Health Workers' Behavior, Patient Reporting and Reputational Concerns: Lab-in-the-Field Experimental Evidence from Kenya*

Isaac Mbiti[+]
University of Virginia

Danila Serra[±]
Texas A&M University

This draft: March 2021

Abstract

We examine the effectiveness of accountability systems that rely on patient reporting in Kenyan health clinics. Patients and health care providers from public and private health clinics participate in a lab-in-the field experiment focusing on the relationship of trust between patient and provider. Patients decide whether to trust providers, providers have discretion over their reciprocity, and patients can complain. We compare the effectiveness of: 1) a client reporting system where patients' complaints are disclosed to the providers' professional peers, possibly leading to non-monetary penalties, 2) a system where complaints lead to monetary penalties, and 3) a system that, like a standard complaint box, attaches no tangible consequences to complaints. Overall, our findings suggest that citizen reporting systems that leverage peer pressure and reputational concerns can improve service delivery.

JEL Classification: C90, I15, M59

Keyword: Health services, bottom-up accountability, patient reporting, peer shaming.

* We thank Willa Friedman, Guy Grossman, Pam Jakiela, Molly Lipscomb, Tim Salmon, Dan Silverman, Andreas Stegmann for useful comments and suggestions. We are also grateful to seminar participants at George Mason University, Southern Methodist University, the University of Virginia, Virginia Commonwealth University, Utah State University, and participants at the Texas Experimental Symposium, the NEUDC, and the NOVAFRICA conference. The surveys and the experiments were expertly managed by Ron Wendt through Innovations of Poverty Action (IPA). Both Southern Methodist University IRB and IPA IRB reviewed the protocols. Funding for this project was provided through a World Bank Trust Fund on Accountability in Public Services and Southern Methodist University. Daniel Rodriguez-Segura provided additional research support. We thank Gabriel Demombynes, Aaron Thegya and Chris Finch for their support throughout this project.
[+] Email: imbiti@virginia.edu.
[±] Email: dserra@tamu.edu.

1. Introduction

Accountability is often cited as a necessary factor to ensure the efficient provision of public services in developing countries (World Bank, 2004). Low levels of accountability in the health sector are associated with low levels of effort by healthcare providers, resulting in high absence rates ranging from 20 to 40 percent on a given day (Chaudhury et al., 2006 and World Bank SDI, 2013), limited patient interactions (Das et al., 2016), and low quality of service relative to providers' demonstrated ability (Leonard et al., 2007 and Das and Hammer, 2014). Since the cost of monitoring service providers may be significantly cheaper for citizens than for government agents, well-designed bottom-up monitoring schemes could improve service delivery by enhancing accountability (World Bank, 2004). By relying on informal or non-monetary sanctions, they may also be easier to implement than other alternatives, such as top-down accountability systems, especially in countries characterized by weak state and limited bureaucratic capacities.

We test the effectiveness of an accountability system that relies on non-monetary sanctions and leverages providers' reputational concerns by conducting a specially designed lab-in-the field experiment with actual health providers and patients in Nairobi, Kenya. We recruited patients randomly from a set of about 1800 surveyed patients exiting public and private frontline health facilities (health centers and dispensaries). We then invited a subset of providers from the same facilities to participate in behavioral games with the patients. In designing a lab-in-the-field experiment that could parallel some aspects of the patient-provider relationship, our starting point was the recognition that clinical interactions are characterized by information asymmetries, uncertainty (for example, the provider may be absent from work), and uneven distribution of powers. These factors suggest that the level of trust between patients and providers will capture important aspects of clinical interactions. Patients place their trust in providers when they visit a health center. Providers can then reciprocate a patient's trust by being present at the facility and by providing high-quality service. Although there are additional aspects of clinical interactions beyond trust, there is a growing body of literature that shows that patient's trust (or mistrust) in medical providers is a critical determinant of health utilization, health seeking behavior, and ultimately health outcomes in both developed and developing countries (Alsan and Wanamaker, 2017; Alsan, Garrick, and Graziani, 2019; Archibong and Annan, 2021; Lowes and Montero, 2020; Martinez-Bravo and Stegmann, 2020).

These considerations led us to employ a modified trust game (Berg, Dickhaut and McCabe, 1995; Serra et al. 2012) where patients can send money to providers and providers can reciprocate by sending money back. Another reason for employing a trust game setting – as opposed to a dictator game, for instance – is that receiving money from the first mover (the patient) in the game generates a feeling of obligation in

the second-mover (the provider) to send money back, which we believe could capture a provider's contractual obligation to provide care to the patients visiting his or her facility.[1]

Contrary to a standard trust game, in our Reporting Game, we allow patients to file costly complaints against providers if they perceive the amount returned to be too small. We employ different treatments where we modify the consequences of such complaints, with our primary treatment of interest being one where complaints lead to non-monetary sanctions in the form of peer shaming activated through the disclosure of complaints to other providers (Peer Disclosure treatment). Our Monetary Penalty treatment, where patients' complaints lead to a monetary loss for the provider, provides a useful benchmark to assess the effectiveness of the Peer Disclosure system,[2] even though monetary punishment systems are unlikely to be implemented in real life.[3] We contrast both treatments to a baseline treatment where complaints have no tangible consequences (Complaint Box treatment).[4] We also examine whether the possibility of provider retribution against complaining patients negatively affects complaining behavior and reduces the efficacy of non-monetary sanctions (Peer Disclosure with Retaliation treatment). We use the amount of money that providers send back to patients (scaled by the initial transfer amount) as our primary measure of health-worker behavior toward patients. In addition, we use the patient's decision to file a complaint as a function of the amount returned by the provider as our main measure of the patient's willingness to use a reporting system. In light of the growing literature showing that public providers and private providers have different traits and motivations (Cowley and Smith, 2014; Serra, Serneels and Barr, 2011; and Brock, Lange and Leonard, 2016), we also examine the potential for differential responses to accountability schemes across public and private sector health providers.

Our study relates to the small but growing literature that employs lab-in-the-field experiments to generate direct measures of health providers' motivations and responsiveness to different incentive systems in developing countries (Banuri and Keefer, 2016; Barr, Lindelow and Serneels, 2009; Brock et al. 2016; Serra et al., 2011). Contrary to the existing studies, in which providers are the only decision-makers in the experiment,[5] we examine the strategic interactions between providers and actual patients under different

[1] In a validation study with patients in Senegal, Kovacs, Lagarde, and Cairns (2019) show that trust games are have high construct validity. They argue that these games can better measure trust in health providers compared to traditional survey measures.

[2] While there is evidence that monetary penalties are highly effective – and more effective than informal punishment (Masclet et al., 2003) – in lab settings, only a handful of lab experimental studies employ games that introduce monetary punishment in the context of a trust game that resembles our reporting game (Calabuig et al., 2013; Fehr and Rockenbach, 2003; Houser et al. 2008; Rigdon, 2009). We are unaware of any trust games investigating the effectiveness of second-movers' "complaints" while manipulating the monetary or non-monetary consequences attached to such complaints.

[3] Although a reporting system leading to monetary penalties is unlikely to be implemented in the field, it can provide some suggestive insights into the efficacy of a system where customers can punish a provider by switching to another provider, such as under a voucher scheme. In our experiment, the size of the monetary penalty was approximately 10 to 60 percent of providers' daily wage depending on rank and seniority.

[4] Given their ubiquity in public facilities, like government health clinics, the Complaint Box treatment is arguably the most relevant benchmark since complaint boxes are the status quo system.

