CENTER FOR GLOBAL DEVELOPMENT

Information, Loss Framing, and Spillovers in Pay-for-Performance: Experimental Evidence from Health Workers in Nigeria

Sebastian Bauhoff, Eeshani Kandpal

Abstract

We report on an experiment that isolates key components of pay-for-performance contracts: the effect of financial incentives beyond priming; spillovers from incomplete contracts; and the persistence of incentive effects after contracts end. First conducting a lab-in-the-field experiment with 1,359 maternity care workers in Nigeria, we prime them with a checklist of salient clinical actions and randomize them within 690 facilities to receive either no incentives, rewards, or penalties tied to care for hypothetical patients. Relative to priming alone, both rewards and penalties increase performance on paid actions by 7-8 percentage points (pp; 20 percent) on incentivized actions and generate spillovers of 4 pp (14-15 percent) on unincentivized actions. In the second stage, we observe care for real patients in a cross-randomized sub-sample of 400 participants conducted within 72 hours after the lab experiment. Exposure to an incentivized arm improves real-world performance by 16 to 18 pp (21-26 percent) on actions previously incentivized in the lab stage relative to workers not yet exposed to the experiment. These exploratory effects are significant at the 10 percent level after multiple-testing adjustment, providing suggestive evidence of short-run persistence. Together, these findings offer guidance for designing pay-for-performance contracts to increase their effectiveness and cost-effectiveness.

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

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

This paper was first published in January 2024 as "Pay-for-Performance Contracts in the Lab and the Real World: Evidence from Nigeria." The original version is available here.

Information, Loss Framing, and Spillovers in Pay-for-Performance: Experimental Evidence from Health Workers in Nigeria

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

Eeshani Kandpal
Center for Global Development and World Bank
Email: ekandpal@cgdev.org

Sebastian Bauhoff, Eeshani Kandpal. 2026. "Information, Loss Framing, and Spillovers in Pay-for-Performance: Experimental Evidence from Health Workers in Nigeria." CGD Working Paper 767. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/pay-performance-contracts-lab-and-real-world-evidence-nigeria

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. Brian Webster, 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 (Bauhoff, 2018). Errors and omissions are our own.

Replication data are available at https://doi.org/10.7910/DVN/XVH3US

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

Center for Global Development. 2026.

The Center for Global Development works to reduce global poverty and improve lives through innovative economic research that drives better policy and practice by the world's top decision makers. Use and dissemination of this Working Paper is encouraged; however, reproduced copies may not be used for commercial purposes. Further usage is permitted under the terms of the Creative Commons License.

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors, funders of the Center for Global Development, or the authors' respective organizations.

1 Introduction

Pay-for-performance (PFP) contracts are widely used in settings with principal-agent problems, including health service delivery, where they aim to align worker effort with desired outcomes (Miller and Babiarz, 2013; Lazear, 2000; Prendergast, 1999). These contracts typically involve a checklist of targeted actions combined with financial incentives assigned to each action. Such designs engage agents through two primary channels. First, the checklist provides a salient signal of what the principal values, potentially guiding agent behavior through intrinsic motivation, effort allocation, or reduced uncertainty (Arrow, 1963). Second, the financial incentives can directly influence the agent's effort and task prioritization (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). These channels may interact as, for instance, incentives can redouble the effect of the checklist by making its content more salient. Moreover, PFP contracts are often incomplete or time-limited, which can produce spillovers or persistence effects on behaviors outside the incentivized tasks (Celhay et al., 2019; Sherry, 2016; Miller and Babiarz, 2013).

Further, while standard prospect-theoretic intuition predicts stronger effects under penalties than rewards due to loss aversion, several lab and field experiments find that penalty frames do not consistently outperform reward frames once expectations and reference points are equilibrated (Gneezy et al., 2011; Fryer et al., 2012; de Quidt, 2018). Consistent with these findings, de Quidt et al. (2017) highlight that reference points, task characteristics, and the presence of intrinsic motivation can attenuate the advantage of loss framing. Research generally concludes that loss aversion does not consistently enhance incentives (de Quidt, 2018; Milstein et al., 2016; Conrad, 2015), with implementation details often driving outcomes (Kristensen and Jørgensen, 2017). In the health sector, evidence from both high-income and low- and middle-income country (LMIC) settings suggests that the relative effectiveness of reward versus penalty framing is limited or context-dependent (Kairies-Schwarz et al., 2020; Kim et al., 2020; Navathe et al., 2019).

Despite substantial research on PFP, key questions thus remain regarding the relative importance of and interaction between financial incentives and informational primes, the persistence of incentive effects into real-world tasks, and potential spillovers onto non-incentivized tasks. Health

workers, for example, may respond strongly to a checklist even without financial rewards, but it is less clear how such priming interacts with reward or penalty frames, and whether effects of small incentives continue after the contract ends. Empirical evidence from LMICs indicates that PFP can improve adherence to clinical guidelines and quality of care, but spillover effects and persistence remain understudied (Celhay et al., 2019; Diaconu et al., 2021; Basinga et al., 2011).

We address these gaps using a lab-in-the-field experiment with health workers in Nigeria, linked to direct observation of their clinical practice in routine care.[1] We randomize participants to one of three lab arms—information-only (priming), rewards, or penalties—where the rewards and penalties arms are isomorphic.[2] We then observe a subset of participants providing real-world care after the lab experiment ends. Together, these two stages let us separate four things that are usually bundled in PFP contracts: the effect of financial incentives, the effect of priming alone, the effect of reward versus loss (penalty) framing, and whether incentive effects persist and spill over onto actions the contract does not directly pay for. Our design also allows us to test whether misaligned incentives produce harmful overprovision or whether such responses are mitigated by intrinsic motivation and professional norms (Lopez et al., 2022; Das et al., 2016; Jones et al., 2023; Banerjee et al., 2023).

We first conduct a comprehensive survey of all 1,359 maternity health workers in 690 randomly selected primary health clinics across six Nigerian states.[3] From this sample, 1,359 workers are randomized into a lab-in-the-field experiment with three arms: information (priming without incentives, i.e., just a checklist), rewards, and penalties.[4,5,6] In all arms, participants review medical records of five hypothetical labor-and-delivery patients and identify clinically necessary actions. All receive a participation fee and a checklist of seven common clinical actions that are a subset

[1] See Section 2 and Figure 1 for details of the study design.

[2] In other words, the rewards and penalties arms are designed to be behaviorally equivalent, with the same actions leading to equivalent financial losses or gains.

[3] For details on the context, see Section 2 and **??**

[4] Assignment was stratified by facility to ensure no more than one worker per arm per facility; in clinics with fewer than three workers, all eligible participants were included.

[5] The remaining 67 workers completed the survey but were not included in the lab experiment.

