen et al., 2010). Finally, participants who were randomized to financial incentives in the first-stage lab experiment also perform 20-25 percent better on real-world clinical actions that are similar to actions that were incentivized (i.e., not merely primed) in the first stage. By and large, participants assigned to the information arm in the first stage perform similarly to workers who were not part of the first-stage lab experiment. This suggests that the effect of adding incentives may persist into real world tasks whereas the effect of priming does not.

[3] While a design element of the lab experiment precluded the inclusion of a pure control arm without either information or incentives, we use performance on primed actions to benchmark performance in the no-incentives case, noting that these participants still received the same priming intervention. In the information arm, performance on unprimed actions (28 percent) is similar to performance on primed-or-paid actions (26 percent), suggesting that the prime alone may not substantively increase performance.

[4] The reward and penalty arms have differing reference points, because we designed the expected payouts to be equal across the two arms. The penalty arm has a 2.5 times higher reference point that decreases in incorrect performance, while the reward arm has a lower reference point that increases in correct performance. Despite a higher reference point in the penalty arm, we do not find an impact of loss framing relative to reward framing. Theoretically, we show that this result is consistent with loss neutral agents.

Pairing a lab-in-the-field experiment with real-world observation of provider effort has several advantages for studying the attributes of PFP contracts. Probing the mechanisms of PFP contracts is challenging in real-world settings because many PFP programs include ancillary interventions that by themselves can represent substantive reforms, such as providing operating budgets, reforming clinic management practices, and improving accountability (de Walque et al., 2022; Diaconu et al., 2021; Basinga et al., 2011). Moreover, real-world patient care can be complex and difficult to judge. The lab-experimental first stage provides a controlled environment to disentangle nuances of contract design. First, the tasks are designed to allow precise assessments of performance against the universe of potentially relevant clinical tasks. Second, we prime a subset of potentially necessary actions through the checklist and incentivize some of those actions, which allows us to estimate spillovers on actions that are unprimed or unincentivized. Third, incentivized actions differ in the associated value of the incentive, allowing us to examine price responses. Fourth, we can isolate the effects of loss framing by designing isomorphic contracts for the reward and penalty arms. In these arms, all actions lead to equivalently valued gains or losses, and they have the same maximum and minimum payouts. Fifth, our hypothetical first-stage experiment also allows us to cleanly examine responses to a small number of deliberately misaligned incentives that cause unnecessary and potentially harmful treatment in a real-world setting. While there is evidence of such actions in LMICs (Lopez et al., forthcoming; Das et al., 2016), it is not known to what extent providers may be constrained by intrinsic motivation, altruism, and reputational concerns when responding to misaligned incentives. We indeed show a muted response to misaligned incentives, suggesting that overprovision of care may be real but possibly limited in many settings. Finally, the second stage of our study allows us to examine persistence of priming and priming-with-incentives in the real world, after the contract has ended (Fryer et al., 2022). We put our results through a battery of robustness tests and find them to be strongly robust. We also use separate assessments to show that the lab-in-the-field experiment does merely capture knowledge.

While our first-stage lab experiment is hypothetical, i.e. participants recommend actions without performing them, we show that participants behave as though it were real. This is because the experiment design and context are realistic (Harrison and List, 2004; Prendergast, 1999): we use small but meaningful incentives (the maximum payout, USD 6, in our task is about 5 percent of monthly salary, USD 113); the task mimics what participants do routinely in their jobs as maternity care workers; and the study sites are their primary workplaces. Our participants' performance and response patterns align with behavior observed in real-world primary health care provision in low- and middle-income countries (LMICs), including the low overall level of performance and the provision of unnecessary care (Kruk et al., 2016). Our performance measure is significantly and positively correlated with separate assessments of participants' knowledge that were conducted alongside our experiment, but incentives have an additional effect on performance. Moreover, participants' response patterns are not consistent with a mechanical response to the checklist or incentives. For instance, participants engage on the margin even in the information arm where they do not stand to gain financially. In addition, the difference in performance across arms is largest for the middle of the performance distribution rather than the bottom, suggesting that the incentives only did not affect participants who would have not paid any attention otherwise. Similarly, participants frequently identified actions that are clinically necessary but not primed or paid. Finally, we find that these effects persist in the subsequent real-world clinical interactions.

Our findings have important implications for the design of PFP contracts in health care.[5] PFP has long been used in health systems in high-income contexts (Campbell et al., 2007; Mendelson et al., 2017) and is increasingly deployed in LMICs, where the poor quality of health care services stems partly from low effort by health workers (Leonard and Masatu, 2010; Das et al., 2008; Leonard et al., 2007). Penalties are less common in PFP contracts but are nevertheless used, for example, in the United States' Medicare's Nonpayment Program, which withholds reimbursements for costs related to hospital-acquired conditions with the goal of reducing the incidence of these conditions (Gupta, 2021). Our findings support adopting at least small incentives in lieu of information-only interventions, such as job aids, the dissemination of guidelines, or training programs (Rowe et al., 2005). They also suggest that, in practice, there may be little cost to implementing the simpler and more politically palatable rewards frame in real-world PFP contracts. Finally, we contribute theoretical and empirical evidence on the question of spillovers from incomplete contracts. The

[5] Our PFP contract pays for inputs and uses a threshold design where payments are conditional on achieving a goal, and thus mimics PFP programs in many low- and middle-income countries (Kandpal, 2016). Evidence suggests that the design choice of rewarding inputs or outputs can interact with worker characteristics to influence the effectiveness of the intervention (Mohanan et al., 2021). Thus, the results presented here may only apply to PFP schemes that reward inputs and use a threshold design.

limited available literature does not find definitive evidence of spillovers of PFP in health care, even for large schemes such as the United Kingdom's Quality and Outcomes Framework PFP scheme (Campbell et al., 2007) or national PFP schemes in LMICs (Diaconu et al., 2021; Celhay et al., 2019; Sherry et al., 2017). However, concerns have been raised about multi-tasking crowding out effort on unincentivized tasks (Prendergast, 1999). In contrast, we find evidence of positive spillovers and link this to a simple extension of Holmstrom and Milgrom (1991) that describes how PFP designs may leverage production complementarities to generate positive spillovers on actions outside the PFP contract, or conversely can lead to negative spillovers if actions are substitutes in production.

2 Context

Nigeria's maternal mortality ratio was the second highest in the world in 2017— and at 1,127 per 100,000 live births almost 10 times higher than India's (World Bank, 2020). Health service delivery has stagnated over the last three decades and public health workers' performance is inadequate (Okeke and Abubakar, 2020; Khanna et al., 2021). At the same time, 75 percent of health workers in our study reported working seven days a week, for an average of six hours a day. In our sample, the median monthly gross salary is 43,000 Nigerian Naira or about USD 113. Only a third of health workers reported receiving a salary increase in the last two years; a quarter had not received their full pay for the previous month; and 63 percent reported not having received their entire salary for the past year.

2.1 Study design

Our study was embedded in the endline survey of a concurrent cluster-randomized impact evaluation of different health facility financing modalities in Nigeria (Khanna et al., 2021). This larger trial is described in further detail in Appendix B.1. This survey was conducted between July and October 2017. During the survey, we conducted our first-stage lab experiment in 690 of 691 primary health facilities that were randomly selected for the endline survey of the impact evaluation.[6] At each facility, the survey sampled from the roster of health workers on site who routinely provide antenatal or under-five curative care. We administered in-depth interviews and knowledge tests to each of these workers. The experiment was conducted at the end of this interview, and all interviewees who routinely provide antenatal care were eligible to participate. To minimize spillovers, the experiment was conducted simultaneously among all participating health workers within a facility. The survey also included the direct clinical observation of a randomly selected subset of health workers during interactions with actual patients 24-36 hours after the lab experiment. In these observations, health workers were observed while providing routine antenatal care to pregnant women. The workers sampled for these observations were selected through an independent randomization process. These cross-randomized observations of real-world care thus allows us to estimate the persistence of the PFP contract on a real effort task in the second stage of our study.

2.2 First-stage lab experiment

During the health worker survey, we randomized participants in each clinic into one of three trial arms (information, rewards, penalties), stratified by health facility. We asked participants to review records for five fictitious patients presenting for different stages of labor and delivery care, and then to identify all the clinical actions that would be required for each patient.[7] The assessment tool we used to measure performance, called a partograph provided some information about the necessary care required in each case (see Appendix B.2 for details). This design feature thus precluded the inclusion of a pure control arm in this first stage. However, we discuss in Section 3 how we attempt to estimate the effect of information over no intervention in the lab experiment. (As described above, the second stage of the study includes a randomly assigned pure control arm.)

All participants received a printed list of seven randomly selected clinical actions from the universe 15 actions relevant for the five (see Figure B.1 in Appendix B.2). In other words, all

[6] We dropped observations from one primary health facility because an error in the survey software rendered the data unusable.

[7] We designed the five cases to reflect different stages of labor and delivery care, based on examples from medical training materials. Two medical professionals independently identified actions that would be clinically correct or incorrect based on an international standard checklist of essential procedures published by the World Health Organization. They concurred about all actions in all cases.

participants were "primed" on these seven actions while the remaining eight "unprimed" actions were never primed or paid in any treatment arm. The seven selected actions included common treatment steps like monitoring the fetal heart rate or preparing for imminent delivery. In the rewards and penalties arms, five of the seven actions were further randomly selected for financial incentives. In the rewards arm, participants were offered payments if they correctly identified any of the five "paid" actions.[8] In the penalty arm, we deducted the same payment amount for each action that is correct but was not identified by the participant. The reward and penalty contracts are thus isomorphic: the same actions led to the participant being paid the same in either arm. We randomized the payment amounts across the incentivized actions. The primed and incentivized actions include some that are incorrect, i.e., at best they serve no medical purpose and may even be harmful to the patient.

We fielded two types of hypothetical patient cases to represent typical scenarios in our settings. In the two "simple" cases, we asked participants to assess whether a single action suggested by an unnamed colleague is correct or incorrect. In the three "complex" cases, the participant is in charge of the patient and is asked to name all actions that she deems to be correct for that patient. Table B.1 presents all possible actions for the three complex tasks and notes whether a given action is correct or incorrect. Which actions are correct varies across cases but is invariant across participants. We assess the proportion of clinical actions that participants correctly identified for each hypothetical patient case, as well as actions that are incorrect and not named.[9] In other words, the fully correct set of actions consists of all the correct actions and none of the incorrect ones. We then calculate the proportion of correct actions as a share of all possible actions. As the different cases have different numbers of relevant actions, we report results that equally weight each case and those that equally weight each individual item responses. Specifically, in the "across cases" measure, we calculate the proportion of correct responses separately for each of the five cases as well as the average across cases. In the "across responses" measure, we calculate the proportion correct for individual actions in all cases.

[8] While we randomly selected actions to be primed, due to the small overall number of actions, the three groups of actions (unprimed, primed, paid) could still be systematically different. We examine this issue in the balance section below.

[9] In four instances, actions can be ambiguous, i.e., they can be unnecessary or harmful to patients. Hence, we consider these actions to be incorrect. See Table B.1 for further details.

All participants received a flat payment that varied by arm such that the maximum, minimum, and expected average payouts are all the same. The flat payment in the information arm was 1,750 Nigerian Naira or about USD 5.80. The participation fee allows us to account for an endowment effect, which is important because most public sector contracts have fixed remuneration scales whereas many at-scale PFP interventions in LMICs aim to be budget neutral (Fritsche et al., 2014). The rewards arm received a base fee of 1,000 Naira (USD 3.30) and could earn an additional 1,500 Naira for a total of 2,500 Naira (USD 8.30), while the penalties group received a base pay of 2,500 Naira and could lose up to 1,500 Naira (USD 5) for a minimum payout of 1,000 Naira.[10] The incentives for individual actions in the rewards and penalties arm varied between 50 and 300 Naira (USD 0.17 to 1). Whether an action is correct—and hence the incentive is paid out—depends on the specific case. We compensated participants in the form of cellphone airtime after they had completed all five cases. Appendix B.2 presents the complete instructions provided to participants, payout schedule, screenshots of representative interview templates, and all lab assessment tools. The study was pre-registered as AEARCTR-0002482.[11]

Table 1 presents the prices and proportion of correct responses for each action, disaggregated by whether the action is paid, primed, or unprimed.[12] There is considerable variation in performance across actions as well as an overlap in the range of performance for primed and unprimed actions. We observe relatively high levels of performance for many primed and unprimed actions. For example, health workers correctly identify the need for referral to a higher-level facility (an incentivized action) 75 percent of the time and correctly recommend not administering magnesium sulfate (an unprimed action) 96 percent of the time. In contrast, participants often missed other actions that are always correct: monitoring contractions and the amniotic fluid, tracking the fetal heart rate and the mother's vital signs, and recording the fluids and drugs administered.

[10] One possible concern is that our loss framing failed to change participants' reference points by very much. However, while we did not prepay the participation fee in the penalty arm, the instructions explicitly stated that participants stood to lose part of their participation fee.

[11] The primary analysis does not deviate from the pre-analysis plan (PAP) for the lab experiment, but the secondary analysis does differ in a few dimensions from what was described in the PAP. First, we had intended to assess the size of the facility catchment area as a key moderator of responses. However, these data were largely missing, thus leaving us unable to complete this portion of the analysis. We had also intended to include the health worker's education level as a covariate but were unable to do so due to a lack of variation in this variable. In lieu of the health worker's education, we explore tenure and experience as additional dimensions along which to assess balance.

[12] Table B.2 disaggregates performance on each of these actions for the three complex cases and Figure D.2a and Figure D.2b presents the empirical cumulative distributions for the complex cases only. As these results show, our findings are robust to excluding the simple cases.

Performance is also generally better for actions that are clinically incorrect in all cases, such as administering magnesium sulfate or augmenting labor.

2.3 Health worker performance in the real world

In the second stage of our study, we assess the impact of treatment assignments in the lab experiment on related actions during real-world patient-provider interactions for antenatal care.[13] These interactions involve real effort by the worker and real stakes for the patient. Because we did not incentivize the real-world interactions, this analysis allows us to examine whether impacts of the incentives persist after the PFP contract has ended and sheds further light on the validity of our experiment. Specifically, we examine the impact on real-world performance of assignment to one of the two incentives arms versus assignment to the information arm. We only consider actions that were included in the experiment and remain relevant in the real-world task, such as monitoring the fetal condition, performing essential blood and urine tests, and documenting all care provided (see Appendix B.3 for details). This yields our estimate of the persistence of the impact of incentives relative to information alone.

Real-world behavior is recorded by enumerators trained in the direct clinical observation of antenatal care provision using a structured, quantitative checklist. The enumerators recorded whether the health worker performed actions contained in the standard WHO standard of care for antenatal visits. As detailed in Appendix B.3, this includes two actions that were also incentivized in our experiment (listening to the fetal heartbeat and palpating the abdomen), one that was only primed (recording in the patient's file all the care provided), and three actions that were unprimed but measured in the experiment (measuring the pregnant woman's blood pressure, performing a urine test, and conducting a vaginal or pelvic exam). The remaining actions recorded in the direct observation were not relevant to those measured or incentivized in our experiment. Health workers were not incentivized by our experiment for these tasks and they were not hard copies of the checklist to keep after the lab experiment.

[13] While our lab experiment assesses performance on maternity care for labor and delivery, the real-world care provided is for antenatal care. This discrepancy arises because antenatal care visits are more common and hence more feasible to observe. We did not use antenatal care for our lab experiment because there is no equivalent patient chart for antenatal care that includes both problem-solving as to the patient's care needs as well as documentation.

Given the cost and logistical challenges of collecting direct observation data, these observations were performed at a randomly chosen third of the sampled primary health centers, for a total of 230 health centers (Khanna et al., 2021). In each of these 230 health centers, two randomly selected health workers were observed while providing care to at most two patients each. Some facilities were not large enough to have two health workers providing care during the observation. Not all participants of the first-stage experiment were at work and selected for the observation of real-world care provision and, conversely, not all workers selected for observation had participated in the prior first-stage experiment. Our final sample is thus of 344 directly observed health workers who had participated in the earlier first-stage experiment and 67 additional pure control health workers who had not participated in the first-stage but worked in a health facility where at least one of their colleagues had participated.[14] The two groups also have similar characteristics (Table 2).

2.4 Summary statistics and balance

The survey captured a range of health worker characteristics that allow us to assess balance across the treatment arms of the lab experiment, as well as for the randomly selected pure control for the direct observation of real-world patient-provider interactions. This includes the workers' education, when they had last received training in labor and delivery, and their professional grade. It also included a vignette-based assessment of health workers' knowledge of the standard international checklist of essential procedures for antenatal care (Das et al., 2008; Villar et al., 2001). In this assessment, participants were read a narrative about a pregnant woman seeking antenatal care and were asked to list everything that they would do during that visit. Although such vignettes are commonly used to assess provider knowledge (Das et al., 2008), they do not capture worker effort, which is the component most likely to respond to pay-for-performance.

Panel A of Table 2 shows balance across the lab experiment and direct observation arms. There are 445 participants in the information arm and 457 in each of the incentives arms. Overall

[14] The sample size varies slightly across specifications as we exclude actions that are reported to have been performed outside of the exam room, prior to the observed interaction. Moreover, Table D.10 shows that participants in the first-stage experiment who were or were not observed providing real-world care are comparable, except that they are significantly less likely to be male and are somewhat more likely to have greater-than-median experience. Our results for the first-stage experiment are robust to only estimating impacts for the subset of participants who were observed providing care (Table D.11).

performance is low: on average in the information arm, only about half of participants' responses are correct, that is, either mentioned when correct and not mentioned when incorrect. Further, the average payouts in the rewards and penalties arms are comparable to each other, at 1,842 and 1,818 Naira respectively (about USD 6) but are significantly higher than in the information arm (1,750 Naira). The average payouts in the two incentives arms also reveal how much money participants forwent by not performing correctly: 658 Naira in the rewards arm and 682 Naira in the penalties arm. In other words, participants left about one-quarter of the maximum payout on the table by not identifying the correct actions. Hpwever, in terms of assessed covariates, all three treatment arms are balanced, suggesting successful randomization of workers into the three first-stage lab arms.

We also test for balance in the cross-randomized second stage of the study, which is the direct observation of real-world care provision. Panel B of Table 2 tests for balance within the direct observation sample, comparing the pure control observations with those who had previously participated in the first-stage lab experiment. The assessed covariates are all balanced. Further, Table D.10 (discussed in detail in Appendix D) tests for balance among lab participants selected for direct observation in the second stage compared to lab participants not selected for direct observation. This table shows that assessed covariates for workers selected for observation of real-world care provision were balanced compared to the covariates of workers not selected for observation. This table thus suggests that the cross-randomization for direct observation in the second stage was also successful.

3 Empirical strategy

We leverage the randomized first-stage and cross-randomized second-stage treatment assignments and the following OLS model to assess the impact and persistence of PFP contracts. This captures the effect on all possible clinical actions (unprimed, primed, or paid). In this section, we further describe how we estimate both direct effects and indirect effects arising from spillovers on unpaid (but primed) and unprimed actions.

$$y_{if} = \alpha + \beta \cdot Incentives_{if} + \gamma_f + \eta_{if}, \tag{1}$$

where y is the performance of participant i in facility f and the vector $Incentives$ captures whether they were randomly assigned to rewards or penalties. All specifications include facility fixed effects γ_f. α and η are the constant and error term, respectively. We report robust standard errors. We assess robustness with additional regressions that control for the participant-level covariates listed in Table 2.

We also estimate the impact of adding incentives to priming by subtracting the performance in the information arm (priming only) from the performance in the incentivized arms (which provide both incentives and priming). This yields the effects of pure incentives (i.e., incentives net of priming) on paid actions in the presence of possible spillovers. We estimate the spillover effect of incentives on unprimed and primed actions by comparing performance in the rewards and penalties arm (where spillovers from incentives could exist) to performance in the information arm, for primed and unprimed actions. This yields an estimate of the incentive spillovers relative to priming alone.

As discussed above, the experiment does not include a pure control arm which prevents us from cleanly estimating the effect of information. However, we can obtain a rough benchmark of this effect by comparing, within the information arm, performance on primed actions with performance on unprimed actions. There are two important caveats. First, the actions in these two groups may not be comparable, and therefore performance could be different in the absence of the information we provided. Second, there could be spillovers onto these "pure control" actions, for instance, if participants in the information arm shift effort toward actions on the checklist.[15] With these issues in mind, we can estimate the effect of information relative to a pure control as the difference between the primed and unprimed actions in the information arm.

Finally, to address the concern that we measure knowledge or skill rather than effort, we

[15] If there are negative spillovers from information on unprimed actions, our estimate of the effect of information would be an upper bound, while a positive spillover would lead to a lower-bound estimate.

examine whether the responses to our task are explained by the health worker's knowledge of the international standard of care for maternal care. We calculate this level of knowledge as the share of actions or screening tasks that the participants correctly identified in the above-mentioned antenatal care vignette and use a binary indicator of whether the participant scored above the sample median. We assess the impacts of these confounders by interacting our three study arms with $Confounder_{if}$ as follows:

$$y_{if} = \alpha + \beta \cdot Incentives_i + \kappa \cdot Confounder_{if} + \gamma \cdot Incentives_{if} \cdot Confounder_{if} + \eta_{if} \qquad (2)$$

4 Results

In this section, we first discuss our findings related to the direct effects of incentives on overall performance and on actions associated with incentives in the rewards and penalties arms. Then, we estimate the spillover effects on actions that are unprimed or primed, but not paid. Next, we estimate the persistence of incentives on performance in the second-stage real-world patient-provider interactions. Then, we rule out two key confounders: exposure to the larger PFP trial and the participant's clinical knowledge. We conclude with a discussion of the validity of the lab experiment and other robustness checks.