[5] Patients are either passive participants, as in a standard dictator game, or they are not part of the experiment.

incentive systems. We also contribute to a large body of evidence from experiments in both the laboratory and the field testing whether informal mechanisms, such as social observability and public disclosure of performance rankings, can promote prosocial behavior, even in the absence of formal incentives. The evidence is mixed, with some studies (e.g., Andreoni and Petrie, 2004; Ashraf et al. 2014, Carpenter and Myer, 2010; Charness et al., 2014; Della Vigna et al., 2012; Erikson et al., 2009; Gerber et al. 2008; Gill et al. 2017; Karlan and McConnel, 2014; Linardi and McConnel, 2011) showing positive effects and other studies (e.g., Brent et al., 2019, Cason et al., 2014; Charness et al., 2014; Dufwenberg and Muren, 2006; Noussair and Tucker, 2007) finding null or even negative impacts on behavior.[6]

In our Peer Disclosure Treatment, we inform providers that we will share the number of complaints a provider receives, a measure of patient dissatisfaction, with their peers. This differs from other studies which typically focus on sharing information on provider behavior or performance directly with clients. This way, we simulate and evaluate a bottom-up accountability system that could easily be implemented by non-governmental organizations, as it does not require the actual measurement and quantification of providers' performance. In a related study, Ashraf et al. (2014) also compare non-monetary and monetary incentives in a health context. In their study of sales of condoms by hairdressers in Zambia, the authors examine the impact of publicly displaying a "thermometer" that reports condom sales. Our non-monetary accountability mechanism differs from Ashraf et al. (2014) in a variety of ways. First, our Peer Disclosure treatment generates comparisons across providers based on negative rather than positive measures of performance, therefore relying on the activation of non-monetary costs rather than non-monetary rewards. Second, our system relies on subjective evaluations by clients (i.e. complaints) rather than objective performance measures, such as sales.[7] Third, our treatment simulates a mechanism where patients' complaints are only disclosed to providers and not to the public. Hence, our Peer Disclosure mechanism works purely through the activation of reputational concerns (or fear of shaming) among professional peers (the providers), absent any concern about the impact that information about relative performance could have on existing or potential clients.[8]

Through our Peer Disclosure with Retaliation treatment we also add to the literature on harassment and retaliation, which has shown that fear of retribution prevents subjects from engaging in costly punishment or reporting of others both in lab (Abbink and Sadrieh, 2009; Abbink et al., 2014; Balafoutas et al.,

[6] For instance, Charness et al. (2014) show that performance ranking increases performance in a real-effort task under flat-rate compensation. However, when subjects in a group are able to cheat or sabotage others in order to inflate their position in the ranking, they tend to do so.

[7] This is especially meaningful in sectors such as healthcare where performance is linked to the delivery of a service behind closed doors rather than the exchange of a product of fixed quality and price.

[8] The literature of online customer feedback in standard markets in the developed world suggests that ratings significantly affect sales (Anderson and Magruder, 2012; Cabral and Hortacsu, 2010; Chevalier and Mayzlin, 2006) and prices (Houser and Wooders, 2006; Resnick et al., 2006). However, contrary to standard online rating systems, the mechanism we investigate is one where negative ratings are only shown to providers.

2014; Denant-Boemont et al., 2007; Nikiforakis and Engelmann, 2008) and field settings (Balafoutas and Nikiforakis, 2012). Our lab-in-the-field environment is especially interesting because, while it relies on a laboratory experiment, it allows for the norms and expectations that guide the longer-term relationships existing between patients and providers in the field to play a role in the complaining decisions patients make within the experiment.

Our findings show that, relative to a simple Complaint Box, the Peer Disclosure system increases provider reciprocity in the game by 42 percentage points, or approximately 40 percent. The possibility of provider retaliation reduces the effectiveness of the Peer Disclosure reporting system to 26 percentage points relative to the Complaint Box (approximately 25 percent). However, the coefficients are not statistically different from each other. In addition, we find that provider reciprocity tends to be more responsive to the Peer Disclosure system compared to the Monetary Penalty system, although we cannot reject the hypothesis that they are equally effective. We also find that public sector health workers are more generous toward patients and more responsive to our treatments compared to their private sector counterparts. This is consistent with a growing empirical literature (Banuri and Keefer, 2015; Serra et al., 2012; Delfgauuw et al., 2013; Gregg et al., 2011; Kolstad and Lindkvist, 2012) that finds significant differences between public and private sector employees with respect to their prosocial motivations and their responsiveness to monetary and non-monetary incentives. In addition, we find that less prosocial health workers, as measured in a standard Trust Game that preceded the Reporting Game, are more responsive to the threat of both peer shaming and monetary sanctions.

With respect to the complaining decisions, on the intensive margin, patients submit more complaints under all systems that attach tangible consequences to the complaints, relative to the Complaint Box baseline. On the extensive margin, the propensity of patients to file unfavorable reports is lower when complaints lead to tangible monetary or non-monetary consequences for providers.

A note on the external validity of our experiment is warranted. We took a series of measures to enhance the extent to which behavior in the experiment could be informative of behavior in real world settings.[9] First, we conducted the lab-in-the-field experiment with actual health providers and patients rather than university students. Second, we conducted the experiment in a developing country context, where bottom-up accountability is often touted as a potentially cost-effective way to improve service delivery. Third, we involved a representative sample of patients, randomly selected from a larger set which was interviewed while exiting health centers. Although our primary research focus is on examining the behavioral responses to our treatments, we acknowledge that the decisions providers and patients made in the context of our experiment differ from outside-the-lab decisions in a variety of ways. First, due to the transparency of each

[9] See Kessler and Vesterlund (2015) for a general discussion of the external validity of the qualitative results generated by laboratory experiments.

player's actions in the reporting game, patients are able to clearly evaluate and identify the provider behavior that they deem deserving of a complaint. Although this may be more difficult for patients in real life, it is exactly this design feature that allows us to clearly measure patients' *willingness* to file complaints when reporting systems are available. Our study is more relevant for designing policies aimed at curbing observable provider misbehaviors, such as absence from work and request for bribes, which are commonplace in developing countries (Chaudhury et al., 2006, Transparency International, 2011). As a partial test for the external validity of our methodology, we conducted unannounced visits at health facilities and correlated absence data with our lab measure of provider performance. We find that providers – and especially public sector providers – who return more money to patients in our lab-in-the field experiment are less likely to be absent from work, which suggests that our lab measures capture salient aspects of real life behavior.

Overall, our study provides important insights and guidance regarding the design of better accountability systems, as it assesses some of the central assumptions and hypotheses behind these systems. These include the willingness of citizens to actively use them and the responsiveness of providers to patient feedback in a controlled environment. Given the significant fraction of GDP devoted to personnel in the public healthcare sector, improving provider behavior, especially the quality of service, can have significant implications on public finance management.[10] Our findings show that reporting systems that trigger social sanctions through disclosing providers' negative feedback to their professional peers have the potential to significantly and positively affect providers' performance. As many developing countries are actively considering accountability systems that rely on social sanctions, these design insights could be especially relevant in the current policy climate.[11]

2. Research Methodology and Procedures

As part of the study, we conducted: 1) a survey of public and private health facilities in Nairobi; 2) a survey of patients exiting these facilities; 3) a survey of health providers from the same facilities; and 4) lab-in-the-field experiments with a subset of health providers and surveyed patients. In this section we start by describing the Kenyan context (Section 2.1) and explaining the study sampling and survey procedures (Section 2.2). We then provide details about our lab-in-the-field experiments, outline our hypotheses (Section 2.3), and explain how the experimental workshops were implemented (section 2.4). We conclude this section by describing our estimation strategy (Section 2.5).

[10] See Das, Holla, Mohpal, Muralidharan (2015) for a discussion on the fiscal costs of teacher absence in India.
[11] For example, Tanzania's Big Results Now program relies on the dissemination of public sector performance rankings to district officials, summarized in three color bands (red, orange, and green) to improve service delivery (Cilliers, Mbiti and Zeitlin, forthcoming)

2.1 Context

Health expenditures in Kenya grew from 4.5 percent of GDP in 2000 to 5.7 percent in 2014, where public health expenditures drove the majority of spending increases (World Bank WDI, 2017). At the same time, key measures of health have improved. The under-5 mortality rate decreased from 100 per 1000 births in 2000 to 53.5 per 1000 births in 2014 (World Bank WDI, 2017,). Moreover, over the same time span, life expectancy in the country has increased from 51.7 years at birth to 66 years at birth (World Bank WDI, 2017).