[6] As detailed in Section 3 and **??**, the lab experiment does not include a pure control arm; performance in the incentivized arms is benchmarked against the information-only arm.

of all clinically necessary actions needed for the hypothetical record review task.[7] For five of the seven checklist actions, financial incentives are provided in the rewards and penalties arms, which are isomorphic: participants gain or lose the same amount depending on their arm. To equalize expected payouts, the penalties arm starts with a higher endowment from which incorrect answers reduce pay, while the rewards arm starts lower but increases with correct answers. Average payouts are 1,842 Naira in the rewards arm and 1,818 Naira in the penalties arm.[8]

Then, we assess real world care provision by these health workers within a 72-hour window after the conclusion of the lab experiment. For this assessment of real world care, we randomly select a third of all facilities (230 out of 690, for budget reasons) and observe care provided by a maximum of two health workers on call at each facility that day.[9] We observe the real world provision of antenatal care by these health workers. Antenatal care shares many common clinical actions with labor and delivery, thus allowing us to assess the extent to which the PFP contract "sticks" after the incentives have ended. The real-world analysis compares workers exposed to the experiment before observation with workers not yet exposed at observation, including the 67 workers not assigned to the lab experiment and 49 assigned workers observed before their lab exposure.

We report on four key findings from the lab experiment and subsequent observation of real world care. First, participants in the rewards and penalties arms performed 7.2-7.9 percentage points (pp), or about 20 percent, better than those in the information arm. The incentive effect is notable in health care, where workers may be intrinsically motivated (Mohanan et al., 2021; McGuire, 2000; Arrow, 1963) and respond to interventions that highlight the importance of effort, even without incentives (Gauri et al., 2021; Brock et al., 2018; Leonard and Masatu, 2017; Ashraf

[7]We assess performance using the WHO Essential Procedures Protocol, measuring performance as the share of clinically necessary actions correctly identified out of 15 potentially applicable actions.

[8]The flat participation fee in the information arm was 1,750 Naira (USD 5.80), while rewards and penalties arms were structured to replicate endowment effects observed in typical public sector PFP contracts (Fritsche et al., 2014); see Section 2 for details.

[9]This gives us a subset of 400 participants from the universe of 1,359 workers surveyed earlier. Of these 333 also subsequently participated in the lab experiment, while 67 were not selected for the lab experiment. 49 workers participated in the lab experiment after the real-world observations; we consider these as pure-control because they have not been exposed to incentives or primes. We conducted the basic health worker survey for them.

et al., 2014; Kolstad, 2013). Second, despite a higher reference point in the penalties arm, we find no impact of loss framing relative to reward framing, consistent with loss-neutral agents.[10] Third, we find positive spillovers of incentives on two unincentivized actions: relative to the information arm, rewards improve performance by 4.2 pp (15 percent) and penalties by 3.9 pp (14 percent) for actions not on the checklist. Such spillovers may arise due to production complementarities (Sherry, 2016; Mullen et al., 2010). Finally, participants randomized to financial incentives perform 16.3 to 17.8 pp (21-26 percent) better than pure control on real world clinical actions, indicating that incentive effects can persist into real-world tasks even after the contract has ended. However, we do not find a statistically significant difference between the incentivized and information arms in the real world.

Pairing a lab-in-the-field experiment with real-world observation has several advantages for isolating the impacts of key design elements of PFP contracts because many real world PFP programs include additional reforms, such as operating budgets, clinic management improvements, and accountability mechanisms (de Walque et al., 2022; Diaconu et al., 2021; Basinga et al., 2011). Lab experiments provide a controlled environment to disentangle contract design: (i) tasks allow precise performance assessment against all relevant clinical actions and (ii) priming some actions and incentivizing others allows estimation of spillovers on unprimed or unincentivized tasks. In addition, our experiment examines responses to deliberately misaligned incentives causing potentially harmful treatment. While such actions occur in LMICs (Lopez et al., 2022; Das et al., 2016), providers may be constrained by intrinsic motivation, prosociality, altruism, or reputation. We find muted responses, suggesting limited overprovision of care. This aligns with emerging evidence that prosociality mediates PFP responses (Jones et al., 2023) and that clinicians are not purely payoff-maximizing, due to norms, cognitive constraints, or bounded rationality (Banerjee et al., 2023). Finally, real-world observations allow examination of persistence of priming and incentive effects after the contract ends (Celhay et al., 2019; Fryer et al., 2022).

[10]Note that the penalties arm has a 2.5 times higher reference point with payouts decreasing with incorrect performance while rewards have a lower reference point that increases with correct performance. This difference equalizes expected payouts across arms and keeps actions isomorphic. Appendix **??** details payouts for each action.

We also contribute by (i) providing evidence from rural Nigeria, complementing higher-income studies, and (ii) separating persistence of incentives from the information channel, i.e. does tying a payment to a reminder help the reminder stick longer? Even in real-world observations, participants in the incentives arm outperform the pure controls randomized out of the lab. This suggests that, in our context, financial incentives persist even after they have ended (Celhay et al., 2019).

Although our lab experiment involves hypothetical tasks, participants behave realistically (Harrison and List, 2004; Prendergast, 1999). We use small but meaningful incentives (max USD 6, 5% of monthly salary), tasks mimic routine maternity care, and study sites are participants' primary workplaces. Performance aligns with real-world primary health care in LMICs with workers simultaneously providing far from all necessary care while also providing some unnecessary care (Kruk et al., 2016). Performance correlates with separate knowledge assessments, but incentives provide additional gains. Participants engage on the margin even in the information arm, with the largest differences at mid-performance levels, indicating targeted effects. Further, clinically necessary but unincentivized actions are frequently identified, and incentives influence subsequent real-world care. These results are robust and capture more than knowledge alone. Finally, although the experiment technically admits a dominant, payoff-maximizing strategy of selecting all actions, participants generally do not play it, leaving roughly a quarter of the maximum payout on the table on average; as in classic dictator-game settings, such deviations are informative of underlying preferences rather than a design flaw, and——conditional on health workers having some regard for patient health and for doing things well, alongside financial motives——help explain spillovers onto unincentivized actions. (See Section 4.5 and **??** for detailed discussions of these caveats.)

Our findings have four implications for PFP design.[11] First, PFP has long been used in high-income health systems (Campbell et al., 2007; Mendelson et al., 2017) and increasingly in LMICs, where low-quality care partly reflects low effort (Leonard and Masatu, 2010; Das et al., 2008; Leonard et al., 2007). Second, penalties, while less common, are nevertheless used such as in

[11]Our PFP contract pays for inputs and uses a threshold design where payments are conditional on achieving a goal, mimicking many LMIC programs (Kandpal, 2016). Rewarding inputs or outputs interacts with worker characteristics to influence effectiveness (Mohanan et al., 2021). Results may apply only to input-based, threshold PFP schemes.

the United States Medicare Nonpayment Program (Gupta, 2021). Our evidence adds to literature on limited loss framing effects (de Quidt et al., 2017; de Quidt, 2018; Kairies-Schwarz et al., 2020; Kim et al., 2020). Third, small incentives outperform information-only interventions, with little cost to adopting rewards frames (Rowe et al., 2005). Fourth, we provide evidence on spillovers from incomplete contracts: existing literature finds weak or mixed spillover evidence in health care (Campbell et al., 2007; Diaconu et al., 2021; Celhay et al., 2019; Sherry et al., 2017), but we find positive spillovers consistent with production complementarities (Holmstrom and Milgrom, 1991; Sherry, 2016).