4.1 Impact of performance pay in the lab experiment

A comparison of the empirical cumulative distributions of performance in Figures D.2a and D.2b, equally weighing cases and responses, respectively, yields three findings. First, the range of observed performance is comparable across all arms. Second, the two incentive arms perform substantively and statistically better than the information arm (Kolmogorov-Smirnov tests p<0.01) largely due to a shift in the middle of the distribution. Third, the distributions of the rewards and penalty arms are not economically or statistically different.

We observe the same pattern in the regression results for average effects presented in Table 3. Performance in the information arm is 53 percent and 4.4 and 3.4 percentage points (pp) higher across cases in the rewards and penalties arms, respectively, and approximately 2 pp higher across responses. Overall, incentives increase average performance by 3.5 to 8 percent.[16] For both measures, rewards and penalties have statistically indistinguishable impacts on performance, e.g., with a p-value of 0.31 for the "across cases" measure as reported in the bottom panel of Table 3. Such a lack of difference between rewards and penalties is consistent with loss neutral agents, as discussed in Appendix A.

We next examine the effects of incentives on the subset of actions that were paid. We conduct this analysis at the level of individual actions from the complex cases. Table 4 reports the average performance on all three types of actions (unprimed, primed, and paid) by arm; the full regression results are presented in Table D.1. Focusing on paid actions, performance in the two incentive arms is 7.4 (rewards) and 7.9 (penalties) pp or 20 percent higher than for the same actions in the information arm. We do not find detectable differences between the rewards and penalties arms.

In Table 4, we estimate correct performance to be 28.2 percent on unprimed actions and 26.4 percent on the combined primed or paid actions, suggesting priming alone had no meaningful effect in our experiment. Note that if there were positive spillovers in the information arm, then the level of performance on unprimed actions would be higher than in a pure control arm and our estimate of the effect of information would be a lower bound.

Turning to incorrect care (columns 6-7 in Table 4), in the rewards arm, workers increase the provision of incorrect care by 2.8 pp (about 3.5 percent) over the information arm. The impact of penalties is not statistically significant. In all arms, performance on the incorrect actions is higher than for correct actions, i.e., participants are generally less likely to recommend an incorrect action. For example, in the information arm the proportion of correct responses is 91.5 percent for unprimed and incorrect actions as opposed to 28.2 percent for unprimed but correct actions. Compared to the estimated impact (20 percent) on paid correct actions, an impact of 3.5 percent on incorrect actions

[16] In Table D.7, we explore whether workers of different characteristics respond heterogeneously to performance pay, as in Donato et al. (2017) which participants The high performers in our first-stage lab experiment are more likely to be female and are younger than the median worker. However, we do not find cadre or experience to matter.

is small.This difference may arise from several factors. For instance, not naming a wrong action may be easier than naming a correct action. Alternatively, the two unprimed incorrect actions, not performing an unnecessary cesarean section and not referring incorrectly, may be particularly salient. The correct and incorrect actions are also substantively different.

4.2 Spillovers and price response in the lab experiment

Table 4 also summarizes spillovers from the incentives on the unprimed actions and primed actions. For unprimed actions, we find positive spillovers on correct actions: relative to 28.2 percent correct performance in the information arm, performance in the rewards arm is about 15 percent higher in the rewards arm and about 14 percent higher in the penalties arm. This effect is driven by two unprimed actions, measuring the mother's vital signs and the rate of descent of the fetal head. We find no evidence of spillovers for incorrect actions. We also do not find spillovers for primed but unincentivized actions— performance on these actions is between 10.8 and 12.5 percent across the arms— which may be because incentivizing actions increases their salience relative to the primed actions. In the context of the theoretical framework described in Appendix A, we interpret positive spillovers from incentives as consistent with complementarities across actions.

In Appendix C, we estimate that both direct and spillover effort in the lab experiment responds only concavely to price (Figure C.1 and Figure C.2). Going from zero to a small positive price captures most of the gains in performance for both rewards and penalties, suggesting that the incentive may primarily serve to amplify the prime. This result is important because it suggests that PFP contracts may be made more cost-effective by using a token price. Indeed, if one were to expect a linear response to price in any setting, it would be a lab setting where workers simply have to tell us what they would do rather than actually perform the action. It is thus especially striking that even in our context, workers "leave money on the table."

However, caveats apply to this analysis. We did not randomize prices across actions, but instead purposively assigned higher prices to more complex actions. Performance on each action thus reflects responses to both the price and non-price characteristics of the action. These estimated price responses assume that the non-price characteristics of an action are constant between the two types of arms. This may be a reasonable assumption in our setup where performance simply entails identifying the correct action rather than actually performing it.

4.3 Impact of performance pay on subsequent real-world patient interactions

Next, we assess the persistence of experimental gains on six real-world actions conducted after the lab experiment. Of the six, two are incentivized in the experiment (listen for the fetal heartbeat and palpate the abdomen), one is primed (record keeping), and three are neither paid nor primed (perform a urine test, take the pregnant woman's blood pressure, and conduct a vaginal exam). The results in Table 5 show that assignment to the incentives arms in the first-stage lab experiment significantly increases real-world performance on the two actions that are incentivized in our experiment by about 25-27 percent. In contrast to the lab experiment, in the real-world interactions, rewards but not penalties drive most of the persistent gains. We also find that information had no persistent effect over the pure control group, highlighting the incremental effect of performance pay.

For primed (but not paid) and unprimed actions, recall that the information arm is equivalent to the incentive arm in the absence of any spillovers that persist from being exposed to incentives. Ex ante, the impact is unclear for real-world actions proximate to unpaid actions in the lab experiment. On the one hand, both information and incentives received the same prime for unpaid actions, so we should not expect meaningful incremental gains over information for these actions. On the other hand, if the positive spillovers estimated on these actions persist, then the incentive arms may do better than information even after the PFP contract ends. We find that incentives and information led to equivalent improvements on unpaid actions over pure control, suggesting that the spillovers onto unpaid actions do not persist. That even priming and incentives provided in the context of a lab experiment can have persistent effects on real-world care suggests that PFP can affect worker performance even after the incentives end, and that pairing training interventions with financial incentives may lead to lasting effects.

4.4 Validity of the lab experiment and robustness

In this sub-section, we summarize several tests of the validity of the lab experiment as a realistic and meaningful measure of health worker response to the PFP contract. We also present an overview of the various robustness checks we conduct on our estimated impacts. Both the tests of validity and robustness checks are presented in detail in Appendix D.

While the lab experiment revolves around fictitious patients and does not impose actual effort costs on health workers, it has several design features that help make its setup realistic along the lines of a framed field experiment (Harrison and List, 2004). As discussed in detail in Appendix D, the incentives are real, the participants are actual health workers whose daily work—providing labor and delivery care—aligns with our experimental task, and the study was conducted in their primary workplace. We also find that overall performance on our task is comparable to non-experimental assessments of knowledge by the same health workers. Specifically, participants in the information arm have an average score of 53 percent on our task, which is similar to the scores on the knowledge vignette and typical for LMIC settings (see, e.g., Das et al., 2008).

We also perform a battery of robustness checks that are also detailed in Appendix D. The first set of checks shows that results that robust to various alternative formulations of the lab-in-the-field performance score: considering the simple and complex cases separately (Figure D.2a and Figure D.2b and columns 3–7 of Table 3), measuring performance in z-scores instead of as a proportion (Table D.5), and using an Item Response Theory aggregated outcome score (Table D.4) to account for the differences in the characteristics of the various actions required to treat each hypothetical patient. Second, we consider and rule out health worker knowledge as a potential confounder in the lab experiment (Table D.6). Third, we show that our lab results hold, albeit with some expected loss of power, for only the cross-randomized sub-sample that is observed while providing real-world care (Table D.11). Several other robustness checks are presented in the appendix, including covariate balancing and ensuring that performance in the lab experiment is not driven by any one case or only by the simple cases.

5 Discussion

PFP contracts have been implemented in advanced health systems, such as in the United States and the United Kingdom (see, e.g., Doran et al., 2011), and are also increasingly prevalent in LMICs. These contracts can elicit effort through two economic channels: information and financial incentives. In addition, incentives may elicit different levels of effort depending on the magnitude of the incentive and whether they are cast as rewards or penalties. Two related questions are whether these contracts generate spillovers on unincentivized actions, and how effort responds to magnitude of the incentive.

We report on a two-stage experiment designed to examine the effect of adding incentives to the prime; spillovers on actions that are unprimed and primed-only; price response; and whether PFP effects persist for real-world tasks after the contract has ended. The first stage is a lab experiment in which we randomized maternity care workers in Nigeria into three arms (information only, rewards, and penalties) and either primed or paid for a subset care for hypothetical patients. In the second stage, we conducted a cross-randomized assessment of care provision to actual patients after the end of the lab experiment.

We find that incentives matter above and beyond information, with incentives outperforming information in both the lab and the real world. Further, rewards and penalties generally perform similarly compared to information alone. While our lab experiment precluded the inclusion of a pure control arm—the partographs themselves could have conveyed information— the real-world task includes a pure control arm. In this setup, we show that incentives improve performance beyond information alone. In the lab experimental setup, we attempt to benchmark the effect of priming by comparing performance on unprimed and primed actions in the information arm, and similarly do not find an effect of priming alone. In addition, we find that effort increases with price, but only concavely and only while the incentives are being paid.

Finally, our design and findings speak to the role of intrinsic motivation, altruism, and reputation concerns in restricting responses to misaligned incentives. Several studies find that even though health workers perform only about half of all clinically appropriate actions, they also perform actions that are unnecessary and potentially harmful to patients (Lopez et al., forthcoming; Das et al., 2016). Evidence further shows that paying for health care provision can lead to an increase in the provision of unnecessary care (Green, 2014). However, in real-world settings, researchers typically do not know the extent to which the bundle of altruism and reputational concerns keeps health providers from mechanically responding to financial incentives, thus curtailing unnecessary actions. In our lab setting with hypothetical patients, there is no real harm from unnecessary actions, which allows us to examine responses to a small number of deliberately misaligned incentives. Had workers fully responded to the misaligned financial incentives, we would have seen a 100 percent increase in over-treatment. However, we observe only a 4 percent increase in unnecessary actions in the incentive arms relative to the information arm. This finding suggests that over-treatment may be somewhat limited in practice, perhaps via considerations like intrinsic motivation, reputational concerns, and altruism.

In the context of the conceptual framework presented in Appendix A, we interpret the direct effects as evidence that incentives are a critical component of PFP contracts, the positive spillovers as an indication that actions are complements in production, and the similar effects of the rewards and penalties as an indication that participants may be loss neutral. Our findings that performance changes most when the incentives jump from zero to a positive price and that effects persist even after the incentives run out suggest that the incentives may primarily increase the salience of the prime.

While our study design allows us to isolate various aspects of the effects of incentives, there are important caveats to the external validity of our results. Our lab experiment and the elicited responses closely mimic health worker behavior in the real world, but participant responses may not reflect costly effort or trade-offs in clinical practice. Nonetheless, performance on our task is significantly correlated with both performance in actual patient-provider interactions as well as health worker knowledge. Similarly, because the task is hypothetical, we do not capture any possible effects of altruism, which could interact with PFP-type interventions if, say, offering financial incentives erodes altruism (e.g., Lohmann et al., 2016). Third, our study examines responses to PFP among current maternity care workers and cannot speak to the effect on workforce composition. PFP can have substantial compositional effects on the teacher and health provider workforce through differential recruitment, which may lead to lower or higher performance (Deserranno, 2019; Leaver et al., 2021) and, possibly, to different responses to PFP. In practice, public health systems in LMICs often have rigid recruitment and career progression regulations that may limit effect on workforce composition, at least in the short run (Araujo and Maeda, 2013). Finally, while our PFP contract provides high-powered incentives directly to the worker, "real-world" PFP schemes are more complex and may be harder for workers to understand. Our participants were immediately and directly paid at the end of the experiment, while performance-based bonuses are typically calculated at the facility level, transferred to the facility, and only then apportioned among staff. This could dilute the effect of the incentives.

Our analysis contributes new evidence on the priming and incentive channels of PFP contracts, the role of loss framing, the persistence of effects, the price response of effort, as well as the spillovers from incomplete contracts. Disentangling these mechanisms is important for understanding contracting arrangements to resolve principal-agent problems and provide guidance on how to empirically examine the information and incentive channels (Prendergast, 1999). Taken at face value, our results imply that direct, high-powered PFP incentives would outperform health worker interventions that provide the same information but without incentives. We also show that contracts with rewards appear to generate the same performance gains as penalties and may be preferable for administrative ease and political acceptability.

A back-of-envelope calculation in Appendix E suggests that a PFP with our estimated effects is about as effective as increasing the health workforce by one qualified physician per primary health facility in Nigeria. In resource-constrained settings, improving existing contracts with the health workforce may be a more feasible intervention than increasing the size of that workforce. Moreover, small incentives may be sufficient to signal the importance of a task, and PFP designers could leverage this insight to increase the cost-effectiveness of PFP contracts.

References

E. Araujo and A. Maeda. How to recruit and retain health workers in rural and remote areas in developing countries: A guidance note. HNP Discussion Paper, World Bank, 2013.

K. J. Arrow. Uncertainty and the welfare economics of medical care. *American Economic Review*, 53(5):941–973, 1963.

N. Ashraf, O. Bandiera, and B. K. Jack. No margin, no mission? a field experiment on incentives for public service delivery. *Journal of Public Economics*, 120:1–17, 2014.

P. Basinga, P. J. Gertler, A. Binagwaho, A. L. Soucat, J. Sturdy, and C. M. Vermeersch. 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, 2011.

N. Benhassine, F. Devoto, E. Duflo, P. Dupas, and V. Pouliquen. Turning a shove into a nudge? a "labeled cash transfer" for education. *American Economic Journal: Economic Policy*, 7(3), 2015.

J. M. Brock, A. Lange, and K. L. Leonard. Giving and promising gifts: Experimental evidence on reciprocity from the field. *Journal of Health Economics*, 58:188–201, 2018.

S. Campbell, D. Reeves, E. Kontopantelis, E. Middleton, B. Sibbald, and M. Roland. Quality of primary care in England with the introduction of pay for performance. *New England Journal of Medicine*, 357(2):181–190, 2007.

M. A. Cattaneo, C. Oggenfuss, and S. C. Wolter. The more, the better? the impact of instructional time on student performance. *Education Economics*, 25(5):433–445, 2017.

P. A. Celhay, P. J. Gertler, P. Giovagnoli, and C. Vermeersch. Long-run effects of temporary incentives on medical care productivity. *American Economic Journal: Applied Economics*, 11 (3):92–127, 2019.

R. Chetty, A. Looney, and K. Kroft. Salience and taxation: Theory and evidence. *American Economic Review*, 99(4), 2009.

J. Das and J. Hammer. Which doctor? combining vignettes and item response to measure clinical competence. *Journal of Development Economics*, 78(2):348–383, 2005. ISSN 0304-3878. doi: https://doi.org/10.1016/j.jdeveco.2004.11.004. URL https://www.sciencedirect.com/science/article/pii/S0304387805000027.

J. Das, J. Hammer, and K. Leonard. The quality of medical advice in low-income countries. *Journal of Economic Perspectives*, 22(2):93–114, 2008.

J. Das, A. Holla, A. Mohpal, and K. Muralidharan. Quality and accountability in health care delivery: Audit-study evidence from primary care in India. *American Economic Review*, 106 (12):3765–3799, 2016.

J. de Quidt. Your loss is my gain: a recruitment experiment with framed incentives. *Journal of the European Economic Association*, 16(2):522–559, 2018.

D. de Walque, E. Kandpal, A. Wagstaff, J. Friedman, M. Piatti-Fünfkirchen, A. Sautmann, G. Shapira, and E. Van de Poel. *Improving effective coverage in health: do financial incentives work?* World Bank Publications, 2022.

S. DellaVigna and D. Pope. What motivates effort? evidence and expert forecasts. *Review of Economic Studies*, 85(2):1029–1069, 2018.

E. Deserranno. Financial incentives as signals: Experimental evidence from the recruitment of village promoters in Uganda. *American Economic Journal: Applied Economics*, 11(1):277–317, 2019.

K. Diaconu, J. Falconer, A. V. Verbel Facuseh, A. Fretheim, and S. Witter. Paying for performance to improve the delivery of health interventions in low- and middle-income countries. *Cochrane Database of Systematic Reviews*, (5), 2021.

K. Donato, G. Miller, M. Mohanan, Y. Truskinovsky, and M. Vera-Hernández. Personality traits and performance contracts: Evidence from a field experiment among maternity care providers in india. *American Economic Review*, 107(5):506–10, 2017.

T. Doran, E. Kontopantelis, J. M. Valderas, S. Campbell, M. Roland, C. Salisbury, and D. Reeves. Effect of financial incentives on incentivised and non-incentivised clinical activities: Longitudinal analysis of data from the UK Quality and Outcomes Framework. *British Medical Journal*, 342: d3590, 2011.

E. Duflo, R. Hanna, and S. P. Ryan. Incentives work: Getting teachers to come to school. *American Economic Review*, 102(4):1241–78, 2012.

E. Fehr and L. Goette. Do workers work more if wages are high? evidence from a randomized field experiment. *American Economic Review*, 97(1):298–317, 2007.

D. Filmer and N. Schady. Does more cash in conditional cash transfer programs always lead to larger impacts on school attendance? *Journal of Development Economics*, 96(1):150–157, 2011.

National Population Commission. Nigeria demographic and health survey 2018-final report. Technical report, 2019.

G. Fritsche, R. Soeters, and B. Meessen. *Performance-based financing toolkit*. World Bank, 2014.

R. G. Fryer, Jr., S. D. Levitt, J. List, and S. Sadoff. Enhancing the efficacy of teacher incentives through framing: A field experiment. *American Economic Journal: Economic Policy*, 14(4): 269–99, 2022.

V. Gauri, J. C. Jamison, N. Mazar, and O. Ozier. Motivating bureaucrats through social recognition: evidence from simultaneous field experiments. *Organizational Behavior and Human Decision Processes*, 163:117–131, 2021.

U. Gneezy and A. Rustichini. Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3):791–810, 2000.

J. Goldberg. Kwacha gonna do? experimental evidence about labor supply in rural malawi. *American Economic Journal: Applied Economics*, 8(1):129–49, 2016.

E. P. Green. Payment systems in the healthcare industry: An experimental study of physician incentives. *Journal of Economic Behavior Organization*, 106:367–378, 2014.

A. Gupta. Impacts of performance pay for hospitals: The readmissions reduction program. *American Economic Review*, 111(4):1241–83, 2021.

G. Harrison and J. List. Field experiments. *Journal of Economic Literature*, 42(4), 2004.

B. Holmstrom and P. Milgrom. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 7:24, 1991.

T. Hossain and J. A. List. The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167, 2012.

D. Kahneman and A. Tversky. Choices, values, and frames. *American Psychologist*, 39(4):341–350, 1984.

D. Kahneman, J. L. Knetsch, and R. H. Thaler. Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, 5(1):193–206, 1991.

E. Kandpal. Completed impact evaluations and emerging lessons from the Health Results Innovation Trust Fund learning portfolio. *Washington, DC: The World Bank*, 2016.

M. Khanna, B. P. Loevinsohn, E. Pradhan, O. Fadeyibi, K. McGee, O. Odutolu, G. B. Fritsche, E. Meribole, C. M. Vermeersch, and E. Kandpal. Decentralized facility financing versus performance-based payments in primary health care: A largescale randomized controlled trial in Nigeria. *BMC Medicine*, 19(224), 2021.

J. T. Kolstad. Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review*, 103(7):2875–2910, 2013.

M. E. Kruk, E. Larson, and N. A. Twum-Danso. Time for a quality revolution in global health. *The Lancet Global health*, 4(9):e594–e596, 2016.

V. Lavy. Expanding school resources and increasing time on task: Effects on students' academic and non-cognitive outcomes. *Journal of the European Economic Association*, 2016.

E. P. Lazear. Performance pay and productivity. *American Economic Review*, 90(5):1346–1361, December 2000.

C. Leaver, O. W. Ozier, P. M. Serneels, and A. Zeitlin. Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from rwandan primary schools. *American Economic Review*, 111(5):2213–46, 2021.

K. L. Leonard and M. C. Masatu. Professionalism and the know-do gap: Exploring intrinsic motivation among health workers in Tanzania. *Health Economics*, 19(12):1461–1477, 2010.

K. L. Leonard and M. C. Masatu. Changing health care provider performance through measurement. *Social Science & Medicine*, 181:54–65, 2017.

K. L. Leonard, M. C. Masatu, and A. Vialou. Getting doctors to do their best the roles of ability and motivation in health care quality. *Journal of Human Resources*, 42(3):682–700, 2007.

J. Lohmann, N. Houlfort, and M. De Allegri. Crowding out or no crowding out? A self-determination theory approach to health worker motivation in performance-based financing. *Social Science & Medicine*, 169:1–8, 2016.