There are three types of health facilities in Kenya: dispensaries (or clinics), health centers, and hospitals. Dispensaries and health centers are the primary health care facilities in most communities. Dispensaries offer basic curative and preventative services and are staffed by nurses and community health workers. In contrast, health centers offer more advanced curative and preventative services and are staffed by registered clinical officers, nurses, and lab technicians. Hospitals provide more specialized curative and referral services for patients who cannot be treated at dispensaries or health centers. They are staffed by a wide array of medical personnel, including surgeons and specialist doctors. Across the country, there are a total over 5,200 dispensaries, 720 health centers, and 450 hospitals, with the private sector accounting for 40 to 55 percent of these facilities (Kenya SPA report, 2011).

The health system in Kenya is plagued with inefficiencies that impede service delivery. Health worker absence is common and more problematic in public health facilities. According to the World Bank's Service Delivery Indicators (SDI), in 2012, almost 30 percent of public health workers were absent on a given day, compared to just over 20 percent of private health workers (World Bank SDI, 2013). These absence rates were higher than teacher absence rates, which averaged approximately 16 percent, suggesting that accountability may be weaker in the health sector (World Bank SDI, 2013). In addition to lower absence rates, private facilities had better infrastructure and equipment, with 85 percent of private facilities having the minimum infrastructure required to be effective, compared to just under 50 percent of public facilities. Adherence to clinical guidelines, based on hypothetical case studies, was also slightly higher in private facilities (48 percent) compared to public facilities (43 percent) (World Bank SDI, 2013). Private facilities also had greater drug availability (80 percent) compared to public facilities (63 percent) (World Bank SDI, 2013). This is consistent with qualitative reports that documented the acute shortages of drugs in public health facilities in Kenya and highlighted the common practice of providers denying patients drugs in order to resell them on the private market (Transparency International, 2011).

Our survey data, collected from 93 randomly selected public and private health frontline facilities (i.e., health centers and dispensaries) in Nairobi County, show similar patterns of service delivery (Table 1). Public facilities in our sample were generally larger, had higher case-loads, and were more likely to have

laboratories, beds, and surgical facilities compared to private health facilities. However, relative to private health facilities, public facilities were less likely to have a complaint box or a published list of prices, which suggests a lower level of transparency and accountability. Patient satisfaction with providers and the overall quality of care was also lower in public facilities. Factors such as longer waiting times in public facilities, even when scaled by patient case-loads, may drive this result. However, public facilities were significantly less expensive than private facilities as a result of government policies aimed at reducing user fees. The differences in costs, waiting times, and patient satisfaction may partly explain why richer and more educated patients visited private facilities.

2.2 Survey design and sampling

We used the Master Facility List from the Ministry of Health's eHealth website[12] to create an initial list of facilities, which we then pared down to include only public, private, and nonprofit /nongovernmental (NGO) dispensaries and health centers in Nairobi County. We chose to focus on dispensaries and health centers, as these are the primary service providers for most Kenyans. These facilities were further screened for their representativeness; therefore, we excluded any facilities that overwhelmingly served only a subset of the population (e.g., mothers), only offered limited services (e.g., immunizations and voluntary HIV counseling and testing) or had staff that was atypical (e.g., mostly medical students). This final list of facilities was then randomized to include an even number of public and private facilities in the study. These facilities were then grouped by geographic proximity to make clusters of three facilities. Each cluster was then randomly assigned a treatment group, as further explained in Section 2.4.

We conducted facility surveys and exit interviews with randomly selected patients in 93 public and private dispensaries and health centers. Our field teams spent about a week surveying each facility and its patients. The facility surveys collected information on infrastructure, management, the staff roster, and staff absence. We conducted 1,784 patient exit interviews. These surveys collected information on health habits, current health, satisfaction with the care they received on the day of the interview and in the past, and socio-demographic information. The staff rosters and patient exit interviews served as our "sampling frames," which were used to randomly invite individuals to what we called community workshops (the lab sessions and additional survey data collection) on the Saturday of the week in which we conducted the exit survey and facility survey. The participating health providers were surveyed after participating in the experiments. Finally, a week after the workshop, we conducted second unannounced visits to the surveyed facilities to record absence of health workers.[13] The structure and timing of the data collection are displayed graphically in Figure A1 in the appendix.

[12] See http://ehealth.or.ke/

[13] During the workshop we recorded each provider's work schedule for the following week and visited the facility at a time where the provider had stated he or she would be at work.

We aimed to have five health workers and ten patients participate in each workshop. We invited individuals to our workshop using a randomly ordered call list from each cluster of three facilities. When possible given the geographical clustering, we selected two public facilities and one private facility to participate in the workshop. We intentionally over recruited individuals to ensure that we had a sufficient number of participants at each session. We selected the first two providers who arrived from each of the two public facilities in the cluster and the first provider to arrive from the private facility. We selected patients in a similar manner, choosing the first four patients who arrived from the first public facility and three patients each from the second public facility and the private facility. If this participant composition was not possible due to insufficient attendance, we used a simpler "first-come, first-admitted" process. We canceled workshops if less than six patients and three providers attended.[14] We paid participants a fee of 300 Kenya shillings (KES) if they were a patient and 500 KES if they were a healthcare worker. In addition to this fee, patients and providers could earn up to 2400 KES by playing the games during the workshops.[15]

2.3 The lab-in-the-field experiment

Patients and providers are seated in different rooms and randomly matched with each other. Each patient is given an endowment E_P and each provider is given an endowment E_H. Patients can send some of their endowment to the matched provider. If a patient sends an amount X, the provider gets three times that, 3X. The provider then decides whether to return any of the received amount to the patient and how much. The amount sent back, Y, is therefore such that $0 \leq Y \leq 3X$. The patient can then file a complaint against the provider at a fixed cost c, which results in the provider receiving a number of cards with frowning faces on them. Once the patient has paid the fixed cost of complaining, they can express their level of frustration by sending up to five frowning face cards to the provider. F denotes the number of cards that a patient sends, and we limit the number of cards that can be sent to five, so $F \in [0,1,2,3,4,5]$.

Patients first decide how much money they wish to send to the matched health provider, and they then decide whether or not to complain about the amount they received back. Patient payoffs are therefore equal to:

$$P_p = \begin{cases} E_p - X + Y, & \text{if the patient does not complain} \\ E_p - X + Y - C(F), & \text{if the patient complains} \end{cases}$$

Health provider payoffs are equal to:

[14] Only one workshop was cancelled due to low attendance.
[15] At the time of the study, the average hourly wage of a nurse in Kenya ranged from 83KES to 656 KES, depending on rank and seniority. Based on per capita GNI, the average hourly wage of a Kenyan was 33 KES.

$$P_H = \begin{cases} E_H + 3\mathrm{X} - \mathrm{Y}, & \text{if the patient does not complain} \\ E_H + 3X - Y - K(F), & \text{if the patient complains} \end{cases}$$

We conducted four treatments in which we varied K(F), i.e., the cost generated by the complaints on the provider.

1. The Complaint Box (CB) treatment simulates a simple reporting system relying on patients' complaints that are privately read by providers. In this treatment, the provider receives an envelope containing the frowning faces (up to five) sent to him or her by the matched patient. No tangible costs to providers are associated with the complaints; therefore, in this setting K(F) is equal to zero, and if the patient complains, the provider's payoffs are identical to the no complaint case;
2. The Peer Disclosure (PD) treatment simulates a system relying on complaints that have no monetary consequences for providers but that are publicly disclosed to all providers participating in the workshop, hence possibly leading to peer shaming. This was achieved by delivering the envelope containing the frowning face cards to the provider, as before, and by displaying the cards received by each provider on the blackboard in the room where all providers were seated. On the board, each provider was identified by a player number given to him or her at the beginning of the experimental session. As under the CB treatment, there are no monetary consequences attached to complaints; therefore, K(F) = 0;
3. The Peer Disclosure with Retaliation (PD-R) simulates a reporting system that is identical to PD but includes the possibility for the provider to retaliate by imposing a monetary penalty on the patient who filed a complaint. This was achieved by allowing providers, after seeing the information displayed on the board, to retaliate on the matched patient by imposing a monetary penalty R. To reflect the power asymmetry that characterizes the provider-patient relationship, we made the imposition of a penalty costless to the provider. The penalty was levied at the payment stage, if any. This modifies the patient's payoffs to: $E_P - X + Y - C(F) - R$, if the patient complains and the provider retaliates;
4. The Monetary Penalty (MP) treatment simulates a system relying on patients' complaints that generate monetary penalties on providers. As in CB, the "frowning face cards" sent by patients are privately seen by providers, but a monetary penalty k for each received card is then levied at the payment stage. This implies that K(F) is now equal to k times F and provider payoffs in case of complaint are equal to: $E_H + 3X - Y - kF$.