2 Context

This section describes the study setting, experiment design, and measures of worker performance, details of which are provided in **??**.

In 2017, when this experiment was conducted, Nigeria's maternal mortality ratio was 1,127 per 100,000 live births—the second highest in the world and almost 10 times higher than India's (World Bank, 2020). The delivery of health services has stagnated in the last three decades, and the performance of public health workers 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 evaluation began in July 2014 and randomized 52 districts in three states to either performance-based financing (PFP), with quarterly bonuses tied to the quantity and quality of services delivered, or direct facility

financing (DFF), which disbursed a comparable fixed amount without conditioning on performance; three additional states matched by observables served as a "business as usual" control. The larger trial found that both PFP and DFF improved coverage outcomes (e.g. immunization, contraceptive use) relative to control, with comparable effects across the two arms, although gains in clinical quality of care were limited. We use this design as a robustness check: our estimated incentive effects are unaffected by controlling for facility-level assignment to the PFP trial's control, DFF, or PFP arms, and worker-level awareness of the concurrent program does not moderate our results, suggesting our findings are not driven by prior exposure to performance-based financing. This larger trial is described in further detail in **??**.

This survey was conducted between July and October 2017. During the survey, we conducted our lab experiment in 690 of the 691 primary health facilities that were randomly selected for the endline survey of the impact assessment.[12] At each facility, the survey conducted detailed interviews with all 1,359 health workers who routinely provided antenatal or curative care under-five years of age in these clinics (see Figure 1 for more details). If a health worker sampled for the interview was not on duty on the day of the survey, the team returned to survey them the next day that they were on duty. As part of this survey, we administered in-depth interviews and knowledge tests to each of them. The lab experiment was conducted at the end of this interview, and all interviewees who routinely provide antenatal care were eligible for the lab experiment. In facilities with three or more eligible health workers, treatment was stratified by facility to have a maximum of one health worker per lab-experiment arm in each facility. In facilities with fewer than three eligible workers, all workers were randomized to one of the arms of the lab game. To minimize spillovers, the experiment was conducted simultaneously among all participating health workers within a facility.

Following the lab experiment, we observed care provision by a randomly-selected subset of the study sample. These observations were of routine interactions with actual patients within 72 hours after the lab experiment; the survey teams returned to the facility to conduct these observations. Sub-sample selection was as follows. Based on budget considerations, we first randomly selected a

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

third of all health clinics in the study, i.e. 230 out of 690 and observed care. Then we randomly selected a maximum of two health workers per clinic for the assessment of real world care. This gives us a subset of 400 health workers— the maximum permitted by the survey budget— from the 1,359 surveyed workers. Of these 400 workers, 333 had previously participated in the lab experiment while the remaining 67 had not been exposed to the lab experiment. However, they had all received the same basic survey during the survey team's prior visit to the health clinic. In all 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 allow us to estimate the persistence of the PFP incentives in the real world stage of our study.

2.2 Lab experiment

Before turning to the mechanics, it is useful to preview the design: all participants see the same five hypothetical patients and the same list of 15 possible clinical actions per case. What varies across the three arms is only whether, and how, a subset of those actions is highlighted on a checklist and tied to money.

During the health worker survey, we randomized all health workers in each of 690 clinics 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.[13] The assessment tool we used to measure performance, called a partograph, provided some information about the necessary care required in each case (see **??** for details). This design feature thus precluded the inclusion of a pure control arm in the lab experiment. However, Section 3 discusses how we attempt to estimate the effect of information compared to no lab intervention.

[13]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 necessary or unnecessary based on an international standard checklist of essential procedures published by the World Health Organization. They concurred about all actions in all cases.

(As described above, the subsequent real world stage of the study includes a randomly assigned pure control arm.)

All participants received a printed checklist of seven clinical actions. These seven were randomly drawn from a fixed universe of 15 possible actions, and all 15 were later offered as response options when participants reviewed the five hypothetical patient cases. (See **??** in **??**). In other words, all 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 as clinically necessary.[14] In the penalties arm, we deducted the same payment amount for each action that is clinically necessary but was not identified by the participant. The rewards and penalties 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 not clinically necessary, 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 clinically necessary. In the three "complex" cases, the participant is in charge of the patient and is asked to name all actions that she deems clinically necessary for that patient. **??** in **??** presents all possible actions for the three complex tasks and notes whether a given action is clinically necessary. Which actions are clinically necessary varies across cases, but all participants are presented with the same cases. Participants' responses are correct when they choose clinically necessary actions, or when they decline to choose actions that are not clinically necessary. We then calculate scores as the share of correct responses to all 15 actions presented in

[14] 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.

each case.[15] Participants score perfectly when all of their chosen actions are clinically necessary and none of their declined actions are. We report scores "across cases", calculating the proportion of correct responses for each of the five cases as well as the average of all five cases. As the number of clinically necessary actions varies across the different cases, we also report scores aggregated "across responses", i.e. the proportion of correct responses for each individual action in all cases. In this way, we report results that equally weight each case and equally weight each individual item response. We report scores in two ways:

- **Across cases**: the proportion of correct responses within each of the five cases, averaged across all five cases (each case weighted equally).
- **Across responses**: the proportion of correct responses for each individual action, pooled across all cases (each action weighted equally, which matters because the number of clinically necessary actions differs by case).

The next three paragraphs describe, in turn: why we use a participation fee rather than a flat wage, how that fee differs by arm, and a worked example of how the same action is scored under rewards versus penalties. 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. Our use of a participation fee is designed to mimic the fixed-wage baseline that public-sector health workers effectively treat as an endowment. This allows the experiment to capture how workers respond to gains or losses relative to an expected reference point, which is central for interpreting behavior under real-world PFP schemes. Because most large-scale PFP programs in LMICs must be budget neutral, incentives typically reallocate, rather than add, resources to providers (Fritsche et al., 2014). In such settings, endowment effects and loss framing become highly relevant channels through which incentives may operate, making it important that our design reflects this institutional feature.

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

[15] In four instances, actions can be ambiguous, i.e., they can be unnecessary or harmful to patients. We treat these as not clinically necessary. See **??** in **??** for further details.

2,500 Naira and could lose up to 1,500 Naira (USD 5) for a minimum payout of 1,000 Naira.[16] The incentives for individual actions in the rewards and penalties arm varied between 50 and 300 Naira (USD 0.17 to USD 1). As discussed above, whether an action is clinically necessary and hence qualifies for a reward, varies from case to case. An example illustrates: a referral is a clinically necessary action to take in case 2, and any worker in the rewards arm who selects that would receive the payment associated with that action (300 Nigerian Naira) in the rewards arm and incur no penalty in the penalties arm. However, the same action is not clinically necessary in case 3 and, if selected, would not qualify for a reward and would incur a penalty, respectively.