C. Lopez, A. Sautmann, and S. Schaner. Does patient demand contribute to the overuse of prescription drugs? *American Economic Journal: Applied Economics*, forthcoming.

T. G. McGuire. Physician agency. In A. J. Culyer and J. Newhouse, editors, *Handbook of Health Economics*, volume 1A, chapter 9, page 461–536. 2000.

A. Mendelson, K. Kondo, C. Damberg, A. Low, M. Motúapuaka, M. Freeman, M. O'neil, R. Relevo, and D. Kansagara. The effects of pay-for-performance programs on health, health care use, and processes of care: A systematic review. *Annals of Internal Medicine*, 166(5):341–353, 2017.

G. Miller and K. S. Babiarz. Pay-for-performance incentives in low- and middle-income country health programs. Working Paper 18932, National Bureau of Economic Research, April 2013.

M. Mohanan, K. Donato, G. Miller, Y. Truskinovsky, and M. Vera-Hernández. Different strokes for different folks? experimental evidence on the effectiveness of input and output incentive contracts for health care providers with varying skills. *American Economic Journal: Applied Economics*, 13(4):34–69, 2021.

K. J. Mullen, R. G. Frank, and M. B. Rosenthal. 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, Mar 2010.

G. S. Oettinger. An empirical analysis of the daily labor supply of stadium vendors. *Journal of Political Economy*, 107(2):360–392, 1999.

E. Okeke and I. Abubakar. Healthcare at the beginning of life and child survival: Evidence from a cash transfer experiment in Nigeria. *Journal of Development Economics*, 143:102426, 2020.

E. N. Okeke. When a doctor falls from the sky: The impact of easing doctor supply constraints on mortality. *American Economic Review*, 113(3):585–627, March 2023.

C. Prendergast. The provision of incentives in firms. *Journal of Economic Literature*, 37(1):7–63, Mar. 1999.

S. G. Rivkin and J. C. Schiman. Instruction time, classroom quality, and academic achievement. *The Economic Journal*, 125(588):F425–F448, 2015.

J. Rothstein. Teacher quality policy when supply matters. *American Economic Review*, 105(1): 100–130, January 2015.

A. K. Rowe, D. De Savigny, C. F. Lanata, and C. G. Victora. How can we achieve and maintain high-quality performance of health workers in low-resource settings? *The Lancet*, 366(9490): 1026–1035, 2005.

K. E. Semrau, K. A. Miller, S. Lipsitz, J. Fisher-Bowman, A. Karlage, B. A. Neville, M. Krasne, J. Gass, A. Jurczak, V. Pratap Singh, S. Singh, M. Marx Delaney, L. R. Hirschhorn, B. Kodkany, V. Kumar, and A. A. Gawande. Does adherence to evidence-based practices during childbirth prevent perinatal mortality? a post-hoc analysis of 3,274 births in Uttar Pradesh, India. *BMJ Global Health*, 5(9), 2020.

S. Sexton. Automatic bill payment and salience effects: Evidence from electricity consumption. *Review of Economics and Statistics*, 97(2):229–241, 2015.

T. Sherry. A note on the comparative statics of pay-for-performance in health care. *Health Economics*, 25(5):637–644, 2016.

T. B. Sherry, S. Bauhoff, and M. Mohanan. Multitasking and heterogeneous treatment effects in pay-for-performance in health care: Evidence from Rwanda. *American Journal of Health Economics*, 3(2):192–226, 2017.

J. Villar, H. Ba'aqeel, G. Piaggio, P. Lumbiganon, J. M. Belizán, U. Farnot, Y. Al-Mazrou, G. Carroli, A. Pinol, A. Donner, et al. Who antenatal care randomised trial for the evaluation of a new model of routine antenatal care. *The Lancet*, 357(9268):1551–1564, 2001.

W. K. Viscusi and C. J. Masterman. Income elasticities and global values of a statistical life. *Journal of Benefit-Cost Analysis*, 8(2):226–250, 2017.

White Ribbon Alliance. Respectful maternity care: A Nigeria-focused health workers' training guide. Technical report, Futures Group, Health Policy Project, 2015.

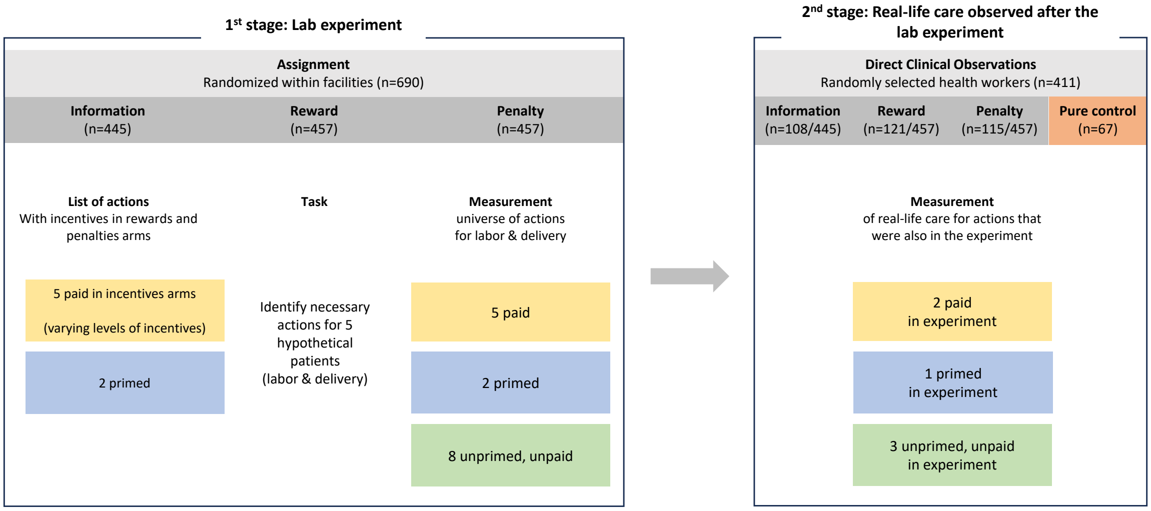
WHO. WHO recommendations for augmentation of labour. Technical report, World Health Organization, 2014.

World Bank. Life expectancy, 2019. data retrieved from World Development Indicators, https://data.worldbank.org/indicator/SP.DYN.LE00.IN?locations=NG.

World Bank. World development indicators. Technical report, 2020.

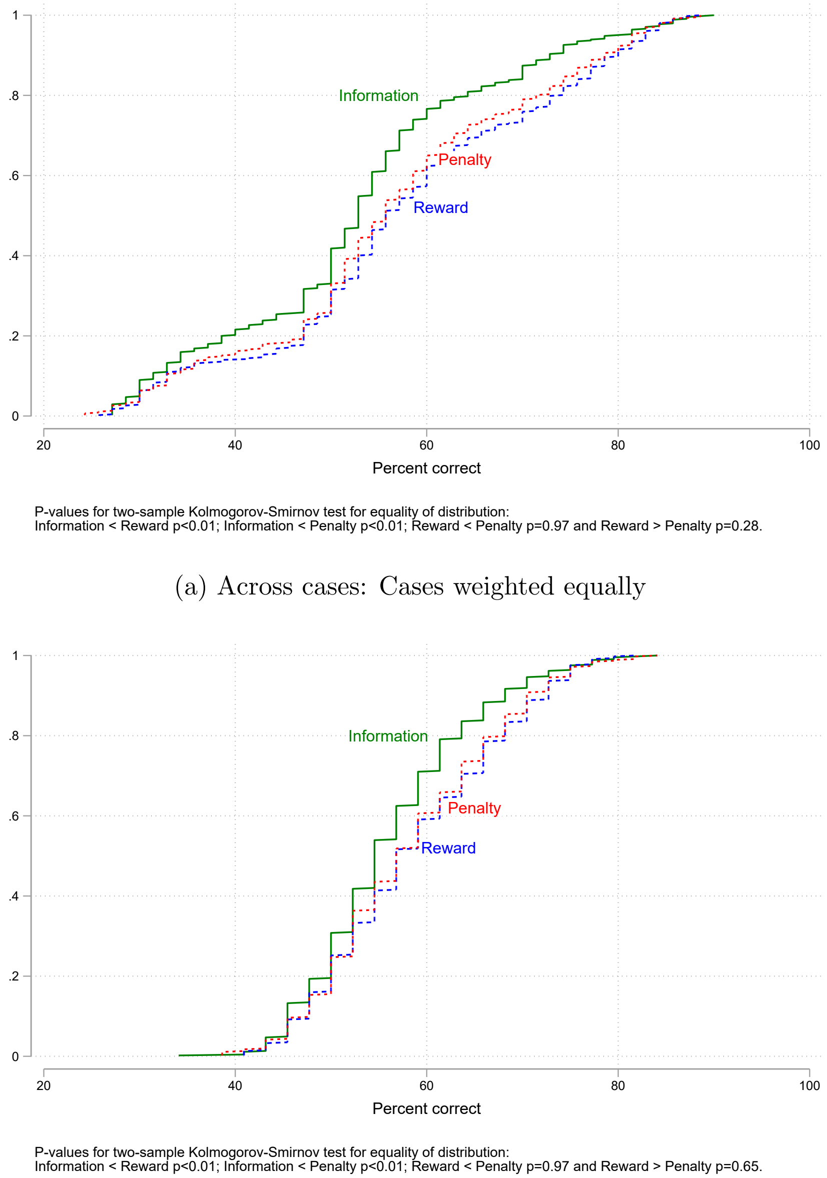
Figures and Tables

Figure 1: Study design


1st stage: Lab experiment
Assignment
Randomized within facilities (n=690)
Information (n=445)
Reward (n=457)
Penalty (n=457)
List of actions
With incentives in rewards and penalties arms
Task
Measurement
universe of actions for labor & delivery
5 paid in incentives arms
(varying levels of incentives)
2 primed
Identify necessary actions for 5 hypothetical patients (labor & delivery)
5 paid
2 primed
8 unprimed, unpaid
2nd stage: Real-life care observed after the lab experiment
Direct Clinical Observations
Randomly selected health workers (n=411)
Information (n=108/445)
Reward (n=121/457)
Penalty (n=115/457)
Pure control (n=67)
Measurement
of real-life care for actions that were also in the experiment
2 paid in experiment
1 primed in experiment
3 unprimed, unpaid in experiment

1

Figure 2: Empirical cumulative distributions



P-values for two-sample Kolmogorov-Smirnov test for equality of distribution:
Information < Reward p<0.01; Information < Penalty p<0.01; Reward < Penalty p=0.97 and Reward > Penalty p=0.28.

(a) Across cases: Cases weighted equally

P-values for two-sample Kolmogorov-Smirnov test for equality of distribution:
Information < Reward p<0.01; Information < Penalty p<0.01; Reward < Penalty p=0.97 and Reward > Penalty p=0.65.

(b) Across responses: Responses weighted equally

Table 1: Prices and performance by type of actions in the lab experiment on pay-for-performance

		Percent correct			
	Incentive (NGN†)	All arms	Information	Reward	Penalty
Paid					
Refer when necessary	300	75	71	78	76
Do not refer when unnecessary	200	70	66	74	71
Palpate the uterus	100	64	64	63	63
Monitor contractions	50	44	39	47	47
Monitor fetal heart rate	100	43	36	46	47
Primed					
Monitor color and consistency of liquor		16	15	16	18
Record fluids/drugs administered		6	6	6	7
Unprimed					
Administer magnesium sulfate		96	97	97	96
Measure urine and test for protein/glucose		94	94	95	94
Augment labor		91	91	92	91
Repeat cervical exam now		82	84	81	81
Administer antibiotics		65	65	66	65
Prepare for imminent delivery		53	53	54	51
Measure rate of descent of fetal head		48	46	48	49
Measure mother's vital signs		37	32	39	38

Based on the three complex cases out of the five hypothetical cases in the lab experiment that forms the first stage of this two-stage study. Estimated for 1359 maternity health workers in 690 primary health clinics in Nigeria.
†NGN refers to Nigerian Naira. 1 NGN equals 0.0012 USD.

Table 2: Summary statistics and balance across the study's first stage (lab experiment) and second stage (real-world care provision)

Panel A: Lab experiment in first stage

	Information	Reward	Penalty	T-test Difference		
	(1)	(2)	(3)	(1)-(2)	(1)-(3)	(2)-(3)
Outcomes†						
Across cases	52.84	57.76	56.62	-4.92***	-3.79***	1.13
	(0.69)	(0.72)	(0.71)			
Across responses	56.17	58.72	58.34	-2.55***	-2.17***	0.38
	(0.40)	(0.44)	(0.44)			
IRT weighted score§	-0.11	0.04	0.07	-0.15*	-0.18***	-0.03
	(0.04)	(0.04)	(0.05)			
Total payout	1,750.	1,841.9	1,817.9	-91.90***	-67.94***	23.96
	(0.00)	(14.70)	(15.23)			
Covariates						
Male	27.64	22.10	20.79	5.54	6.85	1.31
	(2.12)	(1.94)	(1.90)			
Older than median*	49.89	54.49	53.83	-4.60	-3.94	0.66
	(2.37)	(2.33)	(2.33)			
Doctor or nurse	7.42	10.94	10.28	-3.53	-2.87	0.66
	(1.24)	(1.46)	(1.42)			
Above median experience‡	49.89	49.89	51.86	-0.00	-1.97	-1.97
	(2.37)	(2.34)	(2.34)			
Above median knowledge◇	52.36	52.08	50.98	0.28	1.37	1.09
	(2.37)	(2.34)	(2.34)			
Number of observations	445	457	457			

Panel B: Direct clinical observation in second stage

	Control	Experiment	T-test Difference
	(1)	(2)	(1)-(2)
Male	0.19	0.14	0.06
	(0.05)	(0.02)	
Older than median*	0.57	0.58	-0.02
	(0.06)	(0.03)	
Doctor or nurse	0.04	0.04	0.00
	(0.03)	(0.01)	
Above median experience‡	0.66	0.56	0.10
	(0.06)	(0.03)	
Above median knowledge◇	0.48	0.45	0.03
	(0.06)	(0.03)	
Number of observations	67	344	

Notes:This table reports mean values and robust standard errors (in parentheses) as well as pairwise t-test differences in means across groups. Levels are reported in percent, unless otherwise noted. All specifications include facility fixed effects. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels. Panel A estimated for 1359 maternity health workers in 690 primary health clinics in Nigeria. Panel B estimated for 411 direct observations of care provision.† Across cases weighs each case equally; across responses weighs each response equally. § reports the Item Response Theory (IRT) weighted score to account for difference in task characteristics. ∗ median age is 38.0 ‡ median experience is 9.5. ◇ the median score on the assessment of knowledge of antenatal care protocol is 51.5 percent.

Table 3: Overall performance in lab experiment

	Aggregate scores		Disaggregated by case				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Cases weighted equally	Responses weighted equally	Simple 1	Simple 2	Complex 1	Complex 2	Complex 3
Reward	4.36***	2.04***	5.35*	11.21***	1.25*	1.25**	2.74***
	(0.97)	(0.45)	(3.02)	(3.08)	(0.75)	(0.51)	(0.84)
Penalty	3.43***	1.87***	2.58	9.55***	1.21	1.20**	2.61***
	(1.00)	(0.45)	(3.36)	(3.34)	(0.84)	(0.54)	(0.79)
Constant (Information)	53.15***	56.44***	68.64***	26.50***	62.33***	51.98***	56.29***
	(0.59)	(0.26)	(1.87)	(1.91)	(0.47)	(0.31)	(0.49)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-value Penalty v Reward	0.310	0.698	0.376	0.584	0.960	0.923	0.866
N respondents	1,359	1,359	1,359	1,359	1,359	1,359	1,359
R-squared (overall)	0.019	0.015	0.003	0.012	0.005	0.006	0.010

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors. Estimated for 1359 maternity health workers in 690 primary health clinics in Nigeria. The dependent variables in all specifications are expressed as percentage correct.

Table 4: Correct performance in lab experiment, disaggregated by arm and type of action

	All	Performed necessary actions				Did not perform unnecessary actions	
		Unprimed	Primed	Paid	Primed or paid	Unprimed	Paid
Level (%)							
Information	56.6	28.2	10.8	38.1	26.4	91.5	79.4
Reward	58.8	32.4	10.8	45.5	30.6	90.9	82.3
Penalty	58.6	32.1	12.5	46.0	31.7	89.8	80.6
Difference (% points)							
Reward - Information	2.2	4.2	-0.1	7.4	4.2	-0.6	2.8
p-value	0.00	0.01	0.96	0.00	0.00	0.57	0.02
Penalty - Information	2.0	3.9	1.7	7.9	5.3	-1.7	1.2
p-value	0.00	0.02	0.24	0.00	0.00	0.12	0.34
Penalty - Reward	-0.3	-0.3	1.8	0.5	1.0	-1.1	-1.6
p-value	0.64	0.85	0.23	0.78	0.49	0.31	0.18

Differences from unadjusted OLS models; s.e. clustered at worker level. The full output is reported in Table D.1. Based on the three complex cases in the lab experiment. Estimated for 1359 maternity health workers in 690 primary health clinics in Nigeria.

Table 5: Effect of PFP lab experiment on real-world effort assessed through direct clinical observations of antenatal care conducted after the lab experiment

	Related to paid actions				Related to primed actions		Related to unprimed actions					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Fetal heartbeat	Palpate abdomen	Fetal heartbeat	Palpate abdomen	Record information	Record information	Perform or refer for urine test	Blood pressure	Conduct vaginal exam	Perform or refer for urine test	Blood pressure	Conduct vaginal exam
Information	11.78	11.78			9.94*		-4.76	17.65*	-3.66			
	(11.30)	(7.97)			(5.68)		(15.51)	(9.96)	(5.01)			
Incentive	20.78**	20.78***			8.21		0.95	15.26*	-3.59			
	(10.19)	(7.70)			(5.08)		(13.30)	(8.29)	(5.84)			
Information			12.10	12.04		9.96*				-5.92	17.68*	-3.86
			(11.33)	(7.90)		(5.72)				(15.45)	(9.90)	(5.06)
Reward			24.68**	24.01***		8.36				-9.76	18.70**	-6.02
			(11.07)	(8.19)		(5.49)				(13.14)	(8.79)	(6.19)
Penalty			16.02	16.84**		8.03*				14.18	11.60	-0.63
			(10.32)	(7.42)		(4.68)				(14.46)	(8.20)	(6.47)
Constant (Naive)	65.66***	73.04***	65.76***	73.14***	89.36***	89.36***	35.70***	79.01***	10.71**	35.39***	79.06***	10.65**
	(8.50)	(6.29)	(8.56)	(6.22)	(4.36)	(4.37)	(11.27)	(6.85)	(4.59)	(11.23)	(6.79)	(4.64)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-value Info v Incentive	0.142	0.040			0.378		0.476	0.663	0.975			
P-value Penalty v Reward			0.169	0.064		0.800				0.002	0.121	0.237
P-value Info v Reward			0.078	0.020		0.477				0.649	0.857	0.456
P-value Info v Penalty			0.530	0.238		0.290				0.035	0.320	0.380
N respondents	411	410	411	410	410	410	403	353	411	403	353	411
R-squared (overall)	0.015	0.025	0.011	0.019	0.000	0.000	0.001	0.000	0.001	0.017	0.000	0.002

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors. Related to actions that are listed and paid in the experiment. Estimated for a randomly-selected subset of 411 directly observed health workers, 344 of whom had previously participated in the lab experiment and 67 of whom had not but worked in a facility where at least one other worker had participated in the lab experiment. These direct observations were conducted in a randomly-selected third of health facilities that were included in the lab experiment, i.e. 230 out of 690 primary health facilities.

APPENDICES

A Theoretical Framework

In this annex, we sketch out a framework to guide our empirical analysis of how the information and incentive channels of PFP operate and interact. We generalize DellaVigna and Pope's (2018) model by (1) considering an agent who is optimizing effort allocation between multiple clinical actions, (2) making the returns to motivation a function of information, and (3) considering cross-price effects, that is, the impact of one action's incentive on the effort allocated to another action. Spillovers in the rewards and penalties arms can be negative or positive: on the one hand, multitasking may increase effort on incentivized actions and reduce effort on unincentivized ones, while on the other hand, some actions may share common inputs or processes so that effort on one action may increase output on others. For exposition, we do not consider more complex issues, such as interactions among multiple incentivized actions (Mullen et al., 2010; Sherry, 2016).