Note that in all treatments the complaint cards are disclosed to providers at the very end of the experiment, after all the decisions have been made. This way, we measure whether the possibility of

receiving complaints,[16] with no consequences or different types of consequences attached to them, affects provider decision-making.[17]

Before conducting the Reporting Game described above, we also conducted an incentivized standard Trust Game (TG) with the participants. This "Baseline Trust Game" had identical payoffs and initial decision stages as the Reporting Game, but did not include the patient complaining stage. This game was included to allow the participants to get familiar with the mechanics of a Trust Game before they played the more complex reporting game. It also provides us with a baseline measure of providers' behavior toward patients in the absence of any reporting mechanisms.[18] The overall timeline of the activities and data collected is shown in Figure A1 in the appendix.

2.4 Implementation

We conducted a total of 24 community workshops in local schools, involving 216 patients and 103 health providers, mostly nurses and clinical officers. Each workshop started with registration, where we verified that participants were either a provider or a patient recruited from a given facility. Participants were each given a colored badge (green for providers and orange for patients), and assigned a number (1 to 5 for providers and 1 to 10 for patients). Patients and providers were then seated in separate rooms. For the duration of the workshop we referred to each participant by his or her color and identification number.

At the end of the experimental session and before the payment stage, we implemented network questionnaires with both patients and providers, in order to register the personal and professional relationships between study participants. Finally, we surveyed the health providers to collect demographic data and information about individual job experiences and to record their work schedules for the following week. We used such self-reported work schedules to implement unannounced visits to facilities the week following the workshop and, hence, generate a measure of providers' absence from work.

Each experimental subject participated only in one of the four treatments, i.e., we employed a between-subject design. We conducted at least five sessions of each of our four different treatments of the Reporting Game, as shown in Table 2. For each workshop, we managed to recruit between three to five

[16] While in all treatments mentioning the possibility of complaints could have signaled what was appropriate or inappropriate behavior in the game, such signals are unlikely to be driving treatment differences, given that they were present in all conditions.
[17] It would be interesting to test whether the receipt of complaints affects behavior in subsequent rounds. However, due to their work schedules, health providers had limited availability to participate in our experimental sessions. Thus, we were only able to conduct one-shot games rather than repeated games. Even by employing one-shot games, the workshops lasted about 4 hours. In the Table A.5 of the appendix we use our survey data to conduct some exploratory analysis to examine how actual provider-patient relationships affect the interactions in the game.
[18] Subjects were paid based on their choices in one of the games, which was randomly selected at the payment stage. Subjects were given no feedback after playing the trust game and were informed that in the next game they would be matched with a different participant sitting in the other room.

providers and between seven to ten patients. Table 2 shows the distribution of participants across our different treatment conditions. Facilities were grouped by geographic proximity into clusters of 3 facilities, and each cluster was invited to a workshop (or session). Each cluster (or session) was then randomly assigned a treatment of the Reporting Game. Therefore, the assignment of patients and providers to any given treatment was determined by the random allocation of their health facility to that treatment. We attempted to have one private facility and two public health facilities in each sampling cluster. When this was not possible, we had three public health facilities per cluster and workshop. As a result, the percentage of private health providers in each treatment ranges between 20 and 30 percent.

We set patient and provider endowments, E_P and E_H, each equal to 1000 KES. Patients could send multiples of 200 KES to the matched provider, for a maximum of 800, i.e., $X \in [0, 200, 400, 600, 800]$. All participants knew that a provider would receive triple the amount of money that was initially sent by the patient. Providers had to decide how much they wanted to return to the patient (in multiples of 200). We set the cost of filing a complaint, c, to 50 KES. In the Monetary Penalty treatment, each complaint card, F, led to a penalty of 100 KES for the provider, i.e., k = 100. In the Peer Disclosure with Retaliation treatment, we set the retaliatory penalty, R, equal to 150 KES.

In order to ensure that patients had a good understanding of the rules of the game, we first read the instructions of the game aloud and then conducted one-to-one interviews with each patient – recall that patients and providers were seated in different rooms. The collection of patient data through one to one interviews follows standard practice in the implementation of lab-in-the-field experiments in developing countries involving non-standard populations characterized by low levels of education. In these settings, in order to maximize subject understanding of the rules of the games and the monetary consequences of decision-making in the experiment, it is important to repeat the instructions on a one-to-one basis, provide additional examples, and ask comprehension questions before inviting subjects to make their own choices. During the one-to-one interviews, enumerators repeated the game instructions, asked comprehension questions, and then stood up and distanced themselves from the participants in order to give them privacy when they made their choices.[19]

Given providers' higher level of education, we did not conduct private interviews with them. Instead, we let them make their decisions in private. Moreover, for providers we employed the strategy elicitation method, i.e., we had providers fill out a form stating how much they would like to send back in the case of each possible amount sent to them by the matched patient, before seeing the actual amount sent. The use of the strategy elicitation method ensured perfect comparability across providers as it ensured that each of

[19] While we cannot rule out that the presence of the experimenter in the room where patients made their choices may have influenced their decision-making, we expect experimenters to have had the same impact on patients in all treatments. Thus, under this assumption, our estimated treatment effects would not be affected.

them responded to the same set of possible scenarios. Had their responses been directly elicited, the actual scenario faced by each provider would have varied depending on the amount sent by the matched patient.[20]

We also used the strategy elicitation method to register the complaints of each patient. For each possible amount returned by the provider, the matched patient had to state whether he or she would like to complain and select how many frowning face cards to send to the provider. For instance, if a patient sent over 200 KES, the provider would receive three times that amount, i.e., 600, and the patient would then have to state whether or not he or she would like to spend 50 KES to complain if the amount returned by the provider was 0, 200, 400 or 600. These four decisions to complain were registered during the one-to-one interviews in which field officers went through the previously read aloud instructions, used multiple scripted examples, and tested the patients' understanding with specific comprehension questions before eliciting their complaint choices.

As discussed previously, we first conducted a standard Trust Game (Baseline Trust Game) and then conducted the Reporting Game. This helped promote participant understanding of the games, as we anticipated that explaining all the stages of the Reporting Game would be very difficult without having first ensured that subjects were familiar with the two stages of a standard Trust Game.[21] Participants were not informed about the outcomes of each game until the payment stage at the very end of the workshop, and they were paid the earnings from only one randomly selected game.[22]

2.5 Testable Hypotheses

Theoretically, since filing a complaint is costly (i.e., $C(F) = c > 0$), if individuals are purely money-maximizers we should expect to see no complaints and no differences in providers' behaviors across treatments in our reporting game. In other words, in all treatments we should observe F=0, X=0, and Y=0, and all reporting systems should prove ineffective. However, a large number of experimental studies (e.g., Calabuig et al., 2013; Fehr and Gächter, 2002; Fehr and Fischbacher, 2004; Masclet et al., 2003; Xiao and Houser, 2005) have shown that individuals are willing to incur monetary costs to impose monetary or non-monetary penalties to others and that the possibility of receiving punishment, either formal or informal, significantly affects individual decision-making. It is easy to augment health providers' payoff by including a non-monetary cost, I(F), generated by the receipt of negative feedback by patients, in the form of disapproval cards. In the Monetary Penalty treatment, K(P) would then become equal to $kF + I(F)$, with

[20] Brandts and Charness (2011) reviewed the literature and found that using the strategy method in no cases generated treatment effects that are not also observed when employing the direct-response method.

[21] Note that in each game, subjects from one room (for instance, patients) were matched with a different subject from the other room. Participants were made aware of this feature of the experimental design.

[22] The random selection of the payoff-relevant game happened in front of the participants at the end of the workshop before the payment stage.

$I(F) \geq 0$ and increasing in F. In the other treatments, K(F) would be simply equal to I(F). Our hypotheses follow.

Hypothesis 1: The Peer Disclosure treatment will induce providers to return more money to patients compared to the Complaint Box treatment.