As the above example highlights, the rewards and penalties contracts were isomorphic: all actions lead to the same gains or losses across the two types of contracts. Moreover, the prices, and maximum and minimum payouts are constant across incentivized arms. When we contrast the two frames for the same experimental task using isomorphic contracts, albeit with differing reference points, we find that rewards and penalties have similar direct and spillover effects. The different reference points arise because we designed the expected payouts to be equal across the two arms. The penalties arm has a 2.5 times higher reference point that decreases with incorrect responses, while the rewards arm has a lower reference point that increases with correct responses.

We compensated participants in the form of cellphone airtime after they had completed all five cases. **??** 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.[17]

Table 1 presents the prices and proportion of clinically necessary responses for each action,

[16] 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 penalties arm, the instructions explicitly stated that participants stood to lose part of their participation fee.

[17] 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.

disaggregated by whether the action is paid, primed, or unprimed.[18] 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 clinically necessary: monitoring contractions and the amniotic fluid, tracking the fetal heart rate and the mother's vital signs, and recording the fluids and drugs administered. Performance is also generally better for actions that are not clinically necessary in all cases, such as administering magnesium sulfate or augmenting labor.

2.3 Health worker performance in the real world

In the real-world stage of our study, we assess whether the lab incentives affected actual patient–provider interactions for antenatal care. These interactions involved real effort and real stakes and occurred within 72 hours of the lab experiment, after all incentives had ended. Because the real-world encounters were not incentivized, this stage allows us to test whether any effects of the lab incentives persist in the very short run and provides an additional check on the validity of the experimental design. We compare real-world performance for providers assigned to either an incentivized arm or the information arm, focusing only on actions that closely map onto the lab tasks and remain relevant for antenatal care: monitoring fetal condition, conducting essential blood and urine tests, and documenting all care. This yields our estimate of the short-run persistence of incentives relative to information alone.

An important caveat in interpreting these results is that while our lab experiment assesses performance on maternity care for labor and delivery, the real world care observed is for antenatal visits. This discrepancy arises because antenatal care visits are more common, and hence more

[18] **??** in **??** disaggregates performance on each of these actions for the three complex cases and Figure 4a and Figure 4b presents the empirical cumulative distributions for the complex cases only. As these results show, our findings are robust to excluding the simple cases.

feasible to collect data on. Conversely, 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. See **??** for a detailed discussion of the validity of the lab experiment.

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 WHO standard of care for antenatal visits. As detailed in **??**, this includes two actions that were 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 issued hard copies of the checklist to keep after the lab experiment.

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. As discussed earlier, in 67 instances, the participants of the lab stage experiment were not at work on the day of the observation and were replaced with the health worker responsible for antenatal care on the day of the survey. Our final sample includes 333 directly observed health workers assigned to participate in the lab experiment and 67 additional workers who were not assigned to participate. Of the 333 assigned participants, 49 were observed before exposure to the experiment. We therefore classify 116 workers as not exposed at observation and 284 as exposed at observation in the real-world analysis.[19] The two groups also have similar characteristics (Table 2).

[19]The sample size varies slightly across specifications as we exclude actions that are reported to

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 incentivized arms. Overall performance is low: on average in the information arm, participants correctly identified only about half of clinically necessary actions and correctly avoided only about half of clinically unnecessary actions. 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 incentivized arms also reveal how much money participants forwent by failing to identify clinically necessary actions or avoid clinically unnecessary ones: 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 responding correctly. However, in terms of assessed covariates, all three treatment arms are balanced, consistent with successful randomization of workers into the three lab arms.

---

have been performed outside of the exam room, prior to the observed interaction. Moreover, **??** in **??** shows that participants in the lab 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 lab experiment are robust to only estimating impacts for the subset of participants who were observed providing care (**??** in **??**).

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 lab experiment. The assessed covariates are all balanced. Further, **??** in **??** tests for balance among lab participants selected for direct observation in the subsequent real world 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 is consistent with successful cross-randomization for direct observation in the real world stage.

3 Empirical strategy

We use the following OLS model to assess the impact and persistence of PFP incentives in the randomized lab experiment and cross-randomized real world stage treatment assignments. 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 standard errors clustered at the facility level.[20] We assess robustness with additional regressions that control for the participant-level covariates listed in Table 2. For Table 4 and Table 5, we additionally report p-values adjusted for multiple hypothesis testing using the Romano-Wolf step-down procedure. Rather than correcting across all tests jointly, we define families by the substantive claim each set of tests supports

[20] Regressions at the item level are clustered at the worker level but are robust to being clustered at the facility level.

(e.g., direct effects on paid actions; spillovers on unprimed actions; persistence in the real world), and report family-wise adjusted p-values alongside unadjusted p-values within each table.

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 arms (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 by itself. However, we can obtain a rough benchmark of this effect by comparing performance on primed actions with performance on unprimed actions in the information arm. There are two important caveats in interpreting these estimates. 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. Signing the direction of the effect of such spillovers, we posit that 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. 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 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:

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

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 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 4a and 4b, equally weighing cases and responses yields three findings. First, the range of observed performance is comparable across all arms. Second, the two incentivized 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 penalties 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 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.[21] For both measures, rewards and

[21] In **??** in **??**, we explore whether workers of different characteristics respond heterogeneously

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 laid out in the framework presented in **??**.

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 **??** in **??**. In contrast to the 8 percent effect on overall performance above, paid actions specifically show a larger gain: performance in the two incentivized 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 in the information arm, 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. (See **??** in **??** for the detailed results underlying this table.)

Turning to clinically unnecessary care (columns 6-7 in Table 4), we find suggestive evidence that incentives decrease the provision of unnecessary actions. In the rewards arm and relative to the information arm, workers are 2.8 pp (about 3.5 percent) less likely to correctly avoid an unnecessary action, though the impact of penalties is not robust to Romano-Wolf multiple hypothesis correction. The impact of penalties is not statistically significant. In all arms, performance on unnecessary actions is substantially higher than on necessary actions. For example, in the information arm workers correctly avoid unnecessary actions 91.5 percent of the time, compared to correctly identifying necessary actions only 28.2 percent of the time for unprimed correct actions.

to performance pay, as in Donato et al. (2017). The high performers in our 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.

This suggests that professional norms and intrinsic motivation constrain overtreatment even in the presence of financial incentives. The modest and largely insignificant effects on unnecessary actions, set against the 20 percent gain on paid correct actions, are consistent with providers responding selectively to incentives rather than simply doing more of everything.[22]

4.2 Spillovers 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 clinically necessary actions: relative to 28.2 percent correct performance in the information arm, performance is about 4.2 pp (15 percent) higher in the rewards arm and about 3.9 pp (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 clinically unnecessary 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 **??**, we interpret positive spillovers from incentives as consistent with complementarities across actions.

As the spillovers are largely attributable to performance on (1) measuring the mother's vital signs, and (2) the rate of descent of the fetal head, it may be that features of these tasks drive the spillovers and that a large set of unprimed actions would yield a substantively different estimate of spillover effort. For instance, these tasks may differ from others in salience, difficulty, or baseline frequency, and these differences could plausibly contribute to their greater responsiveness. As such it is important to note that although these patterns are consistent with complementarities between the incentivized behaviors and related clinical actions, the evidence should not be generalized beyond this subset of tasks.