Beginning with the rewards arm, consider the risk-neutral agent's optimization problem when facing a flat participation fee, Π_r, and two actions that are each associated with a non-pecuniary "reward," s, which is a function of information about that action, i, and a price, r, that is paid to the agent if she performs the action:

$$\max_{e_1 \geq 0, e_2 \geq 0} \Pi_r + [s(i_1) + r_1]e_1 + [s(i_2) + r_2]e_2 - c(e_1, e_2). \tag{3}$$

We assume a convex cost of effort function, $c(e)$; that is, $c'(e) > 0$ and $c''(e) > 0$ for all $e > 0$. Optimal effort e^* is then increasing in both the non-pecuniary and per-unit pecuniary rewards. First-order conditions can be written as

$$s(i_1) + r_1 - \frac{\partial c(e_1^*, e_2^*)}{(\partial e_1)} = 0, \tag{4}$$

$$s(i_2) + r_2 - \frac{\partial c(e_1^*, e_2^*)}{(\partial e_2)} = 0. \tag{5}$$

Second-order conditions can be written as

$$\frac{\partial^2 c(e_1^*, e_2^*)}{\partial e_1^2} \geq 0, \tag{6}$$

$$\frac{\partial^2 c(e_1^*, e_2^*)}{\partial e_2^2} \geq 0, \tag{7}$$

$$\left[\frac{\partial^2 c(e_1^*, e_2^*)}{\partial e_1 \partial e_2}\right]^2 - \frac{\partial^2 c(e_1^*, e_2^*)}{\partial e_1^2}\frac{\partial^2 c(e_1^*, e_2^*)}{\partial e_2^2} \leq 0. \tag{8}$$

Taking total derivatives of equations (4) and (5) with respect to r_1,

$$-\frac{\partial^2 c}{\partial e_1^2}\frac{\partial e_1^*}{\partial r_1} - \frac{\partial^2 c}{\partial e_1 \partial e_2}\frac{\partial e_2^*}{\partial r_1} + 1 = 0, \tag{9}$$

$$-\frac{\partial^2 c}{\partial e_1 \partial e_2}\frac{\partial e_1^*}{\partial r_1} - \frac{\partial^2 c}{\partial e_2^2}\frac{\partial e_2^*}{\partial r_1} = 0. \tag{10}$$

Equation (10) can be rewritten as

$$\frac{\partial e_2^*}{\partial r_1} = -\frac{\frac{\partial^2 c}{\partial e_1 \partial e_2}\frac{\partial e_1^*}{\partial r_1}}{\frac{\partial^2 c}{\partial e_2^2}}. \tag{11}$$

Plugging equation (11) into equation (9), we get the following expression for the response of optimal effort on an action to its own price:

$$\frac{\partial e_1^*}{\partial r_1} = -\frac{\frac{\partial^2 c}{\partial e_2^2}}{\left(\frac{\partial^2 c}{\partial e_1 \partial e_2}\right)^2 - \frac{\partial^2 c}{\partial e_1^2}\frac{\partial^2 c}{\partial e_2^2}}. \tag{12}$$

We know from the second-order conditions that the denominator on the right-hand side of equation (12) is negative. From our assumption of a convex cost function, $\frac{\partial^2 c}{\partial e_2^2} > 0$. Thus, we have $\frac{\partial e_1^*}{\partial r_1} > 0$, meaning that, holding information constant, providers increase effort allocated to an action in the price of that action. Plugging this into equation (11) gives us the following expression for $\frac{\partial e_2^*}{\partial r_1}$:

$$\frac{\partial e_2^*}{\partial r_1} = \frac{\frac{\partial^2 c}{\partial e_1 \partial e_2}}{\frac{\partial^2 c}{\partial e_1 \partial e_2}^2 - \frac{\partial^2 c}{\partial e_1^2}\frac{\partial^2 c}{\partial e_2^2}}. \tag{13}$$

We know from equation (10) that the denominator is negative, so if actions are complements, the sign of $\frac{\partial^2 c}{\partial e_1 \partial e_2}$ is negative and we have $\frac{\partial e_2^*}{\partial r_1} > 0$. On the other hand, if actions are substitutes, $\frac{\partial^2 c}{\partial e_1 \partial e_2}$ is positive and we have $\frac{\partial e_2^*}{\partial r_1} < 0$. Intuitively, if actions are completely unrelated, effort on an action is independent of the price of other actions.

In the penalty arm, the provider's optimization problem with a flat participation fee, Π_p, two actions, and penalties, p, is

$$\max_{e_1 \geq 0, e_2 \geq 0} \Pi_p + s(i_1)e_1 - \lambda(\bar{e_1} - e_1)p_1 + s(i_2)e_2 - \lambda(\bar{e_2} - e_2)p_2 - c(e_1, e_2), \tag{14}$$

where λ is a parameter of loss aversion such that a loss-averse individual has $\lambda > 1$, while a loss-neutral individual has $\lambda = 1$. The first- and second-order conditions are analogous to those for a positive price for effort. As DellaVigna and Pope (2018) note, actually estimating the loss aversion parameter requires a third treatment (gain or loss) condition.

Solving for own- and cross-price elasticities of effort yields the following expressions:

$$\frac{\partial e_1^*}{\partial p_1} = -\lambda\frac{\frac{\partial^2 c}{\partial e_2^2}}{\left(\frac{\partial^2 c}{\partial e_1 \partial e_2}\right)^2 - \frac{\partial^2 c}{\partial e_1^2}\frac{\partial^2 c}{\partial e_2^2}}, \tag{15}$$

and

$$\frac{\partial e_2^*}{\partial p_1} = \lambda\frac{\frac{\partial^2 c}{\partial e_1 \partial e_2}}{\left(\frac{\partial^2 c}{(\partial e_1 \partial e_2)}\right)^2 - \frac{\partial^2 c}{\partial e_1^2}\frac{\partial^2 c}{\partial e_2^2}}. \tag{16}$$

For a loss-neutral person, equations (15) and (16) for penalties are identical to equations (12) and (13) for rewards. Thus, in the absence of loss aversion, workers choose the same optimal levels of effort in response to a reward or an equivalent penalty. In contrast, loss aversion would imply that a given increase in the penalty on action 1 leads to an increase in effort on action 2 when the actions are complements and a decrease in action 2 when they are substitutes.

In sum, our model predicts that, holding information constant, incentives should increase effort on the incentivized actions and the degree of complementarity between actions determines the sign of any spillovers between actions. In particular, incentives on one action will raise effort on complementary actions (positive spillover) but decrease efforts on actions that are substitutes (negative spillover). Finally, loss-averse agents are more responsive to a penalty than an equivalent reward. The direction and degree of complementarity between actions and the degree of loss aversion are empirical questions that we examine below.

B Study Details

This appendix describes the context in which it was conducted and provides details about the lab and real world assessments.

B.1 Health financing trial in Nigeria

The health financing trial randomized all 52 districts in three states to two arms. A total of 1,389 public primary and secondary care facilities were either assigned (1) to PFP with quarterly bonuses based on the quantity and quality of primary health services they provided or (2) to direct facility financing (DFF) that disbursed half of the average PFP bonus without conditioning the payment on performance. In both arms, district supervisors administered a checklist to assess quality of care on a quarterly basis, and an independent agency verified performance in the PFP facilities. A "business as usual" control group was established by selecting three observables-matched states from the same geopolitical zone. (See Khanna et al. (2021) for details on the selection of control states.) Figure B.1 shows the intervention states (Adamawa, Nasarawa, Ondo) and the control states (Taraba, Benue, Ogun) as well as the locations of the health facilities in our study.

The PFP trial started in July 2014, when the two financing interventions were rolled out. The evaluation endline survey that contained our experiment was conducted between August and October 2017. The impact evaluation found that districts with PFP or DFF performed better than those in the control group and the impacts of PFP and DFF were comparable, with few exceptions (Khanna et al., 2021). For example, both arms significantly increased fully immunized child coverage and modern contraceptive prevalence. However, clinical quality of care, which may be most directly related to provider effort, showed limited gains. For the impact evaluation of the concurrent PFP trial, one primary or secondary health facility was randomly chosen per ward in each of the districts, for a sample of 786 facilities of the 1,389 facilities participating in the trial. Our lab experiment was conducted in all 690 primary health clinics in this sample of 786 primary and secondary facilities, while the direct observation was conducted in a randomly selected third (n=230) of these 690 primary health clinics.

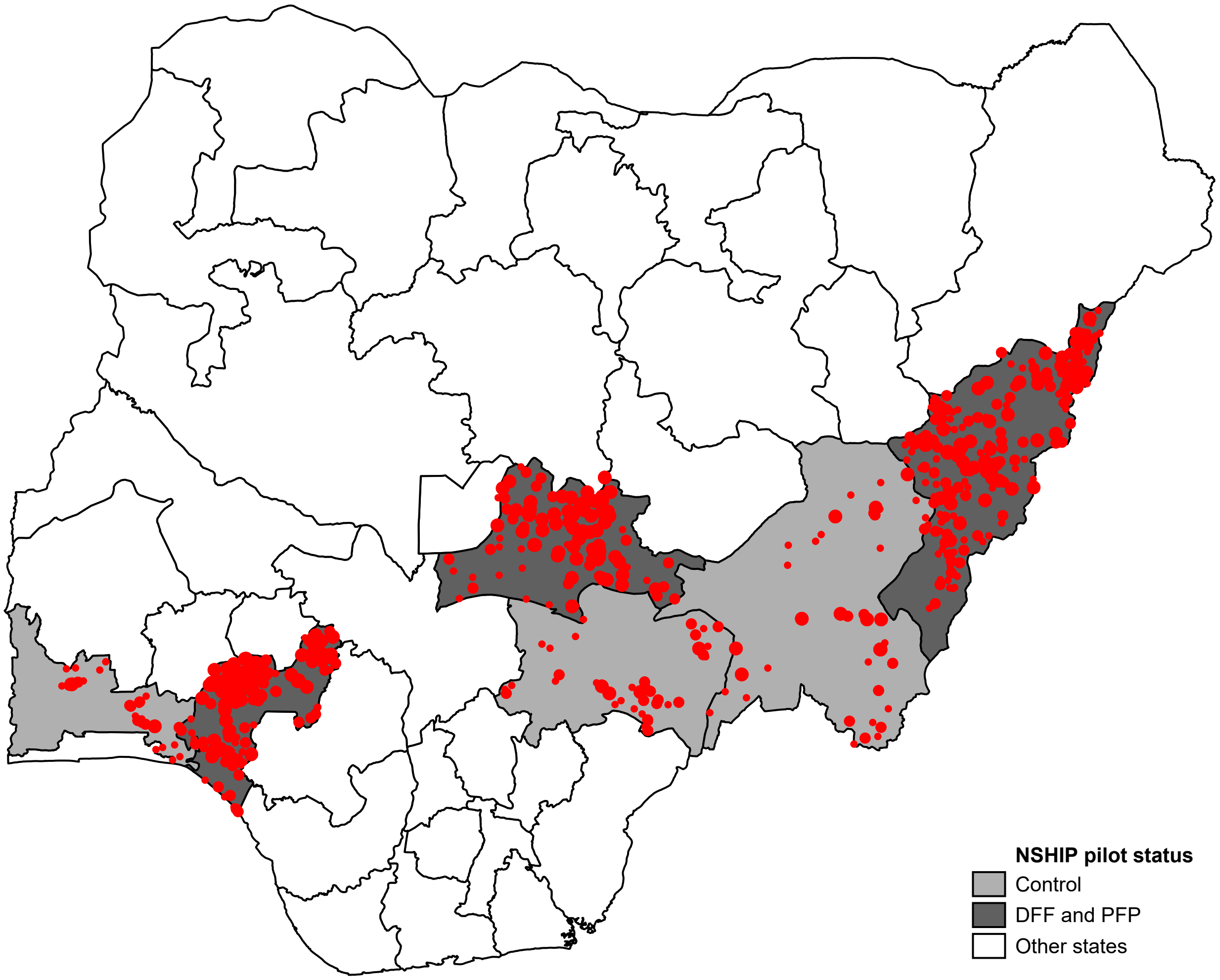
One concern about this concurrent PFP trial may be that participants in the larger PFP program may have been comparatively more attuned in responding to incentives (Leaver et al., 2021). We assess the robustness of our treatment impacts to the inclusion of participation in the arms assigned in the concurrent cluster-randomized trial: control, PFP, or DFF. Table B.1 presents interacted regressions of treatment assignment in our experiment with the larger trial's arms, control, DFF and PFP. Our estimated impacts of incentives are robust to the inclusion of assignment to the PFP trial and interaction terms. Participants assigned to our rewards and penalties arms always perform better than those in the information arm. As in the larger impact evaluation, participants in the DFF or PFP arms perform better than those in the matched control arm. The lowest-performing group are participants in the control arm of the larger trial who were assigned to the information arm in our study.

Further, Khanna et al. (2021) show that awareness of the PFP program was a significant mediator of the its effectiveness.[17] We therefore also examine the effect of a binary measure

[17] However, even in the health facilities assigned to either the PFP or DFF trial arms, a majority of health workers

of participants' self-reported awareness that their clinic is participating in the concurrent PFP program on their performance in the experimental task. We create a binary measure for awareness using responses to a survey question of whether the participants' health facility participates in the trial, and a second a binary measure of (above-median) understanding based on questions about how many indicators are incentivized in the larger trial's PFP arm. Results are presented in Table B.2. Awareness and understanding are associated with higher performance (as measured across responses), but this effect does not covary with our study arms. Thus, neither facility-level participation in the larger trial nor worker-level awareness and understanding has a significant moderating effect on performance on the hypothetical patient task. Nonetheless, the robustness of the main estimated impacts suggests that prior exposure to PFP does not drive the responses to our task. Replicating the qualitative findings from the larger trial bolsters our confidence in our measure of performance.

Figure B.1: Location of study clinics



The size of the marker is proportional to the number of participants (range 1–4). Trial status refers to the concurrent cluster-randomized trial. The intervention states are Adamawa, Nasarawa, and Ondo; the control states are Taraba, Benue, and Ogun.

had either not heard of the trial or did not understand its structure.

Table B.1: Participation in the concurrent national PFP pilot as a potential confounder of performance in the lab-in-the-field experiment

	All cases		By case				
	(1) Across cases	(2) Across responses	(3) Simple 1	(4) Simple 2	(5) Complex 1	(6) Complex 2	(7) Complex 3
Reward	9.82***	2.73**	15.50*	28.15***	3.05	-0.89	3.29*
	(2.71)	(1.18)	(8.87)	(8.78)	(1.86)	(1.39)	(1.93)
Penalty	4.34*	1.40	4.67	13.97*	2.43	-0.26	0.88
	(2.54)	(1.13)	(8.86)	(8.29)	(1.80)	(1.33)	(1.77)
DFF	5.04**	3.62***	15.34**	-0.42	7.82***	-1.85	4.33***
	(2.05)	(0.99)	(7.28)	(6.51)	(1.63)	(1.15)	(1.59)
PFP	5.94***	4.99***	13.26*	1.85	8.86***	0.50	5.24***
	(2.10)	(1.00)	(7.27)	(6.51)	(1.61)	(1.13)	(1.64)
Reward × DFF	-5.96*	-0.55	-12.50	-17.71*	-2.72	3.20**	-0.06
	(3.06)	(1.48)	(9.95)	(9.91)	(2.31)	(1.61)	(2.33)
Reward × PFP	-5.27*	0.14	-9.30	-19.52**	0.08	2.44	-0.03
	(3.15)	(1.52)	(9.96)	(9.95)	(2.29)	(1.63)	(2.44)
Penalty × DFF	-1.23	0.66	-4.88	-3.94	-1.01	1.90	1.81
	(2.91)	(1.44)	(10.00)	(9.51)	(2.29)	(1.58)	(2.20)
Penalty × PFP	0.12	1.31	0.88	-4.64	0.20	1.72	2.47
	(2.99)	(1.48)	(9.96)	(9.53)	(2.28)	(1.58)	(2.31)
Constant (Information)	48.06***	52.43***	55.93***	25.42***	54.60***	52.42***	51.94***
	(1.80)	(0.77)	(6.48)	(5.69)	(1.27)	(0.98)	(1.29)
Facility fixed effects	No	No	No	No	No	No	No
P-values from tests of coefficients							
Control: Penalty v Reward	0.043	0.277	0.206	0.116	0.740	0.634	0.197
DFF: Penalty v Reward	0.124	0.425	0.430	0.177	0.456	0.407	0.411
PFP: Penalty v Reward	0.089	0.459	0.290	0.147	0.960	0.655	0.301
N respondents	1,359	1,359	1,359	1,359	1,359	1,359	1,359

* p<0.10, ** p<0.05, *** p<0.01. OLS models with robust standard errors.

Table B.2: Interaction with awareness that facility participates in NSHIP pilot and understanding of PBF program (percentage points)

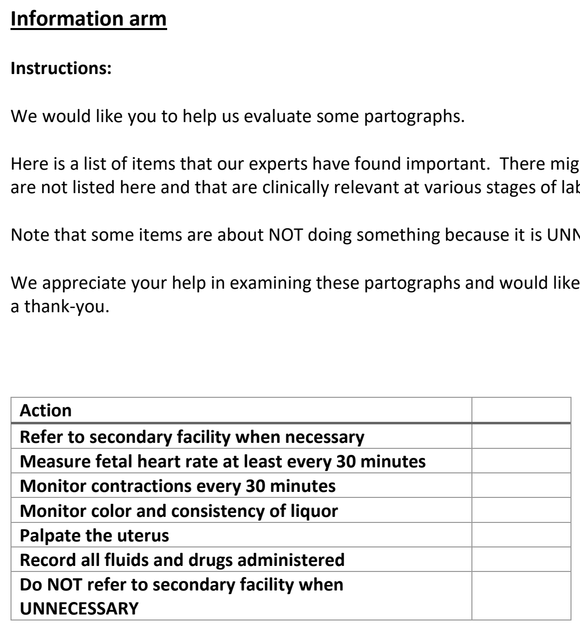
	Awareness		Understanding		Understanding if aware and part of PBF	
	(1)	(2)	(3)	(4)	(5)	(6)
	Across cases	Across responses	Across cases	Across responses	Across cases	Across responses
Reward	2.96	1.81*	2.94**	2.15**	2.08	2.20
	(1.92)	(1.06)	(1.49)	(0.88)	(4.30)	(2.28)
Penalty	2.75	1.43	3.29**	2.65***	4.41	3.75
	(1.98)	(1.05)	(1.45)	(0.88)	(3.96)	(2.41)
Aware that facility is part of NSHIP	0.03	1.43*				
	(1.58)	(0.87)				
Reward × Aware	1.73	0.99				
	(2.31)	(1.33)				
Penalty × Aware	1.50	1.33				
	(2.34)	(1.32)				
High understanding			0.36	2.14**	1.09	3.35**
			(1.48)	(0.88)	(2.75)	(1.67)
Reward × High understanding			2.47	0.68	2.64	0.36
			(2.13)	(1.29)	(4.66)	(2.53)
Penalty × High understanding			1.08	-0.50	-0.35	-1.95
			(2.13)	(1.29)	(4.35)	(2.65)
Constant (%)	55.39***	56.76***	55.27***	56.84***	55.76***	56.75***
	(1.49)	(0.78)	(1.17)	(0.70)	(2.63)	(1.55)
State FE	Yes	Yes	Yes	Yes	Yes	Yes
NSHIP arms	DFF PFP	DFF PFP	DFF PFP	DFF PFP	PFP	PFP
N respondents	1,178	1,178	1,178	1,178	527	527
R-squared	0.029	0.041	0.031	0.044	0.043	0.057

* p<0.10, ** p<0.05, *** p<0.01. Unadjusted OLS models with robust standard errors. High understanding defined as above-median number of correctly named health care services that are incentivized in PBF intervention (median = 6).

B.2 Lab instructions and partograph cases

This subsection provides details on the structure and scoring of the lab experiment. We present below the checklists provided to each of the three lab arms. Next, we show the scoring scheme and payout matrix for each of the cases. Finally, we provide illustrative examples of the tool we used in the study. For the experiment, we created five clinical tracking tools for maternity patients. The client tool—called a partograph—is a standard tool that is "strongly recommended" by the World Health Organization (WHO) and used by Nigeria's Federal Ministry of Health to train health workers providing maternal health services (WHO, 2014; White Ribbon Alliance, 2015).

Figure B.1: List provided to participants in "Information" arm



Information arm

Instructions:

We would like you to help us evaluate some partographs.

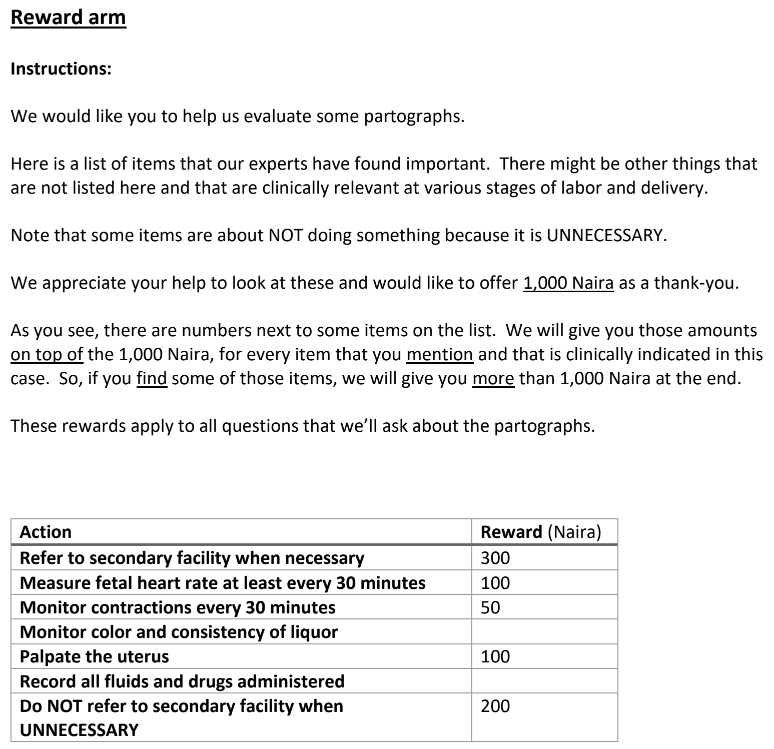
Here is a list of items that our experts have found important. There might be other things that are not listed here and that are clinically relevant at various stages of labor and delivery.