The Peer Disclosure treatment combines non-monetary penalties (as in CB) with a social comparison and peer shaming mechanism. We are unaware of other empirical studies – based on observational or experimental data – testing the effectiveness of accountability systems that rely on individuals' reputational concerns through the disclosure of received complaints to professional peers. However, many laboratory and field experiments have shown that individuals' behavior may respond positively to the possibility of social observability and judgment (e.g., Andreoni and Bernheim, 2009; Andreoni and Petrie, 2004; Ariely et al., 2009; Brock et al 2016; Gerber et al., 2008; Linardi and McConnell, 2011; Xiao and Houser, 2011). Our Peer Disclosure treatment is also related to a number of studies that examine the effect of public disclosure of performance ranking on job productivity and education outcomes (e.g., Ashraf et al., 2014; Azmat and Iriberri, 2010; Blanes i Vidal and Nossol, 2010; Delfgaauw et al., 2013; Eriksson et al., 2009; Gill et al., 2015). The extent to which providers care about peer judgment could be easily reflected in the provider's payoff function by assuming that the non-monetary cost generated by complaints is also a positive function of the visibility of such complaints to peers, i.e., $I = I(F, v)$ where v is equal to zero in CB and MP, since complaints are read in private by providers, but it is equal to 1 in the PD treatment, where complaints are shown to other providers. This implies that patients' complaints generate a higher non-monetary cost in PD as compared to CB.

Hypothesis 2: The Peer Disclosure with Retaliation treatment will induce providers to return lower amounts of money to patients as compared to the Peer Disclosure treatment.

The comparison between Peer Disclosure and Peer Disclosure with Retaliation is straightforward. Since the possibility of retaliation increases patients' costs of complaining, this should lower providers' expectation of receiving a complaint for any given amount Y that they transfer back to patients. As a result, the amount sent back to patients in the PD-R treatment should be lower than under PD.

Hypothesis 3: The Monetary Penalty treatment will induce providers to return more money to patients compared to the Complaint Box treatment.

Hypothesis 3 follows from the fact that in the Monetary Penalty treatment, the cost of receiving patient complaints is equal to kF + I(F), with I(F) ≥ 0 and increases in F. It is therefore larger than the corresponding cost I(F) suffered by providers in the Complaint Box treatment for any positive F.

While the comparison of provider behaviors in the MP and CB treatments is straightforward, we cannot formulate a clear hypothesis with respect to the comparison of the PD and MP treatments. Since, in our framework, the cost generated by complaints to the provider is equal to kF + I(F) under MP versus I(F, v) under PD, the PD treatment would be more effective than the MP treatment if and only if I(F, v) > kF + I(F), i.e., the smaller the monetary penalty k and the higher the provider's sensitivity to peer judgment, $\frac{\partial \mathrm{I(F,v)}}{\partial v}$.

Hypothesis 4: Public sector providers will return more money to patients as compared to private sector providers. In addition, public and private sector providers will respond differently to the accountability treatments.

A growing number of studies such as Brock et al, (2015), Kolstad and Lindkvist (2012), and Serra, Serneels and Barr (2011) have found significant differences in prosocial (or intrinsic) motivation between public and private sector workers. In line with these studies, we expect that private and public providers will differ in their baseline level of reciprocity toward patients. We measure this baseline level by conducting a standard Trust Game (the Baseline Trust Game) before the more complex Reporting Game. In line with the existing literature, in the Trust Game, absent the threat of patient complaints, we expect that relative to private providers, public providers will send back more money to patients. This would suggest that public workers are more prosocial than private workers. In addition, public and private providers may differ in their responsiveness to the accountability treatments. This may be due to differential reputational concerns (for the PD and PD-R treatments) and/or differential marginal utility of money (for the Monetary Penalty treatment).[23]

Our predictions with respect to patients' willingness to file reports against providers are less clear. Laboratory experiments that allow for costly punishment have provided evidence of widespread willingness to punish others because of social preferences or the desire to enforce social norms. For simplicity, we can assume that although filing a complaint is costly (i.e., C(F) = c > 0), individuals may gain a non-monetary benefit from expressing disapproval toward a health professional that returned less than what the patient considers an appropriate or fair amount. The existing literature suggests that a patient would complain when

[23] As we discuss in Section 2.6, we do not have sufficient statistical power to test whether private and public providers differ in their responsiveness to the individual treatments. Rather, we focus on the public versus private differential to the Pooled Accountability treatments, relative to the Complaint Box.

the non-monetary benefit outweighs the cost C(F). Moreover, if the non-monetary benefit of complaining depends positively on the severity of the provider's punishment, we should expect patients to be more willing to complain under Monetary Penalties and Peer Disclosure than the Complaint Box. However, our experimental environment differs from a typical lab setting in several ways. First, in our reporting game, anonymity is partially lifted. Providers and patients see each other before being sent to different rooms to play the game. If they know and recognize each other,[24] their interaction in the lab-in-the-field setting is in fact conditioned by the beliefs, norms, and expectations that guide the longer-term relationships existing between patients and providers in the field. This implies that fear of outside-the-lab retaliation may lower reporting in our experiment compared to standard lab settings.[25] Moreover, in our experiment, the potential "punisher" is not a "peer" of the individual being punished; on the contrary, the parties involved in the game are likely to be separated by a large social status gap. This might, in turn, further reduce the likelihood of patient reporting.

Formally, the above discussion implies that patients may suffer a non-monetary cost when complaining against a provider. These costs are likely increasing in the number of disapproval cards sent (F), as well as the actual cost incurred by the provider in the event of a complaint (K). Since the Monetary Penalty and the Peer Disclosure treatments both increase K, it is possible that patients' willingness to file complaints will be lower in these treatments than in the Complaint Box treatment, due to an increase in patients' non-monetary cost of complaining. Given that our MP and PD treatments are likely to increase both the non-monetary benefit and the total cost of filing a complaint, relative to the CB treatment, we cannot formulate a clear hypothesis on citizens' willingness to complain under these two treatments.

2.6 Empirical Strategy

As treatments are randomly assigned at the session level, we can estimate the treatment effects of the interventions using OLS specifications to examine both patient and provider behavior in our experiment. Since all participants in the same session face the same accountability system, we cannot include individual- or session-level fixed effects, as they would absorb the treatment assignment. For providers, we focus on the amount of money sent back to patients as our primary outcome measure and use the following OLS specification to estimate the treatment effects:

$$y_{ik}^{r} = \beta_0 + T_i'\beta_1 + X_i'\beta_2 + v_i{}'\beta_3 + \varepsilon_{ik} \qquad (1)$$

[24] This is not always the case, as the providers that were sampled (and who participated) from a given facility might not be the ones patients interacted with the day of the exit interview. Since our study included a post-experiment survey component that recorded personal relationships among participants, we are able to control for such relationships in the empirical analysis.
[25] See Balafoutas and Nikiforakis (2012) for an example of field punishment, or lack thereof, in Athens subway stations.

where y_{ik}^{r} is the ratio of the amount provider *i* returned (or sent back) to the matched patient and the amount the patient had sent, $k \in [200, 400, 600, 800]$. Since k was tripled between being transferred to the provider, y_{ik}^{r} is a number between 0 and 3, with 0 indicating that the provider kept the tripled amount for himself/herself and 3 indicating that he or she sent the full tripled amount back to the patient. In the next sections, we refer to this outcome variable as the returned-received ratio (R/R). T is the vector of binary treatment indicators, with the Complaint Box condition serving as the reference (or omitted) group. In our analysis we also include specifications where we pool our accountability treatments (Peer Disclosure, Peer-Disclosure with Retaliation, and Monetary Penalty) to increase our statistical power. X is the vector of controls, which includes demographics (age and gender), whether the provider works in the public or private sector, whether the provider is a clinical officer or a nurse, the number of patients that the provider recognized at registration, and whether the provider knows one or more other providers participating in the experimental session.