[22]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 clinically necessary and unnecessary actions are also substantively different.

4.3 Suggestive analysis of price response of effort

As a secondary and more exploratory analysis, we examine how effort varies with the price attached to an action (**??** and **??** in **??**). We caution that these patterns should not be interpreted causally: prices were not randomized across tasks, and higher prices were purposively assigned to more complex actions. As a result, observed performance reflects both the incentive level and the intrinsic characteristics of each action, and our estimates rely on the assumption that non-price attributes are constant across study arms and can be differenced out—an assumption that may be reasonable in a setting where performance requires only identifying the clinically necessary action, but one that does not fully overcome the fundamental limitation of non-randomized pricing. With this caveat in mind, the observed pattern is that moving from zero to a small positive price is associated with most of the apparent performance gains under both rewards and penalties, with little additional association at higher prices, in both direct and spillover effort. One possible interpretation is that incentives may chiefly operate by heightening attentional focus on the underlying clinical objective or by transiently increasing motivation to meet perceived expectations, rather than by paying for marginal effort; ceiling effects, short-term learning dynamics, or residual priming from the lab task offer alternative explanations that we cannot rule out. We note, suggestively, that this pattern of apparently limited responsiveness to price is consistent with related field evidence that healthcare providers in LMICs do not fully exploit available financial incentives (Banerjee et al., 2023).

4.4 Effect of incentives on subsequent real world patient interactions

Next, we assess the persistence of experimental incentives in real world interactions conducted after the lab experiment. Of the six actions observed, two were 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) but were assessed in the experiment as well. The results in Table 5 show that exposure to an incentivized arm increases real-world palpation by 16.3 pp (21 percent) and fetal-heartbeat monitoring by 17.8 pp (26 percent) relative to workers not yet exposed to the experiment. The adjusted p value is 0.057 for both outcomes, robust to dropping singleton facilities. We do

not find a statistically significant direct contrast between the reward and penalty arms (adjusted p=0.105 for palpation and 0.278 for fetal-heartbeat monitoring); given that we did not pre-specify an equivalence margin, we treat this as a failure to reject rather than evidence that the two arms have equal effects. We similarly do not find a robust statistically significant direct contrast between the information and incentive arms (p=0.392, 0.157 unadjusted; 0.365, 0.269 adjusted).

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 those real world actions that are 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 incentivized arms may do better than information even after the PFP contract ends. We do not detect an incremental effect of incentives over information on primed-but-unpaid actions. This does not establish equivalent effects or rule out persistent spillovers.

Among the unprimed actions, we find no significant effects of incentives or information on blood pressure measurement, urine testing, or the vaginal exam (columns 7–12 of Table 5). For the vaginal exam specifically, we note that the exam is only clinically necessary for patients with a history of vaginal bleeding in a prior pregnancy. Controlling for this patient risk factor, we find broadly similar null results across both subgroups, though estimates are imprecise (Appendix **??**).

That even priming and incentives provided in the context of a lab experiment can have persistent effects on real world care suggests that multiple pathways affect worker performance, including after the incentives end. While previous evidence has shown that very large temporary incentives can persist even after the cash runs out (Celhay et al., 2019), our study suggests that lab-based incentives can have meaningful effects on real-world performance, even after the formal lab intervention has ended (Brock et al., 2018). While the lab tasks involve hypothetical patients and controlled settings, the behaviors practiced appear to translate to actual care. This indicates that carefully designed lab games can act as behavioral triggers or small nudges, influencing motivation and effort beyond the immediate incentive. In particular, performance improvements are observed

despite the absence of continued financial rewards, suggesting that effects are not driven solely by extrinsic monetary incentives.

Given the short window between the lab experiment and real world observation, we interpret improvements in real-world performance as evidence that lab-based incentives can have a short-term influence on provider behavior. Our use of the term "persistence" refers to the continued influence of the incentive rather than long-term behavioral change. As the real world assessment occurs within 72 hours after the lab experiment, alternative explanations are plausible. For example, providers may have been responding to Hawthorne effects arising from recent observation, rather than internalized or stable changes in behavior. We therefore interpret these effects as suggestive of incentive persistence, but note that further inquiry is needed to fully assess the extent to which small experimental incentives may persist.

4.5 Validity of the lab experiment, contamination, 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 tests of validity and robustness that we conduct on our estimated impacts. Tests of the validity of the lab experiment and robustness checks are presented in detail in **??**.

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 **??**, 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).

Another concern regarding the validity of the lab task pertains to our use of a dominant strat-

egy. In other words, the healthcare workers in the penalties and rewards arms had the incentive to indicate all incentivized actions, even ambiguous or potentially harmful ones, as there was no cost or effort associated with indicating an unnecessary action. This is the low-effort, payment maximizing dominant strategy. In fact, however, we find that participants did not play the experiment's dominant strategy of selecting all actions. Incentivized participants left about a quarter of the maximum payout on the table, and only 2.5% achieved the maximum payout. As in classic dictator-game settings, such deviation from a payout-maximizing strategy is itself informative, suggesting participants weigh accuracy and professional norms alongside payoffs. Indeed, conditional on health workers having some regard for patient health—altruism—and for doing things well—prestige—in addition to financial motives, the dominant strategy is unlikely to explain spillovers onto unincentivized actions. This is consistent with incentivized arms showing higher performance and more time-on-task than the information-only arm, and with participants being more reluctant to add a clinically unnecessary action than to omit a necessary one. We discuss the implications of this setup feature in detail in **??** and argue that dominant strategies are a standard feature in many lab games, including the widely used dictator games.

We also expect contamination to be minimal. Within facilities, all workers completed the lab experiment simultaneously and in separate rooms, limiting the scope for treatment–control information exchange. Contamination could generally only arise later, during real-world observations, if treated workers shared what they learned or earned with untreated colleagues. Such sharing could bias our persistence estimates in either direction: if treated workers shared what they learned, this would improve control performance and bias impacts on real-world care toward zero, whereas if treated workers shared that they had been paid for the lab task, this could discourage untreated colleagues and lead us to overestimate persistence. Across facilities, contamination is unlikely given the substantial geographic dispersion of clinics (median distance 7 km; mean 13.5 km) and their large, distinct catchment areas. Moreover, any cross-facility spillovers would affect all study arms similarly. Overall, we believe neither within-facility channel is likely to be large, and contamination does not appear to pose a meaningful threat to internal validity. (See Appendix B.4 for details.)

We also perform a battery of robustness checks that are also detailed in **??**. The first set of checks shows that results are robust to various alternative formulations of the lab-in-the-field

performance score: considering the simple and complex cases separately (Figure 4a and Figure 4b and columns 3–7 of Table 3), measuring performance in z-scores instead of as a proportion (**??** in **??**), and using an Item Response Theory aggregated outcome score (**??** in **??**) 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 (**??** in **??**). 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 (**??** in **??**). 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 the magnitude of the incentive.