Note that some items are about NOT doing something because it is UNNECESSARY.

We appreciate your help in examining these partographs and would like to offer 1,750 Naira as a thank-you.

Action	
Refer to secondary facility when necessary	
Measure fetal heart rate at least every 30 minutes	
Monitor contractions every 30 minutes	
Monitor color and consistency of liquor	
Palpate the uterus	
Record all fluids and drugs administered	
Do NOT refer to secondary facility when UNNECESSARY	

Figure B.2: List provided to participants in “Rewards” arm



Reward arm

Instructions:

We would like you to help us evaluate some partographs.

Here is a list of items that our experts have found important. There might be other things that are not listed here and that are clinically relevant at various stages of labor and delivery.

Note that some items are about NOT doing something because it is UNNECESSARY.

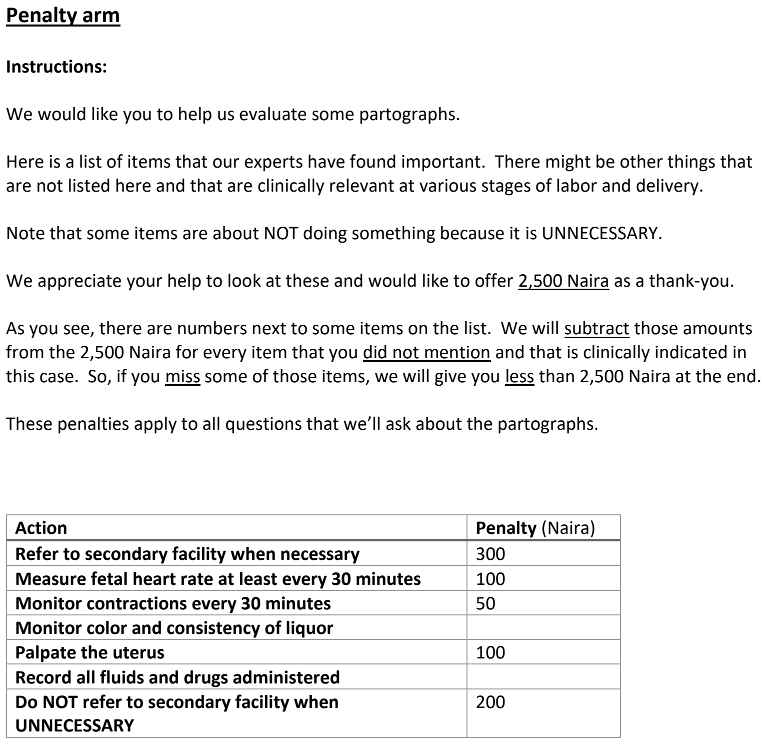
We appreciate your help to look at these and would like to offer 1,000 Naira as a thank-you.

As you see, there are numbers next to some items on the list. We will give you those amounts on top of the 1,000 Naira, for every item that you mention and that is clinically indicated in this case. So, if you find some of those items, we will give you more than 1,000 Naira at the end.

These rewards apply to all questions that we’ll ask about the partographs.

Action	**Reward** (Naira)
Refer to secondary facility when necessary	300
Measure fetal heart rate at least every 30 minutes	100
Monitor contractions every 30 minutes	50
Monitor color and consistency of liquor	
Palpate the uterus	100
Record all fluids and drugs administered	
Do NOT refer to secondary facility when UNNECESSARY	200

Figure B.3: List provided to participants in “Penalty” arm



Penalty arm

Instructions:

We would like you to help us evaluate some partographs.

Here is a list of items that our experts have found important. There might be other things that are not listed here and that are clinically relevant at various stages of labor and delivery.

Note that some items are about NOT doing something because it is UNNECESSARY.

We appreciate your help to look at these and would like to offer 2,500 Naira as a thank-you.

As you see, there are numbers next to some items on the list. We will subtract those amounts from the 2,500 Naira for every item that you did not mention and that is clinically indicated in this case. So, if you miss some of those items, we will give you less than 2,500 Naira at the end.

These penalties apply to all questions that we’ll ask about the partographs.

Action	**Penalty** (Naira)
Refer to secondary facility when necessary	300
Measure fetal heart rate at least every 30 minutes	100
Monitor contractions every 30 minutes	50
Monitor color and consistency of liquor	
Palpate the uterus	100
Record all fluids and drugs administered	
Do NOT refer to secondary facility when UNNECESSARY	200

Table B.1: Scoring scheme for each possible action in the complex tasks

	Complex 1	Complex 2	Complex 3
Paid			
Refer when necessary	Unnecessary	Indicated	Unnecessary
Do not refer when unnecessary	Indicated	Unnecessary	Indicated
Palpate the uterus	Ambiguous	Indicated	Ambiguous
Monitor contractions	Indicated	Indicated	Indicated
Monitor fetal heart rate	Indicated	Indicated	Indicated
Primed			
Monitor color and consistency of liquor	Indicated	Indicated	Indicated
Record fluids/drugs administered	Indicated	Indicated	Indicated
Unprimed			
Administer magnesium sulfate	Unnecessary	Unnecessary	Unnecessary
Measure urine and test for protein/glucose	Unnecessary	Unnecessary	Ambiguous
Augment labor	Unnecessary	Unnecessary	Unnecessary
Repeat cervical exam now	Unnecessary	Unnecessary	Unnecessary
Administer antibiotics	Unnecessary	Indicated	Unnecessary
Prepare for imminent delivery	Indicated	Unnecessary	Indicated
Measure rate of descent of fetal head	Indicated	Ambiguous	Indicated
Measure mother's vital signs	Indicated	Indicated	Indicated

Table B.2: Percent correct by action and case

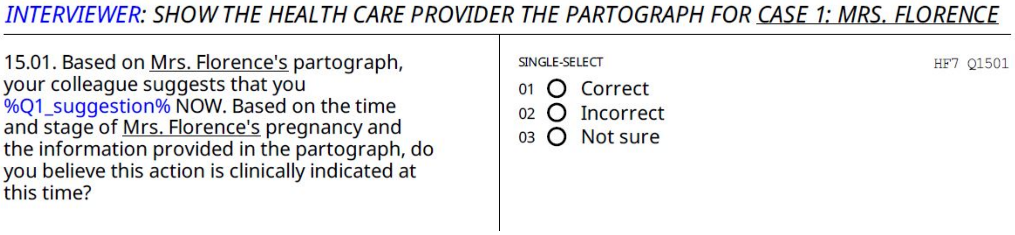
	Overall	Complex 1	Complex 2	Complex 3
Paid				
Refer when necessary	75		75	
Do not refer when unnecessary	70	79		61
Palpate the uterus	64	91	8	92
Monitor contractions	44	58	32	43
Monitor fetal heart rate	43	53	35	42
Primed				
Monitor color and consistency of liquor	16	21	13	16
Record fluids/drugs administered	6	7	6	6
Unprimed				
Administer magnesium sulfate	96	96	96	97
Measure urine and test for protein/glucose	94	93	95	94
Augment labor	91	93	94	87
Repeat cervical exam now	82	78	86	84
Administer antibiotics	65	96	4	96
Prepare for imminent delivery	53	35	90	33
Measure rate of descent of fetal head	48	38	77	28
Measure mother's vital signs	37	45	31	34

Correct captures actions that participants named and are clinically indicated as well as actions that were not named and are unnecessary.

Table B.3: Payout matrix

Action	Reward in Naira	Simple 1	Complex 1	Complex 2	Simple 1	Complex 3	TOTAL
Refer to secondary facility when necessary	300			300			
Measure fetal heart rate at least every 30 minutes	100		100	100		100	
Monitor contractions every 30 minutes	50	50	50	50		50	
Monitor color and consistency of liquor							
Palpate the uterus	100			100			
Record all fluids and drugs administered							
Do NOT refer to secondary facility when UNNECESSARY	200		200		200	200	
Potential max reward / penalty from activities	0	50	350	550	200	350	1,500
Reward arm: Total min payout (only participation incentive)							1,000
Reward arm: Total max payout (participation + activity reward)							2,500
Penalty arm: Total max payout (only participation incentive)							2,500
Penalty arm: Total min payout (participation – activity penalty)							1,000

Figure B.4: Example of a simple case: Assessing whether a suggested action is correct

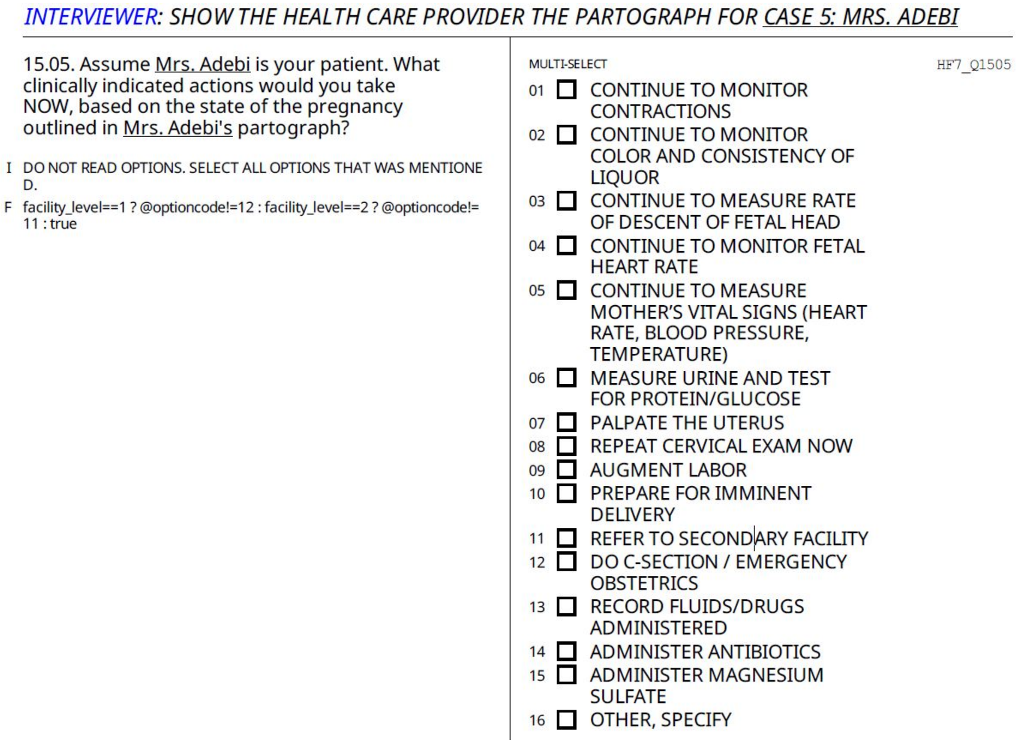


INTERVIEWER*: *SHOW THE HEALTH CARE PROVIDER THE PARTOGRAPH FOR CASE 1: MRS. FLORENCE

15.01. Based on Mrs. Florence's partograph, your colleague suggests that you %Q1_suggestion% NOW. Based on the time and stage of Mrs. Florence's pregnancy and the information provided in the partograph, do you believe this action is clinically indicated at this time?	SINGLE-SELECT HF7 Q1501 01 ○ Correct 02 ○ Incorrect 03 ○ Not sure

Screenshot of CAPI tool for case 1. The action in Q1_suggestion is randomized (within arm) to be (a) "monitor contractions" or (b) "refer to higher level." "Monitor contraction" is correct, while referring is incorrect.

Figure B.5: Example of a complex case: Stating correct action(s) to be taken



INTERVIEWER: SHOW THE HEALTH CARE PROVIDER THE PARTOGRAPH FOR CASE 5: MRS. ADEBI

15.05. Assume Mrs. Adebi is your patient. What clinically indicated actions would you take NOW, based on the state of the pregnancy outlined in Mrs. Adebi's partograph?

I DO NOT READ OPTIONS. SELECT ALL OPTIONS THAT WAS MENTIONED.

F facility_level==1 ? @optioncode!=12 : facility_level==2 ? @optioncode!=11 : true

MULTI-SELECT HF7_Q1505

01 □ CONTINUE TO MONITOR CONTRACTIONS

02 □ CONTINUE TO MONITOR COLOR AND CONSISTENCY OF LIQUOR

03 □ CONTINUE TO MEASURE RATE OF DESCENT OF FETAL HEAD

04 □ CONTINUE TO MONITOR FETAL HEART RATE

05 □ CONTINUE TO MEASURE MOTHER'S VITAL SIGNS (HEART RATE, BLOOD PRESSURE, TEMPERATURE)

06 □ MEASURE URINE AND TEST FOR PROTEIN/GLUCOSE

07 □ PALPATE THE UTERUS

08 □ REPEAT CERVICAL EXAM NOW

09 □ AUGMENT LABOR

10 □ PREPARE FOR IMMINENT DELIVERY

11 □ REFER TO SECONDARY FACILITY

12 □ DO C-SECTION / EMERGENCY OBSTETRICS

13 □ RECORD FLUIDS/DRUGS ADMINISTERED

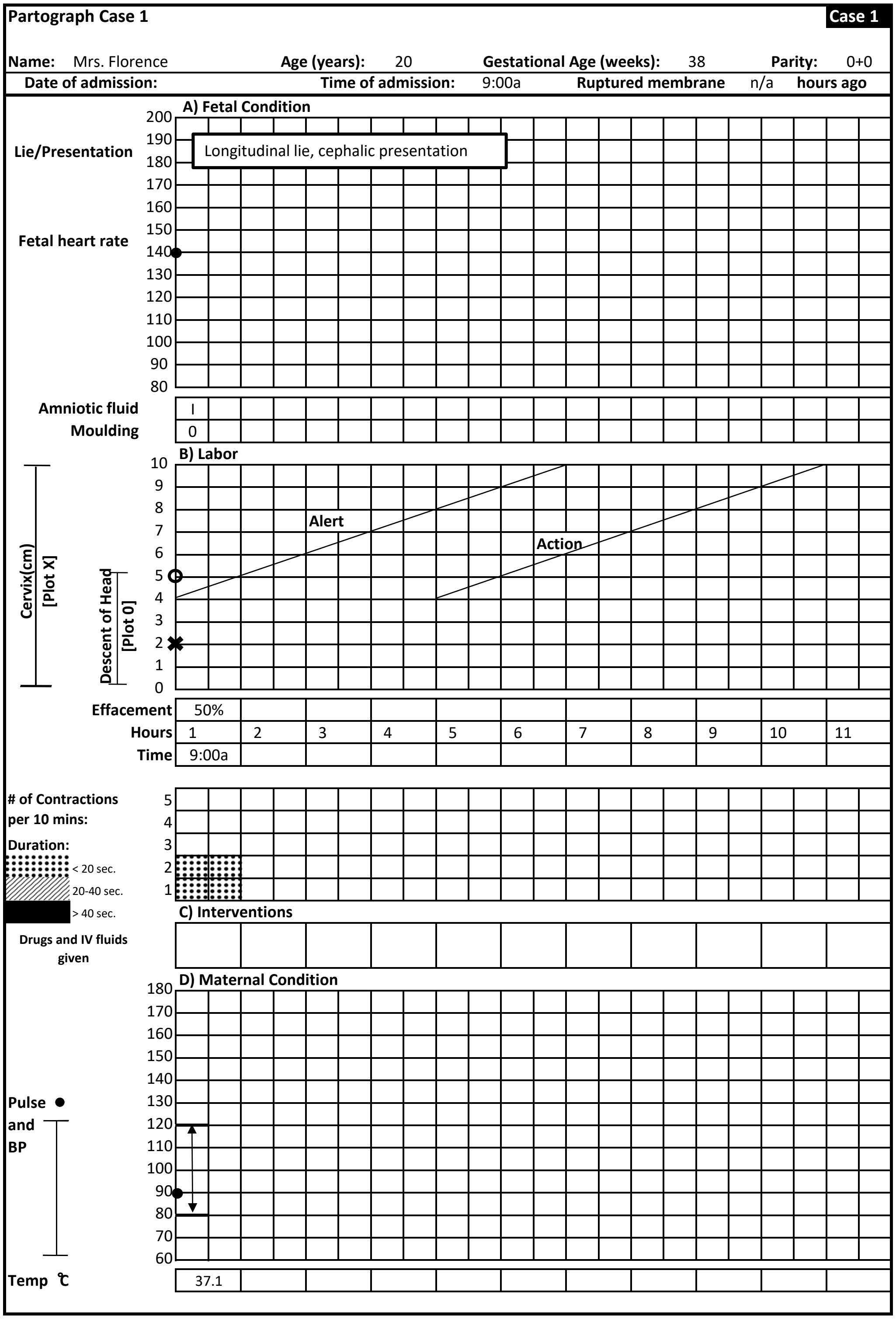
14 □ ADMINISTER ANTIBIOTICS

15 □ ADMINISTER MAGNESIUM SULFATE

16 □ OTHER, SPECIFY

Screenshot of CAPI tool for case 5.

Figure B.6: Partograph case 1



Partograph Case 1 **Case 1**

Name: Mrs. Florence **Age (years):** 20 **Gestational Age (weeks):** 38 **Parity:** 0+0

Date of admission: **Time of admission:** 9:00a **Ruptured membrane** n/a **hours ago**

A) Fetal Condition

Lie/Presentation Longitudinal lie, cephalic presentation

Fetal heart rate

Amniotic fluid I

Moulding 0

B) Labor

Effacement	50%										
Hours	1	2	3	4	5	6	7	8	9	10	11
Time	9:00a										

# of Contractions per 10 mins:

Duration:

▦ < 20 sec.

▨ 20-40 sec.

■ > 40 sec.

C) Interventions

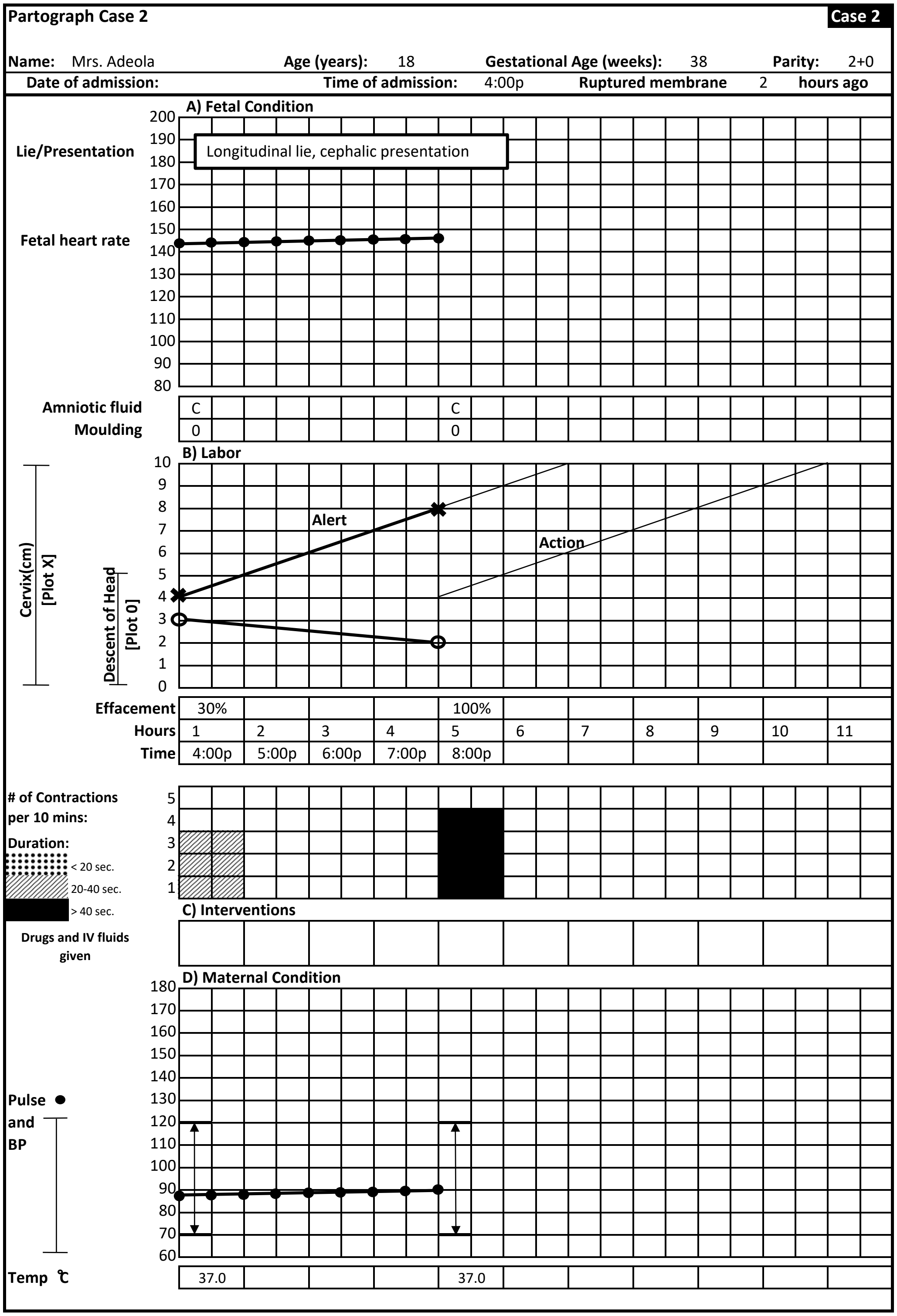
Drugs and IV fluids given

D) Maternal Condition

Pulse ● and BP

Temp ℃ 37.1

Figure B.7: Partograph case 2



Partograph Case 2 **Case 2**

Name: Mrs. Adeola **Age (years):** 18 **Gestational Age (weeks):** 38 **Parity:** 2+0

Date of admission: **Time of admission:** 4:00p **Ruptured membrane** 2 **hours ago**

A) Fetal Condition

Lie/Presentation Longitudinal lie, cephalic presentation

Amniotic fluid	C								C													
Moulding	0								0													

B) Labor

Effacement	30%				100%						
Hours	1	2	3	4	5	6	7	8	9	10	11
Time	4:00p	5:00p	6:00p	7:00p	8:00p						

# of Contractions per 10 mins:

Duration:
- ⠿ < 20 sec.
- ▨ 20-40 sec.
- ■ > 40 sec.