We estimate our regressions using the full set of decisions elicited through the strategy method (see section 2.4). Therefore, each provider has four observations in the data documenting how much money they would return to the patient for each possible amount they could receive from the patient (i.e., 200,400, 600 or 800 KES). We include a set of binary variables (v_i) to control for each of the initial amounts possibly sent by the patient. As participants in a session may face similar shocks, we cluster our standard errors at the session level to account for this possibility. Since there are fewer than 30 clusters, we follow Cameron, Gelbach and Miller (2008, 2015) and use wild-cluster bootstrapping to account for the relatively small number of clusters and report the p-values of the bootstrapping exercise, as well as (regular) clustered standard errors in our results.[26] We also explore the heterogeneity in treatment effects by interacting our treatment variables in equation (1) by a binary variable indicating if the provider is from the public or private sector. We further examine the heterogeneity in treatment effects by provider baseline reciprocity in the standard Trust Game. In these (heterogeneity) specifications, in order to increase our statistical power, we focus on a Pooled Accountability treatment variable, which is equal to one for any reporting system that attaches consequences to complaints.

For patients, we explore their initial sending behavior and their subsequent complaining decisions using OLS specifications. We examine patients' initial sending behavior using the following specification:

$$y_j^s = \delta_0 + T_j'\delta_1 + X_j'\delta_2 + e_j \qquad (2)$$

[26] The wild bootstrap procedure was implemented using the boottest command in Stata (Roodman et. al 2019).

where y^s is the amount sent to providers by patients or a binary variable indicating that the patient sent the provider a positive amount, T is the vector of treatments indicators, and X is the vectors of controls such as demographic variables. We also examine specifications where we report the results for the pooled treatments relative to the complaint box.

As the strategy method elicited patients' complaining decisions (see section 2.4), each patient has several observations depending on the amount they initially sent.[27] We use the following OLS specification to examine the patient complaining decisions:

$$C_{jm} = \gamma_0 + T_j'\gamma_1 + X_j'\gamma_2 + u_{jm} \quad (3)$$

where C is a binary variable indicating whether patient *j* filed a complaint if the provider sent back *m* KES (*m* is discrete). T and X are defined as above. In exploratory work, which we report in the Table A.5 of the appendix, we examine the heterogeneity in treatment effects by the patient's familiarity with (at least one of) the providers participating in the workshop. To do this, we interact the Pooled Accountability treatment variable with a binary variable that indicate whether the patient knows a participating health worker.

Since the decision to file a complaint is conditional on sending a positive amount to the matched provider, our analysis of patients' propensities to complain applies only to the 164 "trusting" patients (out of 216) who decided to send a positive amount to the matched provider. We test for balance in patient sending behavior across the treatments in order to clearly interpret the results in equation 3. Finally, we also examine the number of complaint cards sent by patients as an additional outcome (conditional on filing a complaint).[28]

3. A First Look at the Data

3.1 Patients and Workshop Participants

We surveyed a total of 1,784 patients from public and private facilities. We randomly invited a subset of surveyed patients to participate in the community workshop. A total of 216 patients participated in the workshops. Although we aimed to have ten patients and five providers per workshop, fewer than ten patients participated in 37% of our workshops, and less than five providers participated in 50% of our workshops.

[27] For instance, if a patient sent 200 KES to the matched provider, the patient had to state whether she would like to send up to five cards displaying a frowning face to the provider in each of the following cases: 1) if she received 0 back; 2) if she received 200 back; 3) if she received 400 back; 4) if she received 600 back. If a patient sent 400 KES to the provider, the complaining decision would apply to seven possible scenarios, each corresponding to the provider returning a different amount, in multiples of 200, up to 1200 KES. We use the same methodology in registering complaining decisions when the patient sends 600 and 800 KES to the provider.

[28] There are about 220 instances (out of 980) where patients file a complaint.

Even though we randomly extended invitations to participants, workshop participants may differ from non-participants. In Table 3 we compare the demographic characteristics of our full sample of 1,784 surveyed patients, and the subsample of 216 patients who participated in the community workshops.[29] Our experimental participants do not seem to significantly differ from our full sample of patients in terms of age, education, wealth, marital status and number of children. The only significant difference between the two samples lies in their gender compositions, with a larger percentage of males participating in the experiment than the percentage of males surveyed while exiting the health facility.

Table 4 examines the balance in patient characteristics across the different treatments. The results show that the treatments are balanced in terms of age, marital status, wealth, and employment status. However, compared to the Complaint Box (CB) we find some statistically significant differences (p-value < 0.1) in gender and education among participants in the peer disclosure treatments (PD and PD-R), the Monetary Penalty (MP) treatment, and the Pooled Accountability (Pool) treatment. Because (adult) females have less education than (adult) males in Kenya (World Bank EdStats Data, 2020), the imbalances in schooling are likely a reflection of these gender differences.[30] Given these imbalances, we include covariates as controls in our regressions to account for observable differences across treatments.

3.2 Providers

In each eligible surveyed facility, we randomly selected a subsample of health providers to invite to the community workshop. A total of 103 health providers participated in the experiments. Table 5 provides an overview of their characteristics, and a comparison across their sector of employment.[31]As part of the survey, we registered the health providers' work schedules for the following week, and we used these schedules to conduct unannounced visits to their facilities in order to record their presence at work. This information is reported in the last row of Table 5.

About three-quarters of the sampled health professionals work in the public sector. However, with the exception of age and gender, we do not see significant demographic differences between public and private providers. Public health providers are about 10 years older (p-value <0.01) and 20 percentage points more likely to be women (p-value <0.1). Moreover, they report a higher number of days absent from work in the previous month, although the difference is not statistically significant (p=0.210). Our unannounced visits during the week following the community workshop found more than 40% of sampled health professionals

[29] Since we were unable to survey health professionals at their place of work, we could only survey the health professionals that participated in the experiment. Therefore, for health professionals we cannot formally assess the representativeness of our experimental sample.
[30] There are meaningful differences in the marital status of participants across our treatments. Focusing on the differences between the Complaint Box (CB) and the Pooled Accountability treatment we cannot reject the hypothesis that the composition of these samples is the same using conventional thresholds (P-value < 0.1)
[31] We pool together the public sector and the non-profit sector, since only 7% of our experimental participants worked in the non-profit sector.

absent from work. We do not find any statistically significant differences in absence rates between public and private sector employees.

Table 6 shows the balance in provider characteristics across treatments and the p-values of the formal hypothesis tests. The provider sector of employment, absence, and proportion of clinical officers are all balanced across treatments (p-values all <0.1). There are some imbalances in provider gender and age. The providers in the Peer Disclosure treatment were about 5 years younger than their counterparts in the Complaint Box treatment (p-value =0.05) and the Peer Disclosure with Retaliation treatment (p-value <0.1). The proportion of females in the Peer Disclosure with Retaliation treatment was 20 percentage points higher than the Complaint Box; however, this difference was slightly above the standard thresholds for statistical significance (p-value =0.14). The difference in the gender composition between the Peer Disclosure treatment and the Peer Disclosure with Retaliation treatment was even larger at almost 30 percentage points (p-value <0.05). We test for the overall balance of our sample by comparing the provider characteristics between the Pooled Accountability treatments and the Complaint Box treatment as a summary measure of our overall balance, and we do not find statistically significant differences in the provider characteristics (p-values >0.1). This summary balance tests suggests that the randomization was successful.

3.3 External Validity: Game Behavior, Choice of Sector and Absence from Work

We designed our game to focus on the trust and reciprocity relationship between patients and health providers in a controlled setting. However, the well-known downside of using a lab-type experiment is that the findings may not generalize to the context of specific interest. To ameliorate this concern, we examine how provider behavior in the lab is correlated with their actual real-life behavior at work. Specifically, we examine i) the relationship between provider behavior in the Baseline Trust Game that preceded the reporting game, ii) the sector of employment (private or public),[32] and iii) absence from work during unannounced visits to the health facility of primary employment. Specifically, during the workshop, we recorded each provider's work schedule for the following week and visited the facility at a time where the provider had stated he or she would be at work.