We report on a randomized lab experiment and cross-randomized subsequent observation of real world care designed to examine the effect of PFP incentives relative to treatment with only a 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. First, we conduct 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 of care for hypothetical patients. Then, we conducted a cross-randomized assessment of routine care provision by these health workers to real world patients after the end of the lab experiment.

We find that incentives matter above and beyond information in the lab and at par with information in 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, because 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.

Our findings add to prior evidence suggesting that information (Kolstad, 2013) or even measurement (Leonard et al., 2007) alone can improve care quality. The distinction lies in the nature and visibility of the information. In studies like Kolstad (2013), providers receive feedback on their performance relative to peers, triggering social comparison or reputation concerns. In our experiment, performance information is private; there is no explicit peer benchmarking. Instead, the lab primes, coupled with small incentives, highlight professional norms and the provider's own standards, leveraging intrinsic motivation. This framing suggests that even when explicit information is limited or private, nudges can encourage the internalization of high-quality behavior and meaningfully influence short-run real-world performance.

Two related, non-mutually-exclusive mechanisms may help explain why lab-based incentives persisted in real-world care days after the incentives ended. First, the lab experiment's primes, linked to even modest incentives, may have made providers more aware of the standards expected in their professional role, leading them to internalize and carry forward high-quality behaviors (Brock et al., 2018). Second, engaging with the lab tasks may have strengthened providers' identification with their own competence and responsibility, consistent with the view that internalized professional norms shape real-world behavior (Leonard and Masatu, 2010). Given our relatively short 72-hour observation window, we interpret these as short-lived behavioral triggers rather than evidence of durable change, but the pattern suggests that pairing salient primes with modest incentives may have effects that outlast the incentive itself.

Finally, our design and findings speak to the role of intrinsic motivation, altruism, and reputa-

tion 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., 2022; 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 reputation 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 2.8 pp (3.5 percent) increase in unnecessary actions in the incentivized arms relative to the information arm. This finding suggests that over-treatment may be somewhat limited in practice, perhaps via considerations like intrinsic motivation, reputation concerns, and altruism. In the context of the conceptual framework presented in **??**, 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.

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 effects 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-the-envelope calculation in **??** suggests that a PFP with our estimated effects is roughly 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 use this insight to increase the cost-effectiveness of PFP contracts.

Future research could build on these findings along several fronts. First, it will be important to understand why health workers leave money on the table even in simplified laboratory settings, where payoffs are transparent and effort costs are minimal; identifying whether bounded rationality, attention limits, fatigue, or professional norms drive these choices could help refine the design of incentive schemes. Second, future work could examine the external validity of these behaviors by linking lab-based measures to real-world performance data, including how providers respond

to more complex or ambiguous incentive environments. Finally, integrating these insights into policy experiments that vary incentive design, information framing, and administrative complexity would allow researchers to test which features of performance-pay contracts most effectively elicit high-quality care without imposing excessive cognitive or administrative burdens on health workers.

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.

A. Banerjee, A. Chowdhury, J. Das, J. Hammer, R. Hussam, and A. Mohpal. The market for healthcare in low income countries. *Harvard Business School*, 2023.

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.

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.

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.

D. A. Conrad. The theory of value-based payment incentives and their application to health care. *Health Services Research*, 50(S2):2057–2089, 2015.

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.

J. de Quidt, F. Fallucchi, F. Kölle, D. Nosenzo, and S. Quercia. Bonus versus penalty: How robust are the effects of contract framing? *Journal of the Economic Science Association*, 3(2):174–182, 2017.

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.

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

R. G. Fryer, S. D. Levitt, J. List, and S. Sadoff. Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. Technical report, National Bureau of Economic Research, 2012.

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, S. Meier, and P. Rey-Biel. When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, 25(4):191–210, 2011.

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. B. Jones, M. Tonin, M. Vlassopoulos, and K. P. Winichakul. Paying for what kind of performance? performance pay, multitasking, and sorting in mission-oriented jobs. *Games and Economic Behavior*, 142:480–507, 2023.

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.

N. Kairies-Schwarz, M. Schreyögg, and T. Stargardt. The effects of bonus–malus incentives on physician behavior. *Health Economics*, 29(2):177–192, 2020.

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.

H. Kim, A. Ryan, et al. Effects of penalty-based pay-for-performance: A systematic review. *Health Services Research*, 55(3):395–405, 2020.

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

T. Kristensen and S. S. Jørgensen. Design and implementation of penalty-based incentives in health care. *Health Policy*, 121(5):451–457, 2017.

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.

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*, 14(1):225–260, 2022.

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.

A. Milstein et al. Pay-for-performance in healthcare: History, evidence, and future directions. *Medical Care Research and Review*, 73(4):437–441, 2016.

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.

A. S. Navathe et al. The impact of financial incentive framing on physician behavior: A randomized clinical trial. *JAMA Network Open*, 2(8):e198722, 2019.

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.

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

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

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.

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

Figures and Tables

Figure 1: Study design

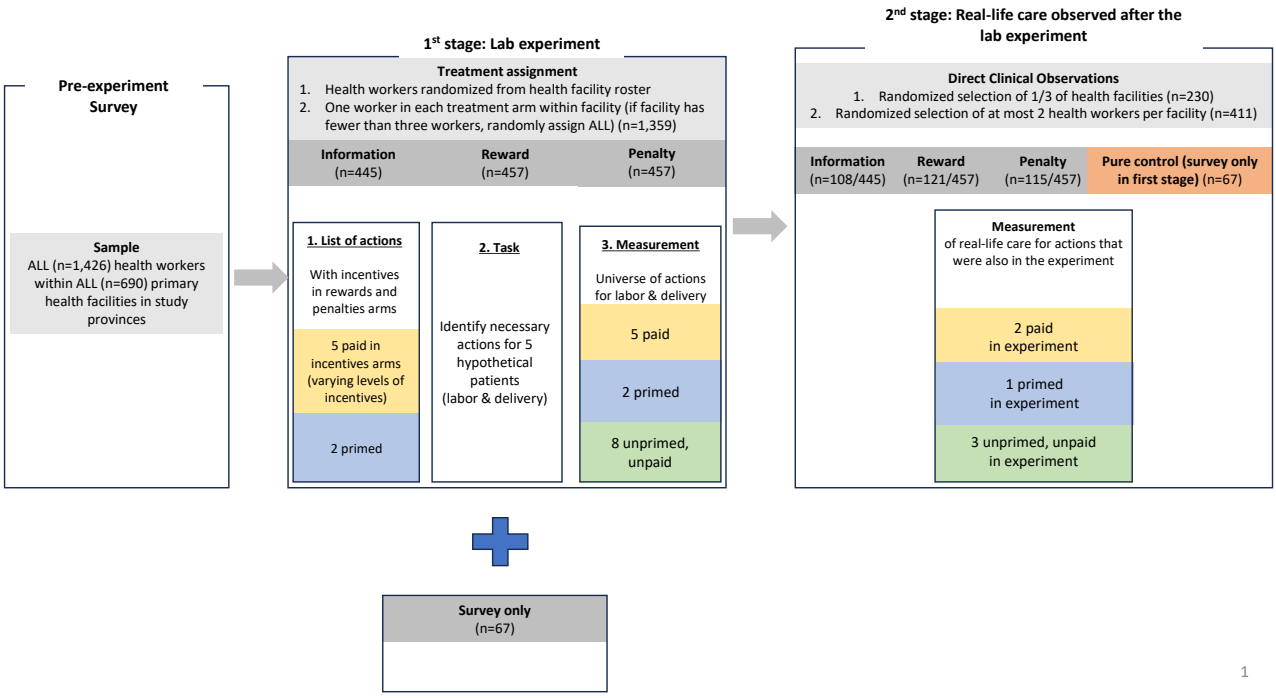


Figure 2: Detailed design of lab experiment

1. Instruction sheet and list of actions (provided to by each participant)

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.