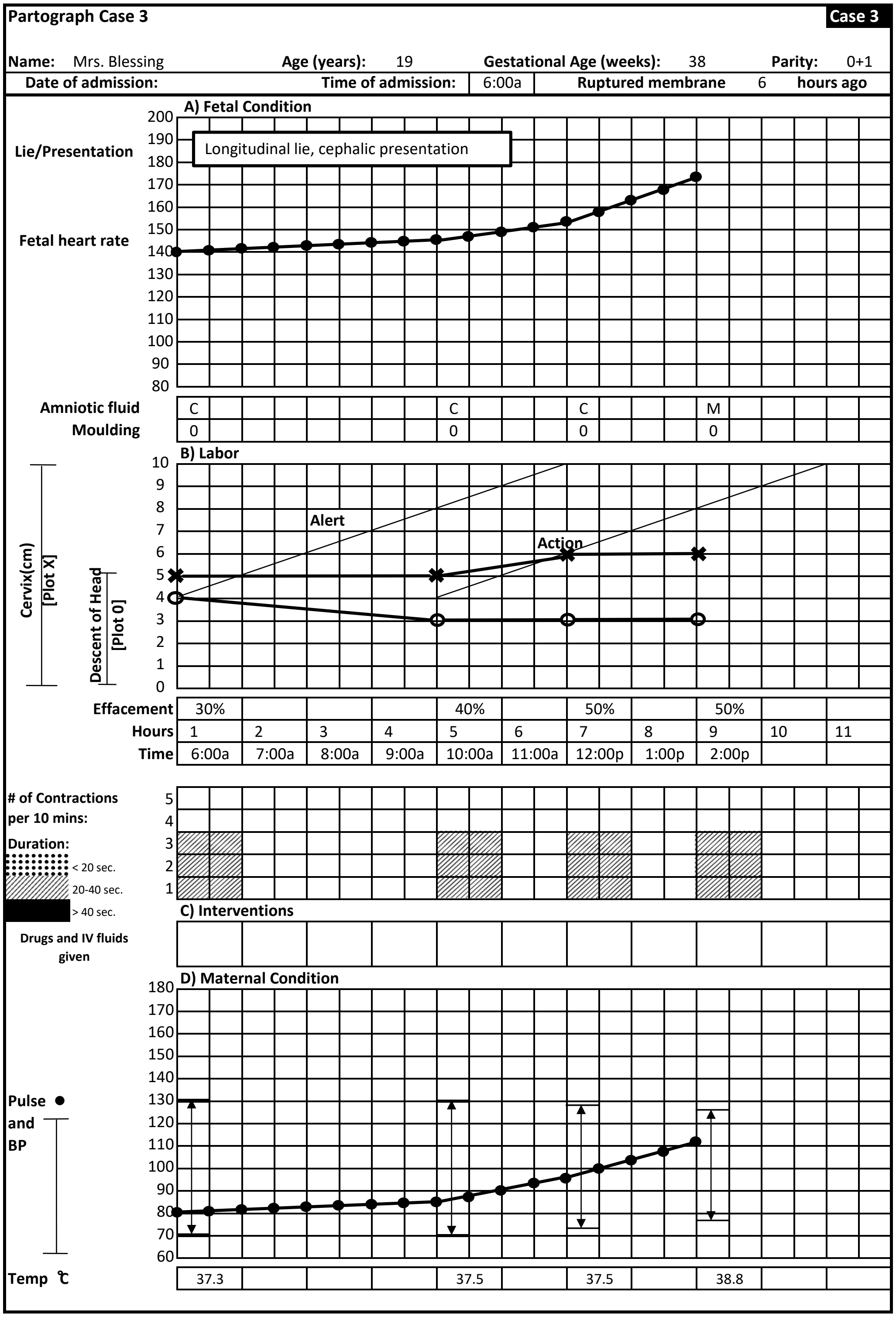
C) Interventions

Drugs and IV fluids given

D) Maternal Condition

Temp ℃	37.0				37.0						

Figure B.8: Partograph case 3



Partograph Case 3 — **Case 3**

Name: Mrs. Blessing **Age (years):** 19 **Gestational Age (weeks):** 38 **Parity:** 0+1

Date of admission: **Time of admission:** 6:00a **Ruptured membrane** 6 **hours ago**

A) Fetal Condition

Lie/Presentation: Longitudinal lie, cephalic presentation

Amniotic fluid	C				C		C		M		
Moulding	0				0		0		0		
Effacement	30%				40%		50%		50%		
Hours	1	2	3	4	5	6	7	8	9	10	11
Time	6:00a	7:00a	8:00a	9:00a	10:00a	11:00a	12:00p	1:00p	2:00p		
Temp ℃	37.3				37.5		37.5		38.8		

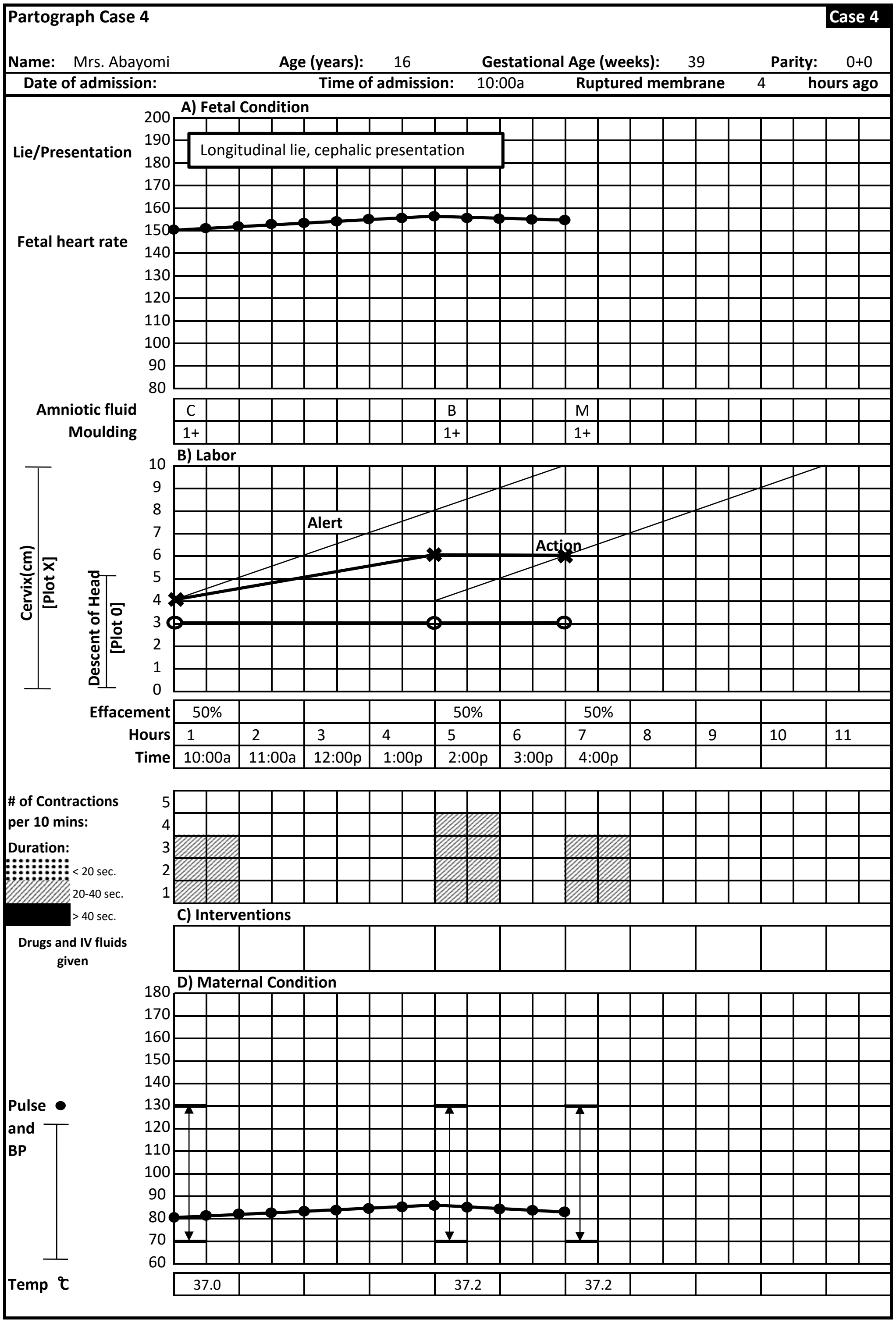
B) Labor

C) Interventions

Drugs and IV fluids given

D) Maternal Condition

Figure B.9: Partograph case 4



Partograph Case 4 **Case 4**

Name: Mrs. Abayomi **Age (years):** 16 **Gestational Age (weeks):** 39 **Parity:** 0+0

Date of admission: **Time of admission:** 10:00a **Ruptured membrane** 4 **hours ago**

A) Fetal Condition

Lie/Presentation Longitudinal lie, cephalic presentation

Amniotic fluid	C				B		M				
Moulding	1+				1+		1+				

B) Labor

Effacement	50%				50%		50%				
Hours	1	2	3	4	5	6	7	8	9	10	11
Time	10:00a	11:00a	12:00p	1:00p	2:00p	3:00p	4:00p				

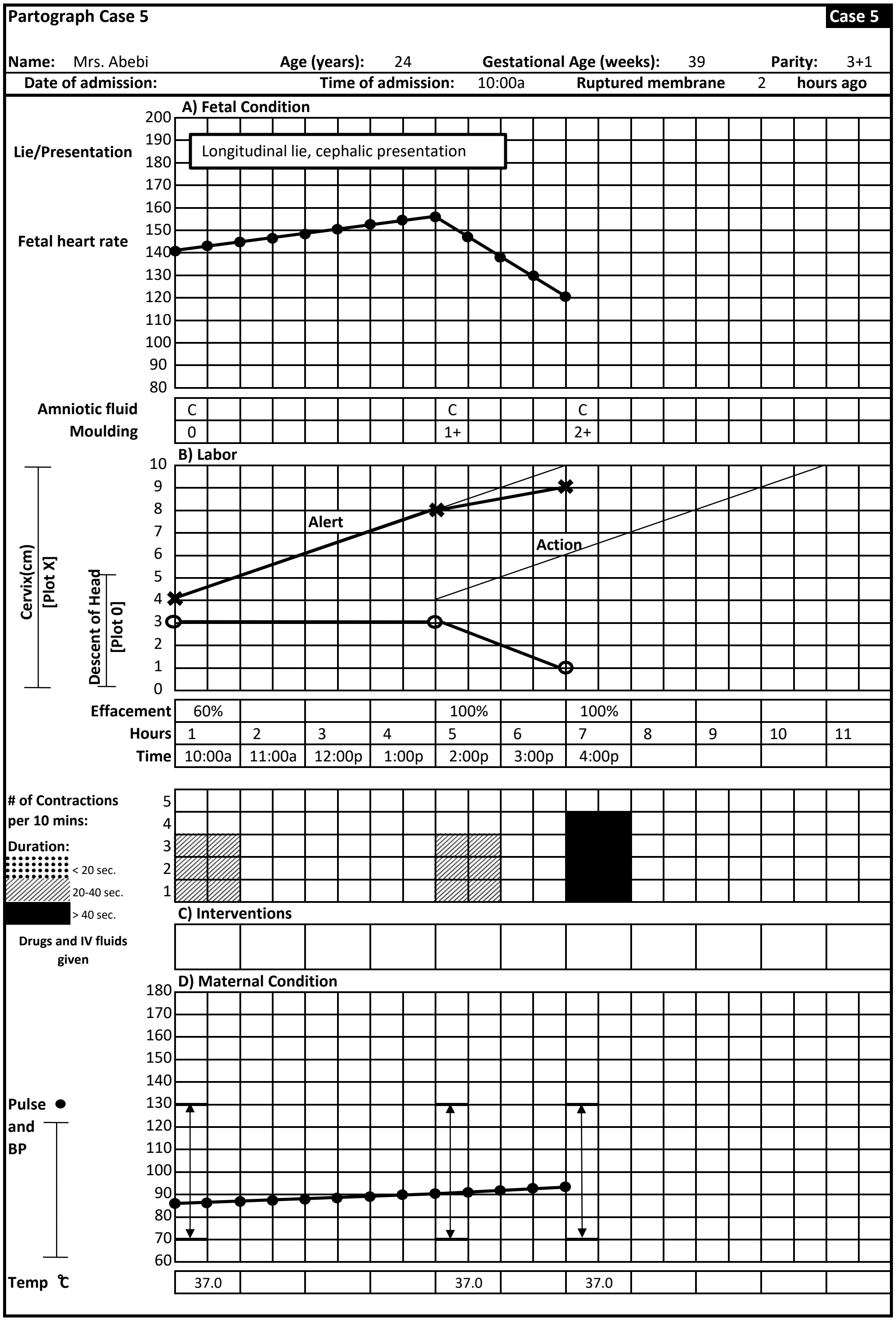
C) Interventions

Drugs and IV fluids given

D) Maternal Condition

Temp ℃	37.0				37.2		37.2				

Figure B.10: Partograph case 5



Partograph Case 5 **Case 5**

Name: Mrs. Abebi **Age (years):** 24 **Gestational Age (weeks):** 39 **Parity:** 3+1

Date of admission: **Time of admission:** 10:00a **Ruptured membrane** 2 **hours ago**

A) Fetal Condition

Lie/Presentation Longitudinal lie, cephalic presentation

Amniotic fluid	C				C		C				
Moulding	0				1+		2+				

B) Labor

Effacement	60%				100%		100%				
Hours	1	2	3	4	5	6	7	8	9	10	11
Time	10:00a	11:00a	12:00p	1:00p	2:00p	3:00p	4:00p				

C) Interventions

Drugs and IV fluids given

D) Maternal Condition

Temp ℃	37.0				37.0		37.0				

55

B.3 Overlap between actions in the experimental task and the direct clinical observations

The following is a complete list of items covered during the real-world provision of antenatal care and as measured by our observations of such care. The items in teal, blue and red are similar to (respectively) paid, primed and unprimed actions in the experiment.

1. Documentation
   - Ask about intake information, e.g., whether this is the first visit to this facility for this pregnancy.
   - Ask about maternal risk based on prior pregnancies, e.g., prior stillbirth or heavy bleeding during delivery
   - Ask about maternal risk during current pregnancy, e.g., bleeding, fever, blurred vision
   - Look at client's health card during or prior to exam
   - Write on client's health card
2. Procedures and lab orders
   - Vital signs: check blood pressure and weigh
   - Examine for anemia
   - Examine for edema
   - Palpate abdomen or conduct ultrasound (for fetal presentation, uterine height)
   - Check fetal heart rate
   - Examine breasts
   - Conduct vaginal exam
3. Perform or refer for test
   - Urine and anemia
   - Syphilis
   - HIV
   - Counseling for HIV if tested positive
4. Treatments (give/prescribe and explain)
   - Iron or folic acids
   - Tetanus injection
   - Anti-malarial
   - Intermittent preventive therapy (IPT)
   - Insecticide treated net
5. Counseling
   - Diet
   - Risk factors, e.g., bleeding, fever, blurred vision
   - Plan for delivery, e.g., whether client plans to deliver
   - Family planning after birth
6. End of the consultation
   - Outcome, e.g., sent home or referred
   - Where did client go after the consultation

C Price Response

We estimate the effort response to price for both primed and unprimed actions. For the former, we can exploit variation in the price assigned to tasks to examine how performance changes with the amount of incentive. While we randomly assigned actions to be incentivized, we purposively assigned higher prices to more complex actions. For instance, palpating the uterus is priced at 100 Naira, while monitoring contractions is priced at 50 Naira. Performance on a given action thus reflects responses to both the price and non-price characteristics of the actions. We can recover the price response by netting out the level of performance on each action in the information arm, where actions only differ in their non-price characteristics. We can then use these estimates of effort response to calculate price elasticities. Similarly, we can assess the response of effort on unprimed tasks to price per unit paid effort.

A caveat is that this exercise assumes that the cost of effort does not change between the information and incentives arms. If this assumption holds, then the non-price characteristics of an action are constant between the two types of arms and can be netted out using performance in the information arm. This assumption would be problematic if, say, there were an interaction between price and the non-price characteristics of the action. However, as performance simply entails identifying the correct action rather than actually conducting it, we believe this assumption to be reasonable. Moreover, the fact that we find identical patterns in rewards and penalties suggests that any interaction of non-price characteristics with the price would have to be symmetric in rewards and penalties.

Figure C.1 plots the percentage point difference in the incentive arms relative to the information arm, for actions from the three complex cases that are primed and correct. Going from zero price to the lowest price of 50 Naira increases effort by 6-8 pp, and the impact does not increase further in the incentive amount. This pattern also indicates the validity of our experimental task: since participants do not actually need to perform the action they identified, one might expect them to respond to the price to a greater degree than when they would have to incur substantial effort costs.

Figure C.1 suggests that effort responds relatively more when going from no incentive to a low price, compared to incremental increases in price. This tapering off suggests that the key role of the financial incentive may be to signal the importance of the task or increase its salience, and is consistent with evidence from public finance and environmental economics on the interaction of salience and financial (dis)incentives (Chetty et al., 2009; Sexton, 2015). It also aligns with evidence that anti-poverty cash transfers to households act as nudges to increase the salience of the behavior on which the transfer is conditioned (Benhassine et al., 2015) and larger transfers may not necessarily increase the behavioral response (Filmer and Schady, 2011). Many of the real-world PFP programs discussed above incentivize facilities rather than workers. In contrast, our contracts incentivize actions that are within the locus of control for the workers, have low effort costs, and directly link worker performance and payout. In this way, our experiment setup may be particularly conducive to generating responses even to small prices. While findings by Gneezy and Rustichini (2000) suggests that too small of an incentive can in fact reduce performance, our smallest price may lie above the range that lowers performance. Together, these findings suggest that PFP contracts can set prices to be relatively small but may need to exceed a lower bound to elicit performance.

Based on these estimates, we estimate that the price elasticities of effort lie between 0.08 and 0.50 for each of the paid and correct actions (Table C.1), and are comparable to the range of wage elasticities estimated by Oettinger (1999) and Goldberg (2016).

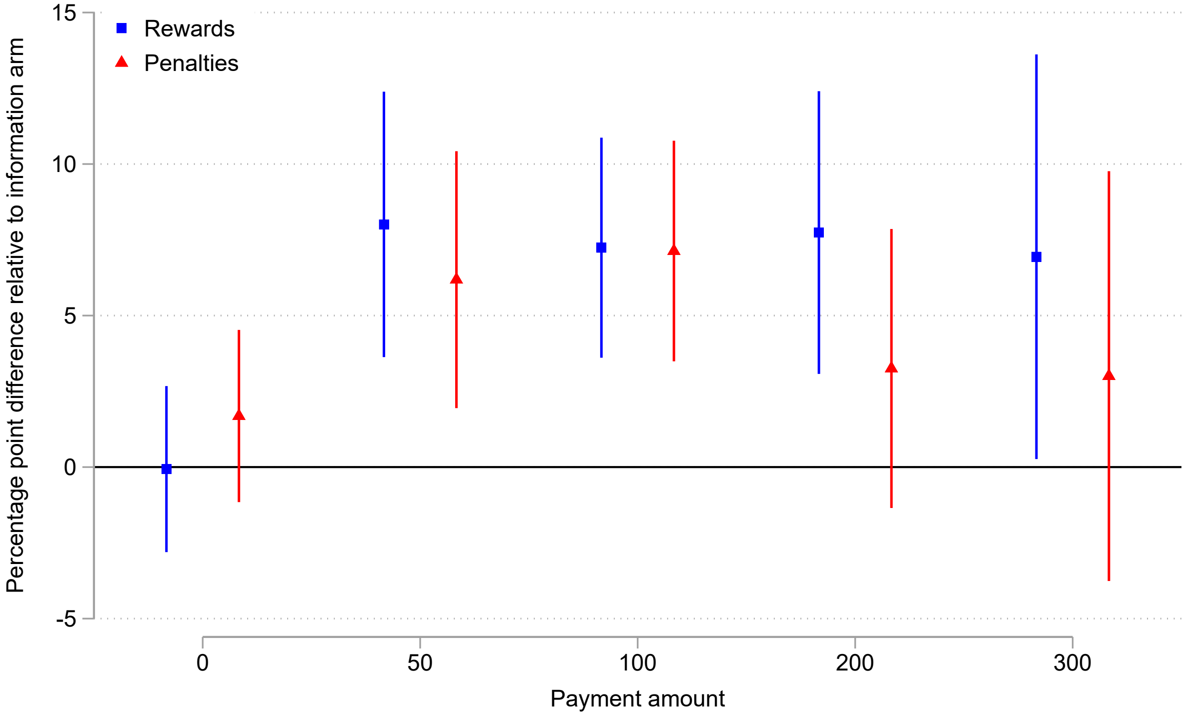
C.1 Pricing in spillovers on unprimed actions

Next, we consider how spillover effort varies in price per unit realized winnings. Doing so allows us to speak to optimal pricing in the presence of spillovers. Figure C.2 presents one way of estimating the spillover effort response to paid performance. The horizontal axis presents the monetary returns per unit paid performance, while the vertical axis presents percent correct performance on the unprimed actions. Much as with the price response of paid effort, we find that spillover performance also varies concavely in price paid, with most gains being captured at the lowest returns per unit performance on paid actions.