Our data show that public sector providers return more money to patients as compared to private sector providers. On average, the return-received ratio (R/R) of public sector providers in the baseline trust game is equal to 1.4. This means that public providers return about 144% of the tripled amount received from patients. This is significantly different from the R/R of 1 which we observed for private sector providers (Wilcoxon rank-sum non-parametric test p-value equal to 0.013). The distributions of public and private sector providers' behavior can be seen in in Figure 1, which shows that the majority of public sector

[32] Serra et al., (2012) find that the behavior of medical and nursing students in a modified Trust Game, similar to the one employed here, correlates with the same subjects' actual choice of sector and job location three years after graduation.

providers return at least as much as patients had sent them, i.e., the average returned-received ratio (R/R) is greater than or equal to 1 for 78% of public sector employees. On the other hand, almost half of the private sector providers return less than what was sent to them, i.e., the returned-received ratio (R/R) is less than 1 for nearly 50% of private providers. It is also noticeable that most public sector providers return on average more than half of the tripled amount sent by matched patients. Estimates from regression analysis[33] confirm that public sector providers return a higher percentage of the amount they receive from patients. These findings are in line with a growing theoretical (Prendergast, 2007; Besley and Ghatak, 2005; Francois, 2000) and empirical literature (Banuri and Keefer, 2016; Delfgauuw, J., & Dur, R., 2008; Gregg et al., 2011; Kolstad and Lindkvist, 2012) suggesting that private and public sector employees have different motivations and objective functions, which may lead them to respond differently to the same set of incentives.

Table 7 reports the correlation between provider behavior in the baseline trust game and absence from work. This was recorded during an unannounced visit to the provider's facility the week following his or her participation in the community workshop at a day and time when the provider had stated he/she would be at work. We report the absence rates of: 1) providers who returned less than what they received from patients (R/R<1 on average); 2) providers who returned at least as much as what was sent to them by patients (R/R>=1); 3) providers who returned at least half of the tripled amount (R/R>=1.5) on average; 4) providers who, for each possible amount sent by patients, stated that they would return what was originally sent or more (R/R>=1 for each amount that could be sent by a patient).[34] Table 7 shows that providers who return more money to patients tend to be less absent during our unannounced visits, especially if working in the public sector. In fact, among public sector providers, 57% of those who return less than what was sent by patients on average are absent from work, versus 38% of those who always return at least as much as what was sent by patients. Although the correlations tend to be not statistically significant, possibly due to the small sample size,[35] we interpret the decline in absenteeism as we move to more and more demanding measures of provider "good performance" in the game as an indication that the lab-in-the-field experiment is capturing some salient aspects of providers' motivations to serve their patients. Overall, the findings that providers who return more money to patients in the game are more likely to be working in the public sector and tend to be less absent from work provide some validation that our game can mimic the trust and reciprocity relationship between patients and providers.

[34] Recall that providers played the game using the strategy method. Therefore, they had to state how much they would return if the patient sent 200, 400, 600, or 800, after the money had been tripled. This means that each provider has 4 R/R data points, each corresponding to the possible amount sent by a patient. In columns 1 to 3 of Table 7, we categorize providers based on their average R/R, where the average is computed across the 4 potential amounts sent by patients. In column 4 of Table 7, we focus on providers who, for each amount, stated that they would return at least that amount or more.

[35] For logistical reasons we were unable to revisit all facilities. We have post-experiment absence data for 94 providers.

4. Treatment Effects

4.1 Responsiveness of Providers to Different Reporting Systems

Figure 2 reports providers' return-received ratios (R/R) in the Reporting Game across treatments. As in Figure 1, we look at three broad categories of reciprocity: 1) providers that on average give back less than what was sent to them; 2) providers that on average give back at least as much as what was sent by the patient but less than half of the total (tripled) pie; 3) providers that on average give back more than half of the total pie. Overall, the figure shows that accountability systems, and in particular the PD and PD-R systems, shifted the distribution of R/R to the right, i.e., they led to higher provider reciprocity, as compared to the Complaint Box treatment.

In Table 8 we report the OLS estimates from equation (1) where we exploit the richness of the data generated by the strategy elicitation method. As discussed previously, the strategy method elicits four observations from each provider, one for each corresponding amount (200, 400, 600, 800) possibly received from the matched patient. We report two specifications in Table 8: a specification using the Pooled Accountability treatment variable, and a specification reporting each individual treatment effect. In both specifications the CB treatment serves as the control (or omitted category). In columns 1 and 2, we use the Returned-Received (R/R) ratio corresponding to each amount sent by patients as our dependent variable. In columns 3 and 4 we estimate linear probability models where the dependent variable is a dummy equal to 1 if the average R/R is greater than or equal to 1 – meaning that providers returned at least as much as what the patients sent them – and 0 otherwise. In columns 5 and 6 the dependent variable is a dummy equal to 1 if the provider's average R/R is greater than 1, meaning that the provider returned more than what was sent to him or her. In all specifications, we control for the provider's R/R in the Baseline Trust Game, which preceded the Reporting Game, as this measures a provider's baseline reciprocity toward patients in the absence of any accountability or complaining system. We also control for the amount of money sent by patients, the sector of employment (private or public), the number of patients that each provider said he or she personally knew, whether the provider knew at least one other provider in the workshop, job title (nurse, clinical officer, etc.), and demographic characteristics, such as gender. We report both clustered standard errors in parentheses and Wild bootstrap p-values in square brackets.

The specifications in Table 8 using the Pooled Accountability treatment variables in Columns 1, 3, and 5, show that, on average, the accountability treatments increased provider reciprocity. In Column 1, the results show that on average, the accountability treatments increased the Returned/Received (R/R) ratio by 33 percentage points, a 32 percent increase relative to the mean R/R of providers in the CB group. In Column 3, we find that the Pooled Accountability treatment increased the probability that providers

returned at least as much as they received by approximately 15 percentage points, a 24 percent increase relative to the mean in the CB group. We find similar but noisier results in Column 5.

The full specifications in Column 2, 4, and 6 of Table 8 allow us to test Hypotheses 1, 2, and 3, described in Section 2.5. We report the results in the same order as we listed the Hypotheses.

Result 1: *The Peer Disclosure reporting system significantly increased the amount returned by providers to patients and reduced the likelihood of the amount returned being lower than the amount originally sent.*

We test Hypothesis 1 using Column 2, 4, and 6 of Table 8. Focusing on Column 2, we find that relative to the Complaint Box treatment, the PD increased the provider return ratio by 42 percentage points, which translates to a 40 percent increase relative to the Complaint Box mean (p-value <0.05). In Column 4, we find that the PD treatment increased the probability that providers returned at least as much as they received by 14 percentage points (p-value <0.1), a 22 percent increase relative to the CB mean. The results in Column 6 are similar in magnitude but noisier. Overall, these results show that the PD system improved provider reciprocity relative to the CB system.

Result 2: a) *The Peer Disclosure reporting system increases the amount returned by providers to patients even under the possibility of provider retaliation.*

b) The effects of the PD and PD-R systems on provider behavior are not significantly different from each other.

We test Hypothesis 2 using Columns 2, 4, and 6 of Table 8. In Column 2, we find that the PD-R system led to a 26 percentage point increase (or roughly a 25 percent increase relative to the CB mean) in the provider return ratio relative to the Complaint Box (p-value <0.05). In Column 4, we find that the PD-R treatment increased the probability that providers returned at least as much as they received by 13 percentage points (p-value <0.05), a 21 percent increase relative to the CB mean. The results in Column 6 are similar in magnitude but noisier. Overall, these results show that the PD-R improved provider reciprocity relative to the CB system.

The potential for provider retaliation could dampen the effectiveness of the PD-R system relative to the PD system. We test for this possibility by conducting formal statistical tests of equality between the PD and PD-R coefficients. In Columns 2, 4, and 6, we find that the coefficients for the PD-R treatment are systematically smaller than the PD treatment coefficients. However, we cannot reject the hypothesis that

the coefficients are equal (see the bottom of Table 8), even though the coefficient on PD is almost 16 percentage points larger that PD-R in Column 2.[36]

Result 3: *The Monetary Penalty reporting system significantly increased the amount returned by providers to patients and reduced the likelihood of the amount returned being lower than the amount originally sent.*

We also test Hypothesis 3 using Table 8. In Column 2, we find that the MP system increased the provider return ratio by 33 percentage points relative to the Complaint Box system (p-value<0.05), which translates to a 32 percent increase. In Column 4, we find that the MP treatment increased the probability that providers returned at least as much as they received by 18 percentage points (p-value <0.05), almost a 30 percent increase relative to the CB mean. The results in Column 6 are similar in magnitude but noisier. Overall, these results show that the PD-R improved provider reciprocity relative to the CB system. Formal tests of equality of all three treatment coefficients (PD, MP, and PD-R) show that the estimated effect of the Monetary Penalty system is not statistically different from the effects of the Peer Disclosure treatments.