Action	**Information arm** N = 445 *We appreciate your help in examining these partographs and would like to offer 1,750 Naira as a thank-you.*	**Rewards arm** N = 457 *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.*	**Penalties arm** N = 457 *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.*
Refer to secondary facility when necessary		300	300
Measure fetal heart rate at least every 30 minutes		100	100
Monitor contractions every 30 minutes		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

Eight additional unprimed actions were recorded but not shown to participants.

2. Five tasks presented to participants (same cases for all participants)

Two partographs of simple hypothetical cases

Based on Ms. X's partograph, your colleague suggests that you do [one specific task that is either correct or incorrect, randomized within study arm]. *Based on the time and stage of Ms. X's pregnancy and the information provided in the partograph, do you believe this action is clinically indicated at this time?*

[Correct / Incorrect / Not sure]

Three partographs of complex hypothetical cases

Assume Ms. Y is your patient. What clinically indicated actions would you take now, based on the state of the pregnancy outlined in Ms. Y's partograph?

Interviewer had a list of 15 explicit options, plus 'other', that were not read aloud.

Interviewer marked all actions mentioned by the participant.

3. Assessment (same scoring scheme for all participants)

Actions considered

- Five paid and two primed that were on the instruction sheet.
- Eight unprimed that were not on the instruction sheet.

Scoring

- Share of clinically indicated actions correctly identified, out of all indicated actions.
- Share of unnecessary actions that the participant did not indicate.

2nd stage Real-life direct clinical observations

Figure 3: Detailed design of real-life care observed after the lab experiment

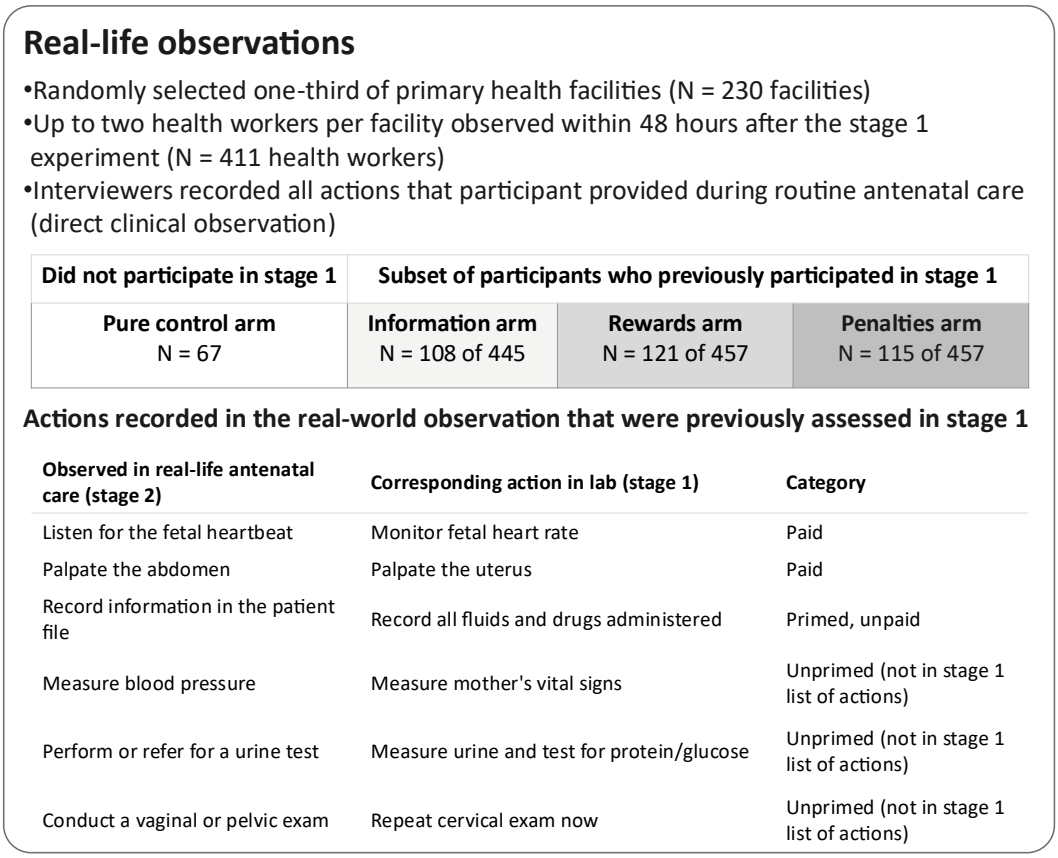
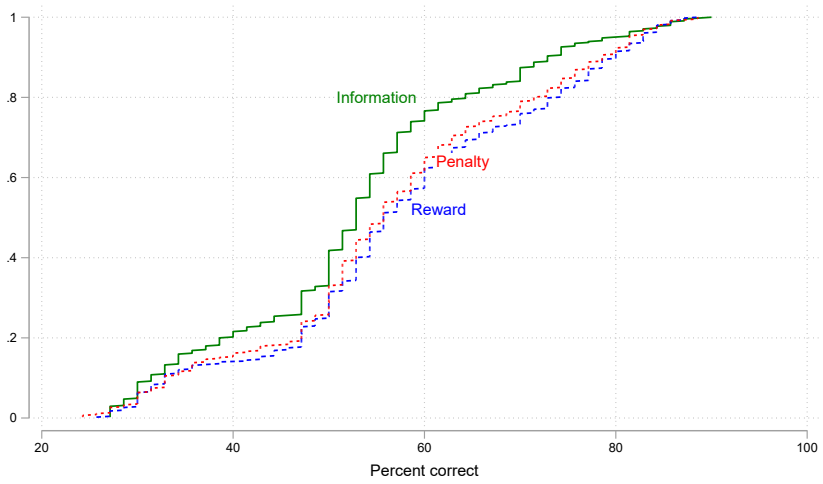
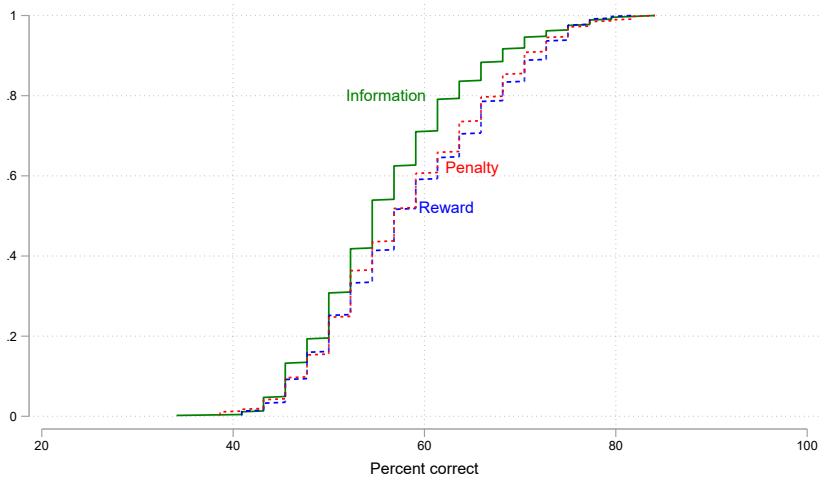