The two concave effort response functions in Figure C.1 and Figure C.2 suggest that in PFP contracts in LMIC healthcare settings, price may primarily function to signal importance rather than something that workers respond to linearly.[18] If a small price captures most of the gains from paying for effort on incentivizes and unincentivized actions, then there may be scope to make pay-for-performance contracts more cost effective.

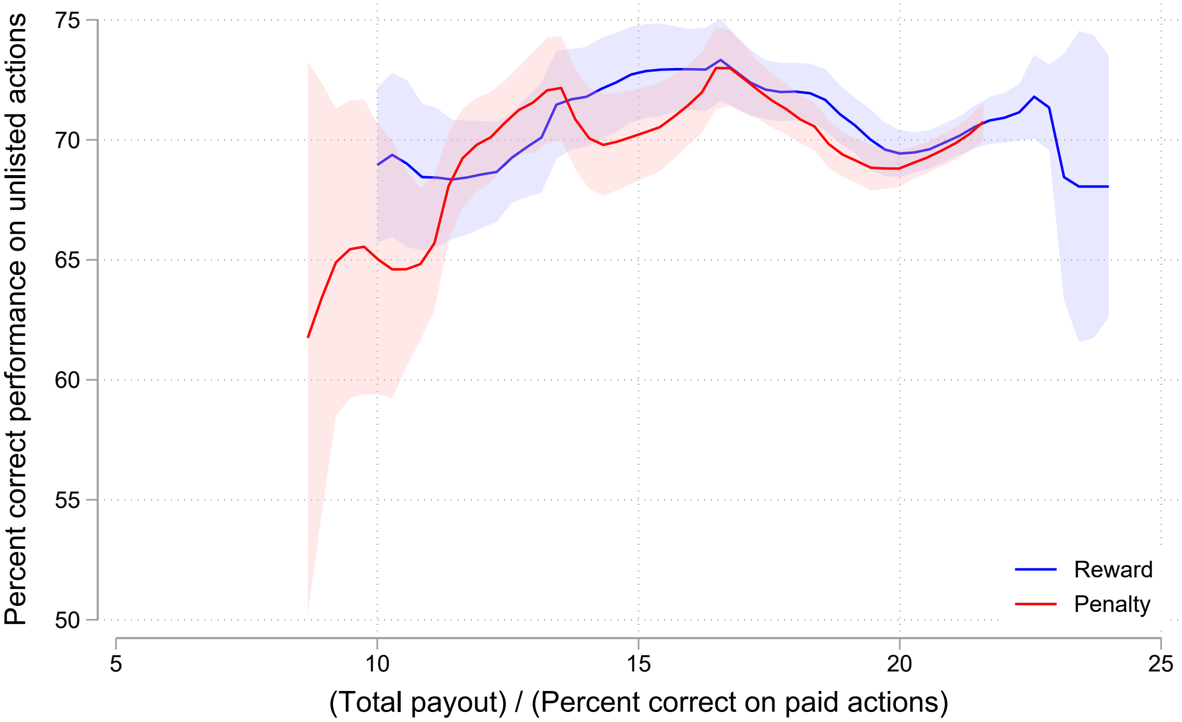
[18] The detailed results underlying Figure C.1 are presented in Table C.2.

Figure C.1: Price response of performance on paid or primed tasks



Based on correct and primed actions in the three complex cases. For the complete regression results, see Table C.2.

Figure C.2: Spillover performance by earnings per unit paid performance



Based on actions in the three complex cases. This graph relates dollar per unit correct performance on paid actions on the x axis to overall correct performance on unprimed actions. Local polynomial smooth plots with 95% confidence intervals.

Table C.1: Price elasticities of effort

Price	Level in info	%ΔReward-info	%ΔPenalty-info	"E_p" info	$E_p$reward	$E_p$penalty
0	28.30	24.93	19.21	0.14	0.12	0.10
50	18.06	33.34	32.49	0.27	0.50	0.49
100	55.26	13.07	5.45	0.83	0.20	0.08
200	60.74	10.66	4.57	1.52	0.27	0.11

Elasticities from unadjusted OLS models; s.e. clustered at worker level. Based on the three complex cases. While we randomly selected a subset of tasks to be listed or paid, we systematically set higher prices for more salient tasks. Hence, the "price" elasticity in the information arm (where there were no task-specific incentives) reflects the elasticity of effort in response to the salience of a task.

Table C.2: Percent correct by arm for actions with different incentives

	(1) Correct
Reward	-0.07
	(1.40)
Penalty	1.69
	(1.45)
Payment=50	28.35***
	(1.51)
Payment=100	18.05***
	(1.29)
Payment=200	55.24***
	(1.73)
Payment=300	60.64***
	(2.56)
Reward × Payment=50	8.01***
	(2.23)
Reward × Payment=100	7.24***
	(1.85)
Reward × Payment=200	7.74***
	(2.38)
Reward × Payment=300	6.94**
	(3.40)
Penalty × Payment=50	6.18***
	(2.16)
Penalty × Payment=100	7.13***
	(1.85)
Penalty × Payment=200	3.25
	(2.35)
Penalty × Payment=300	3.00
	(3.45)
Constant (Information × Payment=0)	10.82***
	(0.99)
N actions	21,744
R-squared	0.217

* p<0.10, ** p<0.05, *** p<0.01. OLS models; s.e. clustered at worker level. Listed actions from the three complex cases.

D Detailed results, validity of the lab experiment and robustness

In this appendix, we provide detailed results, and present various tests of the validity of the lab experiment as well as overall robustness of results.

D.1 Detailed results

The full regression output for the results presented Table 4 is reported in Table D.1.

D.2 Validity of the lab experiment

Although the experiment revolves around fictitious patients and does not impose actual effort costs on health workers—participants only state what clinical actions they would perform but do not actually implement them on patients—there are several design features that may help make this setup realistic along the lines of a framed field experiment (Harrison and List, 2004). In particular, the incentives are real, the participants are actual health workers whose daily work—providing labor and delivery care—aligns with our experimental task, and the study was conducted in their primary workplace. We also find that overall performance on our task is comparable to non-experimental assessments of knowledge by the same health workers. Specifically, participants in the information arm have an average score of 53 percent on our task, which is similar to the average scores on the knowledge vignette (about 53 percent). This level of quality of care is typical for LMIC settings (see, e.g., Das et al., 2008).

The response patterns also indicate that participants responded meaningfully to the lab experiment. First, in the two simple cases, participants respond yes or no. If they were randomly selecting a response, we would expect to see responses of approximately 50 percent for each of these cases. Instead, we observe 69 and 27 percent correct performance, respectively, which suggests we are capturing actual variation even in the so-called simple cases. Second, participants in the information arm performed at levels that were broadly in keeping with their levels of assessed knowledge (Table 2), even if they did not stand to gain from it. Third, seen most clearly in the empirical distributions for the three treatment arms (Figures D.1a and D.1b), the gains from incentives come from the middle of the performance distribution, suggesting that the impact of incentives is not driven by participants who were not paying any attention at all.

Fourth, participants in the incentives arms did not merely minimize effort by naming all paid actions, although doing so would have strictly increased their payout (with the exception of incorrect referrals). Participants also did not settle for the lowest possible payout by not naming any action. A related concern may be that our impacts are driven by sophisticated agents who play the payment maximizing strategy, while "naïve" players neither understood the setup nor responded in a meaningful way. In Table D.2, we examine whether overall performance impacts are driven by "sophisticated" individuals who always named the two actions (monitor contraction and fetal heart rate) that are correct in the three complex cases. Our results show that this type of participant did not differently to the experiment, although relatively few workers appear to have played the payment maximizing strategy.

Fifth, as highlighted in Table 1, participants identified both incorrect actions that were paid, mimicking real-world overuse (Lopez et al., forthcoming)), as well as actions that were neither primed nor paid (e.g., measuring vital signs), suggesting they did not simply mimic the checklist provided. Sixth, we find that participants in the incentives arms spend more time on the interview than those in the information arm (Table D.3), which has been used as a measure of performance and effort in similar studies (Cattaneo et al., 2017; Lavy, 2016; Rivkin and Schiman, 2015).[19]

Seventh, our estimated price responses are relatively low and consistent with estimated wage elasticities in LMICs (Goldberg, 2016). In the context of hypothetical tasks with limited effort costs one might expect high price elasticities (Fehr and Goette, 2007); instead our estimated responses are comparable to "real-world" estimates.

Eighth, we find that our results are robust to controlling for participants' knowledge as measured with the antenatal care vignette, suggesting that participants took the experiments seriously and our task induced real attention and effort.

A final concern may be that the experiment does not represent an appropriate test of performance if workers do not know how to use the clinical record that we use for our assessments. In fact, this clinical record—called a partograph—is a standard tool that is "strongly recommended" by the World Health Organization (WHO) and used by Nigeria's Federal Ministry of Health to train health workers providing maternal health services (WHO, 2014; White Ribbon Alliance, 2015). All participants in our experiment routinely provide maternity care service and are supposed to regularly use partographs in their work. Nonetheless, we screened participants for knowledge of the partograph and the experiment was only conducted with those who stated that they knew how to use a partograph. The correct use of partographs is also often part of PFP programs for health care in LMICs (Fritsche et al., 2014), although not in the Nigerian trial that our experiment was embedded in. Thus, the partograph is a valid instrument with which to assess health worker performance in this setting. Our partograph-based assessment is most similar to vignettes, which are a widely accepted tool to assess provider performance (Das et al., 2008).

D.3 Robustness checks

In this section, we examine the robustness of our estimated impacts. One concern about our lab results may arise from the variation in the nature and number of relevant actions across cases. First, to addresses the concern that different actions have different characteristics, i.e. that they may vary in the difficulty or required effort, we calculate an aggregate score using Item Response Theory (IRT) (Das and Hammer, 2005; Das et al., 2016). IRT accounts for these characteristics by weighting actions differently. We construct IRT scores for overall performance, as well as subscores for each type of action. Results scaled by the IRT score results are also robust, except for the null effect on incorrect paid actions in the IRT analysis (Table D.4).

[19] Some interviews were not completed within the day that they were started, and the survey enumerator returned to the facility when the health worker was next on shift to complete the interview. There may also have been instances in which an enumerator failed to promptly record the interview as complete. In these instances, the interview length cannot be calculated correctly from time stamps because the end date was several days after the start, and the interviews were not "paused" in the interim. Similarly, in some cases the enumerator prematurely marked the interview as "ended" while it was still ongoing. We trim observations in the 10th and 90th percentile from the analysis of time on task. This leaves us with a sample of 1,045 observations in Table D.3.

We account for the differences in the number of relevant actions by reporting results that equally weight each case and those that equally weight each individual item responses. We present the empirical cumulative distributions for the complex cases only in Figure D.2a and Figure D.2b while columns 3–7 of Table 3 report the impact estimates for each case separately. The graphs and disaggregated analysis point to the robustness of results, with rewards and penalties outperforming information alone. This said, the two simple cases have higher impacts of rewards–5.4 pp (8 percent) and 11 pp (42 percent), compared with magnitudes of about 1 to 3 pp (3 to 5 percent) for the complex cases. Strikingly, even where the task structure provides a simple yes or no answer, arriving at that answer does not appear to be trivial. Indeed, performance in the information arm is at both its highest and its lowest for the two simple cases, at 69 percent and 27 percent. The results are robust to measuring performance in z-scores instead of these measures of proportion of correct responses (presented in Table D.5).

Next, we examine knowledge as a potential confounder in Table D.6. Scoring above the median on the antenatal care vignette has a small but strongly positively correlation with performance on the experiment. That the more knowledgeable workers actually performed better on our task as well is indicative of the validity of our experiment. However, controlling for knowledge does not change the sign or statistical significance of our main estimates. Participants with a higher knowledge score do not respond differently to the incentives than workers with a lower knowledge score. Knowledge thus does not fully determine performance in our experimental task and our task captures dimensions of performance that extend beyond knowledge.[20]

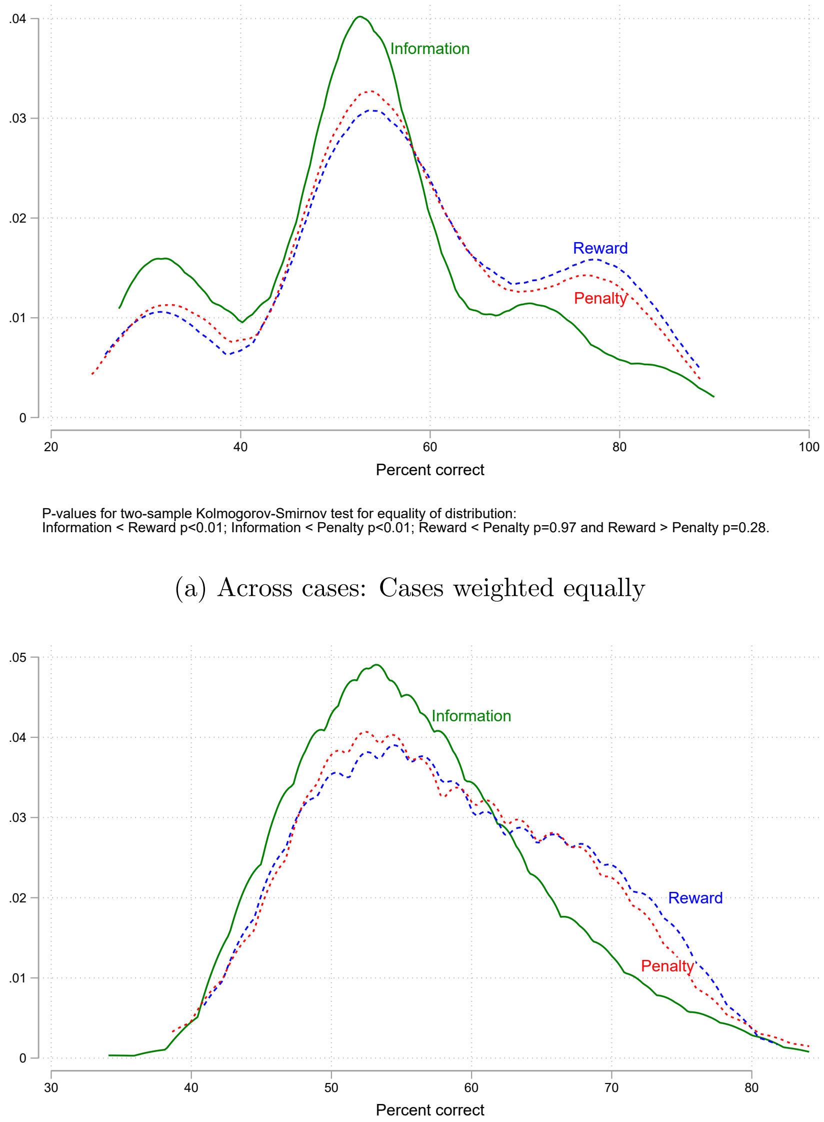
While both study stages were sucessfully randomized as shown in Table 2, another concern may be whether our results are robust to accounting for any potential covariate imbalance. To test the robustness of our experimental results, we present control for covariates for lab-in-the-field results in Table D.8 and real-world results in Table D.9. The signs and magnitudes of estimated impacts are robust. Still, significant results generally remain significant, suggesting that potential covariate imbalance is not a major concern for the robustness of our results.

Similarly, since the real-world impacts are only estimated for a subset of the participants of the first-stage experiment, a question may be whether our main results hold for the direct clinical observations sub-sample. With regards to the first-stage results, Table D.10 shows that first-stage lab experiment participants who were selected for observation of real-world care provision are comparable to the workers not selected for the observation of real-world care provision, except that they are significantly less likely to be male and to have greater-than-median experience. We also see that each lab arm comprises approximately a third each of the experimental participants of the observation sample. (Recall that we additionally observed real-world care provided by some pure control workers, i.e. those who had not participated in the first stage). These comparisons thus indicate the successful cross-randomization of sampling lab participants for direct observation. Our results for the first-stage experiment are robust to estimating impacts for the sub-sample of lab participants who were also observed providing real-world care (Table D.11). With regards to the real-world results, Table 2 shows that the 67 “pure control” and the 344 first-stage participants who were observed providing real-world care are comparable. As Table D.10 shows that lab participants randomly selected for the second stage were less likely to be male than lab workers not selected for

[20] A related consideration is whether certain workers respond more strongly to the PFP contract, as found in Mohanan et al. (2021). In Table D.7, we explore the effect of observables on performance in the lab-in-the-field and report some indication that male workers and those who are younger than the median worker in the sample perform less well than female and older workers.

the second stage, we also use coarsened exact matching to re-estimate the real world impacts. As shown in **??**, our results for the real-world analysis are robust to only estimating the impacts for the sub-sample that is balanced using coarsened exact matching.

Figure D.1: Empirical probability distributions



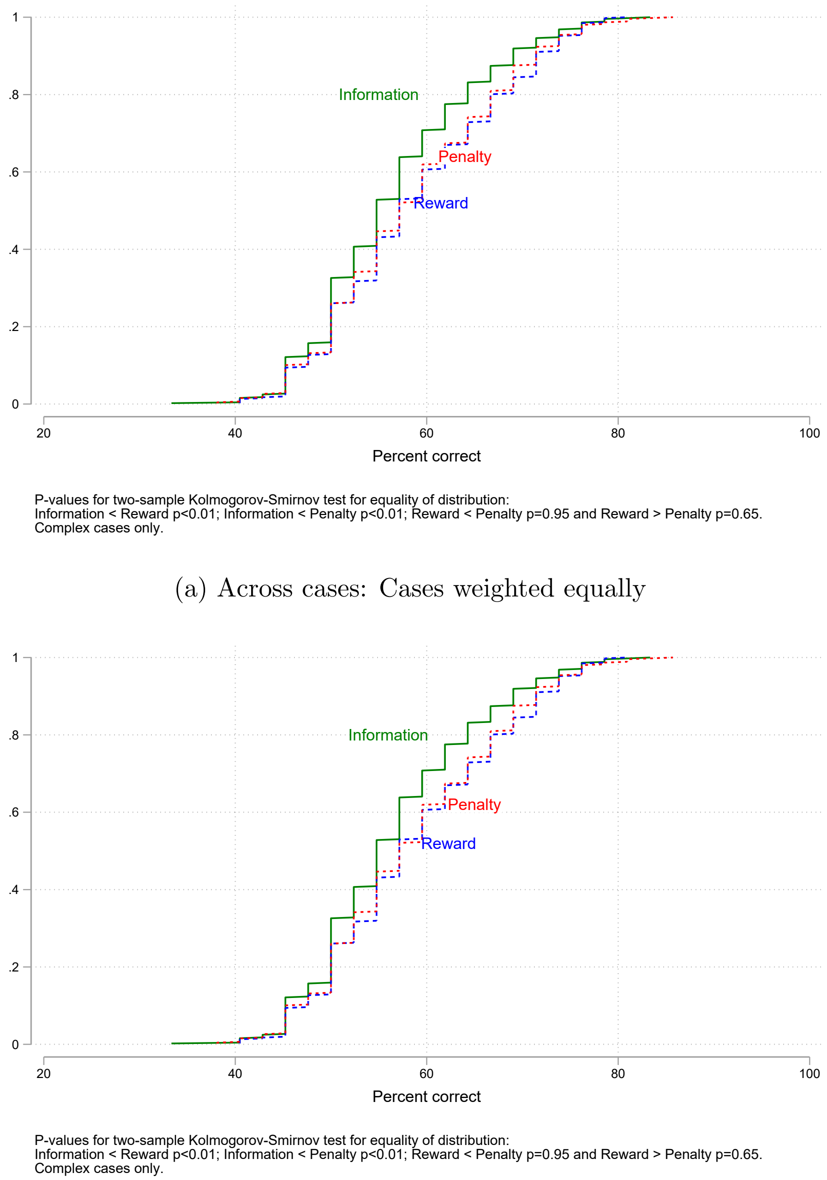
P-values for two-sample Kolmogorov-Smirnov test for equality of distribution:
Information < Reward p<0.01; Information < Penalty p<0.01; Reward < Penalty p=0.97 and Reward > Penalty p=0.28.