Figure 4: Empirical cumulative distributions



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

		Percent correct			
	Incentive (Naira)	All arms	Information	Reward	Penalty
Listed and 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
Listed but unpaid					
Monitor color and consistency of liquor		16	15	16	18
Record fluids/drugs administered		6	6	6	7
Unlisted					
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.

Table 2: Summary statistics and balance across the lab experiment and real world care provision stages

Variable	(1) Information Mean/SE	(2) Reward Mean/SE	(3) Penalty Mean/SE	T-test Difference (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 score	-0.11	0.04	0.07	-0.15*	-0.18***	-0.03
	(0.04)	(0.04)	(0.05)			
Total payout (Naira)	1,750.00	1,841.90	1,817.94	-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)			
Age (years)	38.01	38.24	38.09	-0.23	-0.08	0.15
	(0.45)	(0.45)	(0.43)			
Doctor or nurse	7.42	10.94	10.28	-3.53	-2.87	0.66
	(1.24)	(1.46)	(1.42)			
Qualified (years)	10.85	10.90	10.74	-0.05	0.11	0.15
	(0.39)	(0.41)	(0.37)			
Knowledge score	52.78	53.03	52.46	-0.25	0.32	0.57
	(0.76)	(0.75)	(0.76)			
NSHIP pilot status						
PFP	44.49	42.67	42.45	1.82	2.04	0.22
	(2.36)	(2.32)	(2.31)			
DFF	42.25	45.08	43.11	-2.83	-0.86	1.97
	(2.34)	(2.33)	(2.32)			
Control	13.26	12.25	14.44	1.00	-1.18	-2.19
	(1.61)	(1.54)	(1.65)			
N	445	457	457			

Notes: †Across cases weighs each case equally; across responses weighs each response equally. The median age is 38.0; the median years-since-qualified is 9.5; and the median for knowledge is 51.5 percent. The value displayed for t-tests are the differences in the means across the groups. Standard errors are robust. Fixed effects using variable facid are included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 3: Overall performance (percent correct)

	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	4.36***	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 clustered standard errors.

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
p-value (RW-adj.)	–	0.02	–	0.00	–	0.58	0.09
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
p-value (RW-adj.)	–	0.03	–	0.00	–	0.30	0.57
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 **??**. Based on the three complex cases in the lab experiment. Estimated for 1359 maternity health workers in 690 primary health clinics in Nigeria.

p-value (RW-adj.) reports Romano–Wolf familywise-error-adjusted p-values (`mhtreg`; List, Shaikh, and Xu, 2019; 5,000 bootstrap replications, seed 02138), computed *within* families defined around each of the three substantive claims we make about the lab experiment. Each family and the tests it contains are: (i) *direct effect of incentives on paid actions* — Reward−Information and Penalty−Information on paid necessary actions (2 tests); (ii) *spillovers* — Reward−Information and Penalty−Information on unprimed necessary actions (2 tests); (iii) *overprovision* — Reward−Information and Penalty−Information on paid unnecessary actions and on unprimed unnecessary actions (4 tests). Adjusted p-values are reported hypothesis by hypothesis; because the step-down procedure enforces monotonicity, two hypotheses in the same family can share an adjusted value. These families were constructed for this revision and were not pre-specified in the trial registration (AEARCTR-0002482). Cells marked (–) are not members of any of the three families and are reported unadjusted: the two aggregate columns (All; Primed or paid), which combine actions that appear separately in the families; primed-but-unpaid actions; and the Penalty−Reward contrast, for which we report only that rewards and penalties cannot be statistically distinguished. Both primed-but-unpaid actions (monitoring color and consistency of liquor; recording fluids and drugs administered) are clinically necessary in all three complex cases (see Table A.3), so no primed action is ever clinically unnecessary in our sample; hence, there is no "unnecessary primed" column to report.

Table 5: Effect on behaviors in direct clinical observations of antenatal care (percent)

	Related to listed and paid actions				Related to listed actions		Related to unlisted 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	10.92	8.79			8.28*		-11.87	4.00	-7.23			
	(10.39)	(7.35)			(4.83)		(13.62)	(7.87)	(4.77)			
Incentive	17.80**	16.28***			5.08		-4.78	1.10	-7.29			
	(7.55)	(6.00)			(3.27)		(10.23)	(3.62)	(5.40)			
Information			11.08	8.85		8.29*				-12.04	4.07	-7.24
			(10.31)	(7.27)		(4.85)				(13.55)	(7.85)	(4.78)
Reward			21.61**	19.75***		5.21				-14.21	3.95	-8.06
			(9.26)	(6.93)		(3.88)				(10.54)	(4.63)	(6.29)
Penalty			13.32*	11.95**		4.93*				7.00	-1.38	-6.36
			(7.98)	(5.40)		(2.76)				(11.84)	(4.05)	(5.31)
Constant (Naive)	69.28***	77.80***	69.41***	78.00***	92.29***	92.30***	39.47***	90.39***	11.37***	38.88***	90.33***	11.36***
	(5.46)	(4.25)	(5.45)	(4.17)	(2.55)	(2.54)	(7.38)	(2.74)	(3.63)	(7.31)	(2.80)	(3.62)
Facility fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P-value Info v Incentive	0.392	0.157			0.271		0.468	0.692	0.984			
P-value Penalty v Reward			0.335	0.070		0.884				0.028	0.239	0.715
P-value Info v Reward			0.273	0.085		0.350				0.832	0.986	0.824
P-value Info v Penalty			0.782	0.492		0.225				0.101	0.486	0.818
N respondents	400	399	400	399	399	399	392	343	400	392	343	400
R-squared (overall)	0.018	0.027	0.011	0.019	0.004	0.004	0.002	0.000	0.000	0.011	0.004	0.000

* p<0.10, ** p<0.05, *** p<0.01. OLS models with facility fixed effects and clustered standard errors. Related to actions that are listed and paid in the experiment. Models for vaginal exam control for patient risk (bleeding in prior pregnancy).