(a) Across cases: Cases weighted equally

P-values for two-sample Kolmogorov-Smirnov test for equality of distribution:
Information < Reward p<0.01; Information < Penalty p<0.01; Reward < Penalty p=0.97 and Reward > Penalty p=0.65.

(b) Across responses: Responses weighted equally

Figure D.2: Empirical cumulative distributions for complex cases



P-values for two-sample Kolmogorov-Smirnov test for equality of distribution:
Information < Reward p<0.01; Information < Penalty p<0.01; Reward < Penalty p=0.95 and Reward > Penalty p=0.65.
Complex cases only.

(a) Across cases: Cases weighted equally

P-values for two-sample Kolmogorov-Smirnov test for equality of distribution:
Information < Reward p<0.01; Information < Penalty p<0.01; Reward < Penalty p=0.95 and Reward > Penalty p=0.65.
Complex cases only.

(b) Across responses: Responses weighted equally

Table D.1: Percent correct by arm for different actions

	All actions		Necessary actions								Unnecessary actions			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	All	All	Unlisted	Unlisted	Listed	Listed	Paid	Paid	Listed or Paid	Listed or Paid	Unlisted	Unlisted	Paid	Paid
Reward	2.24***	2.37***	4.18**	4.76***	-0.07	0.53	7.43***	7.89***	4.21***	4.73***	-0.60	-0.98	2.84**	2.66**
	(0.60)	(0.57)	(1.67)	(1.56)	(1.40)	(1.34)	(1.80)	(1.71)	(1.43)	(1.35)	(1.05)	(1.00)	(1.26)	(1.25)
Penalty	1.96***	2.09***	3.85**	4.55***	1.69	2.33*	7.94***	8.50***	5.26***	5.86***	-1.69	-2.17**	1.20	1.04
	(0.60)	(0.57)	(1.68)	(1.57)	(1.45)	(1.39)	(1.82)	(1.73)	(1.48)	(1.40)	(1.09)	(1.04)	(1.25)	(1.24)
Constant (Information)	56.60***	61.49***	28.23***	31.45***	10.82***	6.87***	38.06***	45.29***	26.39***	25.87***	91.52***	98.15***	79.44***	87.50***
	(0.41)	(0.60)	(1.17)	(1.54)	(0.99)	(1.26)	(1.23)	(1.88)	(0.99)	(1.37)	(0.75)	(0.88)	(0.92)	(1.20)
Worker covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Case FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N actions	57,078	57,078	10,872	10,872	8,154	8,154	10,872	10,872	19,026	19,026	21,744	21,744	5,436	5,436
R-squared	0.000	0.011	0.002	0.076	0.001	0.038	0.005	0.059	0.002	0.033	0.001	0.028	0.001	0.020

* p<0.10, ** p<0.05, *** p<0.01. OLS models; s.e. clustered at worker level. Actions from the three complex cases.

Table D.2: Overall performance (percent correct)

	All		Sophisticated		Not sophisticated	
	(1) Across cases	(2) Across responses	(3) Across cases	(4) Across responses	(5) Across cases	(6) Across responses
Reward	4.36***	2.04***	5.79**	-0.64	3.52***	1.74***
	(0.97)	(0.45)	(2.59)	(0.91)	(1.07)	(0.49)
Penalty	3.43***	1.87***	3.27	-0.96	2.61**	1.66***
	(1.00)	(0.45)	(2.39)	(0.91)	(1.18)	(0.51)
Constant (Information)	53.15***	56.44***	64.25***	69.17***	51.51***	54.62***
	(0.59)	(0.26)	(1.73)	(0.64)	(0.64)	(0.28)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
P-value Penalty v Reward	0.310	0.698	0.231	0.665	0.413	0.877
N respondents	1,359	1,359	215	215	1,144	1,144
R-squared (overall)	0.019	0.015	0.017	0.006	0.013	0.008

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors. Sophisticated individuals are those who always named the two actions (monitor contraction and fetal heart rate) that are correct in the three complex cases.

Table D.3: Interview duration

	(1)	(2)
Reward	78.20*	
	(44.99)	
Penalty	129.16***	
	(44.38)	
Incentives (Reward or Penalty)		103.39***
		(39.09)
Male	44.53	46.07
	(54.91)	(55.40)
Older than median	-48.34	-50.84
	(53.96)	(54.19)
Doctor or nurse	-81.79	-83.36
	(97.46)	(97.86)
Above median experience	0.67	2.53
	(54.12)	(54.45)
Above median knowledge	-195.00***	-192.64***
	(69.36)	(69.04)
Constant (Information)	802.74***	801.81***
	(59.56)	(59.63)
Facility fixed effects	Yes	Yes
P-value Penalty v Reward	0.241	
N respondents	1,045	1,045
R-squared (overall)	0.008	0.008

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors. Duration of the full health worker interview in minutes; 10th and 90th percentile trimmed.

Table D.4: Performance on IRT score

	All	Necessary actions				Unnecessary actions	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Unprimed	Primed	Paid	Primed or paid	Unprimed	Paid
Reward	0.11***	0.13***	-0.00	0.14***	0.10***	-0.04	0.01
	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.01)
Penalty	0.14***	0.13***	0.03	0.18***	0.14***	-0.05	0.01
	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.01)
Constant (%)	-0.08***	-0.07***	-0.01	-0.10***	-0.08***	0.04**	-0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.00)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
p-values from tests of coefficients							
Penalty v Reward	0.448	0.979	0.383	0.301	0.300	0.697	0.863
N respondents	1,359	1,359	1,359	1,359	1,359	1,359	1,359
R-squared (overall)	0.007	0.007	0.001	0.012	0.008	0.003	0.003

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors. IRT latent variables estimated with two-parameter logistic models except for unnecessary-paid is based on a one-parameter model. All-action analysis uses actions from all cases; by-type analysis uses actions from the three complex cases.

Table D.5: Overall performance (z-scores)

	All cases		By case				
	(1) Across cases	(2) Across responses	(3) Simple 1	(4) Simple 2	(5) Complex 1	(6) Complex 2	(7) Complex 3
Reward	0.29***	0.22***	0.12*	0.24***	0.09*	0.15**	0.20***
	(0.06)	(0.05)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)
Penalty	0.23***	0.21***	0.06	0.20***	0.09	0.14**	0.19***
	(0.07)	(0.05)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)
Constant (Information)	-0.17***	-0.14***	-0.06	-0.15***	-0.06*	-0.10***	-0.13***
	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-value Penalty v Reward	0.310	0.698	0.376	0.584	0.960	0.923	0.866
N respondents	1,359	1,359	1,359	1,359	1,359	1,359	1,359
R-squared (overall)	0.019	0.015	0.003	0.012	0.005	0.006	0.010

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors.

Table D.6: Accounting for knowledge as a potential confounder of performance in the lab-in-the-field experiment

	All cases		By case				
	(1) Across cases	(2) Across responses	(3) Simple 1	(4) Simple 2	(5) Complex 1	(6) Complex 2	(7) Complex 3
Reward	4.08***	3.08***	0.90	10.64**	3.43***	1.77**	3.66***
	(1.48)	(0.66)	(4.89)	(4.35)	(1.14)	(0.69)	(1.16)
Penalty	3.03*	2.78***	-0.91	7.80	4.25***	1.18	2.81**
	(1.55)	(0.67)	(5.35)	(5.12)	(1.14)	(0.76)	(1.16)
Above median knowledge	1.85	3.11***	1.18	-1.75	6.29***	1.01	2.50
	(2.05)	(0.87)	(6.49)	(6.41)	(1.52)	(0.99)	(1.58)
Reward × above median knowledge	0.57	-1.95**	8.61	1.08	-4.13**	-0.97	-1.72
	(2.06)	(0.97)	(6.42)	(6.74)	(1.64)	(1.07)	(1.87)
Penalty × above median knowledge	0.76	-1.75*	6.69	3.33	-5.83***	0.02	-0.41
	(2.08)	(0.94)	(7.22)	(7.02)	(1.73)	(1.19)	(1.66)
Constant (Information)	52.19***	54.82***	68.04***	27.42***	59.04***	51.45***	54.99***
	(1.26)	(0.52)	(4.06)	(3.73)	(0.91)	(0.54)	(0.90)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-value Penalty v Reward	0.445	0.667	0.704	0.541	0.453	0.451	0.481
N respondents	1,359	1,359	1,359	1,359	1,359	1,359	1,359

* p<0.10, ** p<0.05, *** p<0.01. Unadjusted OLS models; s.e. clustered at facility level.

Table D.7: Predictors of above-median performance

	(1) Across cases	(2) Across responses
Male	-1.73	-9.21*
	(6.38)	(5.39)
Older than median	-5.81	-8.48*
	(5.43)	(4.67)
Doctor or nurse	-0.88	5.53
	(8.63)	(6.22)
Above median experience	4.20	3.00
	(5.41)	(4.61)
Constant	54.10***	62.00***
	(3.65)	(3.24)
Facility fixed effects	Yes	Yes
N respondents	914	914
R-squared (overall)	0.001	0.005

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors. Binary outcome equal to one if the participant scored at or above the median for her group (information, reward, penalty).

Table D.8: Performance across cases and responses (percentage points)

	Without covariates		With covariates	
	(1)	(2)	(3)	(4)
	Across cases	Across responses	Across cases	Across responses
Reward	4.36***	2.04***	4.51***	1.98***
	(0.97)	(0.45)	(0.96)	(0.45)
Penalty	3.43***	1.87***	3.49***	1.77***
	(1.00)	(0.45)	(1.00)	(0.44)
Male			0.22	-1.35**
			(1.26)	(0.58)
Older than median			-2.17*	-0.98*
			(1.13)	(0.53)
Doctor or nurse			0.39	2.44***
			(1.83)	(0.79)
Above median experience			-0.65	-0.48
			(1.12)	(0.50)
Above median knowledge			2.21	1.48**
			(1.58)	(0.63)
Constant (%)	53.15***	56.44***	53.32***	56.57***
	(0.59)	(0.26)	(1.21)	(0.55)
Facility fixed effects	Yes	Yes	Yes	Yes
p-values from tests of coefficients				
Penalty v Reward	0.310	0.698	0.268	0.635
N respondents	1,359	1,359	1,359	1,359
R-squared (overall)	0.019	0.015	0.041	0.050

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors. Omitted category: 15+ years since certification.

Table D.9: Effect on behaviors in direct clinical observations of antenatal care (percent) with covariate controls

	Related to paid actions				Related to primed actions		Related to unprimed actions					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Fetal heart-beat	Palpate ab-domen	Fetal heart-beat	Palpate ab-domen	Record infor-mation	Record infor-mation	Perform or refer for urine test	Blood pres-sure	Conduct vaginal exam	Perform or refer for urine test	Blood pres-sure	Conduct vaginal exam
Information	12.93	12.48*			8.26*		-9.78	10.33	-5.07			
	(10.38)	(7.23)			(4.45)		(15.17)	(7.93)	(5.74)			
Incentive	22.41**	22.91***			5.94*		-4.11	10.43	-4.35			
	(10.03)	(7.58)			(3.45)		(13.57)	(6.35)	(6.05)			
Information			13.11	12.62*		8.28*				-10.83	10.70	-5.27
			(10.44)	(7.21)		(4.47)				(14.69)	(8.04)	(5.79)
Reward			24.75**	25.15***		6.15				-14.34	13.05*	-7.02
			(10.70)	(7.93)		(3.85)				(12.75)	(7.17)	(6.75)
Penalty			19.30*	19.94***		5.68*				9.43	7.98	-0.82
			(10.49)	(7.43)		(3.22)				(14.93)	(6.65)	(6.17)
Male	-1.22	6.77	-1.51	6.67	-8.02	-8.05	-8.47	0.05	0.19	-7.16	0.18	0.52
	(9.40)	(7.94)	(9.37)	(7.95)	(5.12)	(5.19)	(10.97)	(7.01)	(3.61)	(10.29)	(6.99)	(3.69)
Older than median	5.61	5.02	5.18	4.73	4.11	4.08	10.57	12.76*	5.46	12.27	11.56	5.94
	(6.40)	(4.14)	(6.53)	(4.04)	(2.68)	(2.62)	(8.60)	(6.76)	(5.01)	(8.46)	(7.30)	(5.25)
Doctor or nurse	-24.22**	-12.41	-23.19**	-11.45	-0.30	-0.21	10.31	-7.96	-1.06	5.80	-7.35	-2.23
	(11.17)	(8.28)	(11.13)	(8.01)	(3.85)	(4.09)	(14.92)	(7.19)	(4.05)	(13.57)	(7.48)	(4.06)
Above median experience	10.88	3.87	10.65	3.50	-2.79	-2.82	-10.36	-12.14*	-1.07	-9.54	-11.71*	-0.80
	(6.86)	(5.65)	(6.76)	(5.63)	(3.05)	(3.12)	(8.18)	(6.40)	(2.65)	(8.35)	(6.51)	(2.63)
Above median knowledge	-2.79	-6.12	-2.34	-5.81	4.03	4.07	10.75	7.38	2.08	8.75	7.39	1.56
	(7.61)	(6.23)	(7.63)	(6.19)	(3.50)	(3.59)	(11.28)	(7.61)	(6.16)	(10.82)	(7.57)	(5.90)
Constant (Naive)	59.28***	69.69***	59.51***	69.96***	89.82***	89.83***	35.02**	80.02***	8.08**	34.22**	80.26***	7.82**
	(10.25)	(8.06)	(10.34)	(7.98)	(3.89)	(3.88)	(15.25)	(7.52)	(3.15)	(15.09)	(7.62)	(3.32)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-value Info v Incentive	0.124	0.026			0.276		0.459	0.983	0.746			
P-value Penalty v Reward			0.402	0.114		0.799				0.004	0.347	0.182
P-value Info v Reward			0.078	0.014		0.379				0.658	0.618	0.515
P-value Info v Penalty			0.395	0.111		0.226				0.034	0.648	0.251
N respondents	411	410	411	410	410	410	403	353	411	403	353	411
R-squared (overall)	0.023	0.025	0.022	0.022	0.006	0.006	0.004	0.004	0.002	0.018	0.006	0.000

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors. Related to actions that are listed and paid in the experiment.

Table D.10: Comparison of lab experiment participants with and without direct clinical observations (percent)

Variable	(1) Selected for observation Mean/SE	(2) Not selected for observation Mean/SE	T-test Difference (1)-(2)
Outcomes†			
Across cases	54.54 (0.79)	56.18 (0.48)	-1.64
Across responses	55.93 (0.43)	58.38 (0.29)	-2.45
Experiment status			
Information	31.40 (2.51)	33.20 (1.48)	-1.81
Reward	35.17 (2.58)	33.10 (1.48)	2.07
Penalty	33.43 (2.55)	33.69 (1.48)	-0.26
Covariates			
Male	17.15 (2.04)	25.62 (1.37)	-8.46***
Older than median*	57.56 (2.67)	51.13 (1.57)	6.43
Doctor or nurse	12.21 (1.77)	8.67 (0.88)	3.54
Above median experience‡	54.36 (2.69)	49.26 (1.57)	5.10*
Above median knowledge◇	47.38 (2.70)	53.30 (1.57)	-5.92
NSHIP pilot status			
PFP	40.99 (2.66)	43.94 (1.56)	-2.95
DFF	40.12 (2.65)	44.63 (1.56)	-4.51
Control	18.90 (2.11)	11.43 (1.00)	7.47
Number of observations	344	1015	

Notes: † Across cases weighs each case equally; across responses weighs each response equally. ∗ The median age is 38. ‡ High is defined as greater than median; median experience is 9.5 years; and the median score on the assessment of knowledge of antenatal care protocol is 51.52 percent. The value displayed for t-tests are the differences in the means across the groups. Standard errors are robust. Facility fixed effects are included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table D.11: Performance of all second-stage participants compared to a coarsened-exact matched subset (percentage points)

	Real-world sample						Matched sample	
	(1) Across cases	(2) Across responses	(3) Across cases	(4) Across responses	(5) Across cases	(6) Across responses	(7) Across cases	(8) Across responses
Incentive					3.67	3.31***	9.01	7.52***
					(2.87)	(1.11)	(6.56)	(2.27)
Reward	2.82	3.85***	3.47	4.36***				
	(3.45)	(1.40)	(3.30)	(1.28)				
Penalty	4.18	1.97	3.98	1.71				
	(3.18)	(1.59)	(3.45)	(1.49)				
Male			4.17	0.38	4.13	0.58		
			(4.07)	(2.03)	(4.04)	(1.95)		
Older than median			-4.88	-0.28	-4.91	-0.07		
			(3.99)	(1.63)	(3.92)	(1.63)		
Doctor or nurse			4.18	5.59***	4.31	4.89***		
			(4.33)	(1.83)	(4.13)	(1.86)		
Above median experience			3.93	1.22	3.86	1.60		
			(3.47)	(1.43)	(3.48)	(1.44)		
Above median knowledge			2.44	3.21	2.51	2.80		
			(5.73)	(2.15)	(5.68)	(2.26)		
Constant (%)	52.15***	53.92***	50.28***	51.06***	50.33***	50.81***	49.84***	51.99***
	(1.97)	(0.88)	(4.79)	(1.86)	(4.77)	(1.93)	(3.28)	(1.13)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
p-values from tests of coefficients								
Penalty v Reward	0.682	0.222	0.888	0.095				
N respondents	344	344	344	344	344	344	204	204
R-squared (overall)	0.017	0.013	0.027	0.013	0.027	0.011	0.033	0.034

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and robust standard errors. Omitted category: 15+ years since certification. Real-world sample contains all the observations for which data on real world care are available. Matched sample contains subset of observations of real world care that balance information and incentive arms, see Table 2.

E Health impact

We can perform a rudimentary calculation of the potential health gain of our experimental impacts by translating the increase in adherence to a checklist of essential procedures for childbirth to mortality gains for newborns within the first seven days of birth (early neonatal mortality), which is likely most malleable to the effort applied by the health worker. We do so using two estimates of the neonatal mortality reduction from improved adherence to the international standard of care for childbirth. Because there are additional benefits of better adherence to the standard of care for both the mother and newborn, we likely underestimate the health gains.

In a study of delivery care in health facilities in Uttar Pradesh, India, Semrau et al. (2020) estimate that each additional action (out of 10 actions) by the health care provider is associated with a 30 percent decrease in early neonatal mortality. We observe an 8 percent improvement in the rewards arm relative to the information arm, corresponding to 0.8 additional actions. Assuming linearity, this would imply a 24 percent reduction in early neonatal mortality among births in health facilities. Using the observed neonatal mortality of 33 per 1,000 deliveries in Semrau et al. (2020), the 24 percent reduction translates into 8 averted early neonatal deaths per 1,000 facility-based deliveries. Of the approximately 7.6 million births in Nigeria in 2018, 39 percent, or roughly 3 million, occurred in health facilities (National Population Commission, 2019). Thus, if applied at scale, the improvement in adherence we observed would translate into 24,000 fewer neonatal deaths.

We can benchmark the size of this impact in two ways. The first is the economic benefit. The value of a statistical life in Nigeria is estimated to be USD 485,000 (Viscusi and Masterman, 2017) with a life expectancy of 55 years in 2018 (World Bank, 2019). Thus, 24,000 fewer neonatal deaths would translate into an annualized economic benefit of USD 212 million. Second, we can compare our estimated health gain to that from alternative policies. Okeke (2023) reports on a cluster-randomized trial in Nigeria in which either qualified physicians or mid-level professionals were sent to primary care health facilities. He finds that physicians produce significantly higher quality of antenatal and delivery care, translating into a short run intent-to-treat impact of 6–8 fewer early neonatal deaths per 1,000 live births. Okeke (2023) also estimates that more sustained contact with physicians over the course of the pregnancy translates into a mortality reduction of 9–13 deaths per thousand.

Okeke (2023) also notes that this magnitude of improvement is equivalent to the entire newborn mortality reduction Nigeria achieved between 1990 and 2018. Thus, our estimated impact of 8 neonatal deaths averted from improving health worker effort on the intensive margin is comparable to the short-run extensive margin gains from adding an additional physician to a primary health facility. Because only 39 percent of all births in Nigeria occur in health facilities, other approaches may have an even greater impact on neonatal mortality. For instance, Okeke and Abubakar (2020) estimate that a conditional cash transfer in Nigeria prevented up to 85,000 neonatal deaths nationally. Compared to our estimated impact of 24,000, the larger impact of a cash transfer reflects the fact that most neonatal mortality risk is in fact for births outside of health facility settings.

An alternative way of benchmarking our health impact to the estimates provided by Okeke (2023) is by calculating the facility-level improvements in quality from the financial incentives in our lab experiment. We estimate that performance on directly incentivized tasks improved by 20 percent. There are 4.2 health workers on average in the health facilities in our sample. Assuming that the impact on worker quality scales additively, the total improvement in facility quality would be 84 percent.

While we lack the cost data required for a full cost-effectiveness analysis of implementing this experimental PFP contract at scale, we note that the incentives arms in our experimental task had outlays that were about 5 percent higher those in the information arm. This simplified setup abstracts away from the substantial administrative costs incurred by actual PFP programs in LMICs (Fritsche et al., 2014). Such costs go towards training, verifying facility-reported data on performance, and executing payments. In the concurrent PFP trial, about 36.5 percent of total program costs were for administration and operations rather than disbursements to health facilities (Khanna et al., 2021).