The regression results in Table 8 show that private providers returned less money to patients than their public sector counterparts. This is consistent with first prediction outlined in Hypothesis 4. Given this pattern, and the additional public-private differences discussed in Section 3.3, in Table 9 we test for differential responses to our treatments by public or private sector of employment. This serves as a test of the second prediction outlined in Hypothesis 4 in Section 2.3. To increase statistical power, we focus on specifications using the Pooled Accountability treatment variable. This (parsimonious) approach allows us to examine the overall responsiveness of private versus public providers to our accountability treatments.[37]

Result 4: *Public sector providers return more money to patients and are more responsive to the accountability systems compared to private sector providers.*

We test Hypothesis 4 by examining the heterogeneity in treatments effects by sector of employment in Table 9. The key variables of interest are the Pooled Accountability treatment and the interaction term between the Pooled Accountability variable and the private binary variable. Column 1 shows that private providers respond less to the accountability treatments compared to their public sector counterparts. On

[36] Even though the magnitude of the coefficient between PD and PD-R in Column 2 of Table 8 is 16 percentage points, we argue that our results are more likely to indicate that there is actually no statistically significant difference between PD and PD-R because the difference between the PD and PD-R coefficients in Columns 4 and 6 are much smaller.

[37] This approach is further justified by the formal statistical tests in Table 8 where we fail to reject the hypothesis that the coefficients on PD, PD-R, and MP are all equal.

average the accountability treatments increase the R/R ratio for public providers by 42 percentage points (p-value <0.01) relative to the CB group, whereas the accountability treatments increased the R/R ratio for private providers by approximately 10 percentage points (p-value>0.1), a 33 percentage point difference that is statistically significant (p-value <0.1).[38] We find similar qualitative patterns in Columns 2 and 3, where the response among private providers is lower that the public providers, although the coefficients on the interaction are noisy due to our limited sample of private providers. Across all three columns, we find that the total effect of the accountability treatments on private sector providers (treatment + treatment x private) are qualitatively close to zero, and formal statistical tests show they are indistinguishable from zero (see the p-values reported in last row of Table 9). Overall, the results suggest that public providers are much more responsive to the accountability interventions than their private sector counterparts. These results are consistent with the growing literature showing differential motivations of public and private providers.

In Table 10, we move from testing our formal hypotheses to conducting exploratory analysis. We explore the treatment effect heterogeneity by providers' behavior (R/R) in the Baseline Trust Game, which was played before the reporting game. The providers' R/R in the baseline trust game is a continuous measure that ranges between 0 and 3. We label this variable "baseline reciprocity." In Figure 1, we found that the average baseline reciprocity among private providers was significantly lower compared to public providers (see Section 3.3 for the discussion).[39] This suggests that the patterns in Table 9 are at least partly driven by differences in the baseline reciprocity of public and private providers.

In Table 10, we present specifications using the Pooled Accountability treatment variable, as in Table 9. The results in Columns 1 through 3 in Table 10 show that the interactions between the treatments and providers' baseline reciprocity are consistently negative and statistically significant. This shows that providers who returned more (less) to patients in the baseline trust game responded less (more) to the treatments. In Column 1, we find that the accountability treatments increased the R/R ratio by 56 percentage points among providers who did not return anything at baseline (relative to the CB group). In contrast, those who returned what they received (R/R =1) at baseline were less responsive (by 23 percentage points) to the accountability treatments. Both effects were statistically significant. Moreover, the sum of the coefficients on the Pooled Accountability variable and the interaction between Pooled Accountability and baseline reciprocity were statistically significant. We find similar qualitative patterns in Columns 2 and 3.

Our results in Table 10, combined with the differences in baseline reciprocity of public and private providers, suggests that baseline reciprocity (a measure of prosocial behavior) could partly explain the private-public differential responses that we document in Table 9. Further, since public providers who

[38] The treatment effects for the private providers are found by summing the coefficients on Pooled accountability and Pooled Accountability X Private.

[39] On average private providers returned the same amount that was sent, while public providers returned 1.4 times the amount sent (results not shown in a table). This difference was statistically significant (p<0.01), as discussed in Section 3.3.

returned more to patients in the trust game tend to have lower rates of absenteeism (as shown in Section 2.3), this suggests that the treatments may promote more prosocial behavior among low-performing (or unmotivated) workers who tend to be absent from work.[40] We also explored heterogeneity by the number of other health workers known, as this would vary the cost of the PD and PD-R treatments. We did not find any significant interactions of our treatments with knowledge of other workers.

4.2 Patients' Willingness to Complain

Recall that in the Reporting Game, patients had two decisions to make. First, they had to decide how much money they wanted to send to a randomly matched provider in the other room. Then, using the strategy method, they had to decide whether or not to complain for every possible amount they could receive back. If they complained, they could send the provider up to five cards displaying a frowning face, incurring a fixed cost of 50 KES. Given the ambiguous theoretical predictions of the treatments on complaining behavior outlined in Section 2.3, we examine patient complaining behavior using equations (2) and (3) in section 2.5. The results are shown in Tables 11 and 12, where our specifications use both the Pooled Accountability treatment and the individual treatments. Our specifications include controls for the patient's age, gender, wealth, education, recognition of a health worker, satisfaction with care they received during exit survey, and the prestige or social status ranking they attach to the medical profession. The results using a parsimonious set of controls are shown in Tables A.3 and A.4 in the appendix.

On average, patients sent about 254 KES to their matched providers.[41] In Column 1 of Table 11, we find that our accountability treatments increase the probability of sending a positive amount to a provider by 13 percentage points (p-value <0.1) relative to the CB group. This is mainly driven by the PD-R and the MP groups (see Column 2). The results also show that the only individual characteristics that significantly predict patients' decisions to send a positive amount to a provider are individual wealth and patients' perceptions of the social status of the healthcare profession (coefficients not shown in table).[42] Wealthier individuals and individuals who hold the health profession in high regard are more likely to send a positive, and a larger amount to the matched provider in the game.

Since the decision to file a complaint can only be made if the patient sends a positive amount to the matched provider, our analysis of patients' propensities to complain is only relevant for the 164 "trusting"

[40] The coefficients on the interactions between our treatments and actual absenteeism were negative but imprecisely estimated, as the absenteeism data was collected on a smaller set of providers due to logistical constraints.
[41] In the Trust Game that preceded the Reporting Game, the average amount sent to providers is 279 KES with no statistically significant difference across treatments.
[42] The perceived social status of health workers was measured during exit survey through a set of vignettes displaying 11 different professions and eliciting patients' opinions about how respected each profession was. We categorize a patient as having high regards for the health profession if he or she ranked the profession of doctor as first or second.

patients who decided to send a positive amount to the matched provider. Figure 3 shows the percentages of patients sending each possible positive amount (200, 400, 600 or 800) to their matched providers under each reporting system.[43] In columns 3 and 4 of Table 11, we examine the amounts sent to providers by the patients who sent a positive amount. Using the Pooled Accountability variable in Column 3, we find that, on average, the accountability treatments had a limited impact on the amount send to the provider (among this sample of patients). The coefficient is small (30 KES) and not statistically significant. In Column 4, we find qualitatively similar results, with small and statistically insignificant coefficients across all three accountability treatments. This implies that that the amounts sent by the "trusting" patients are balanced across the accountability treatments. Recall that the amounts sent by patients determine the amounts that providers could send back, which in turn determine patients' willingness to complain. Thus, the balanced sending behavior among "trusting" patients suggests that we can more confidently examine the effects of the treatments on patient complaining behaviors.[44]

In Table 12, we assess patients' willingness to file complaints under the different treatment conditions using the specification outlined in equation (3). We exploit our use of the strategy elicitation method to register patients' binary complaining decisions for each amount of money they could receive back from the matched provider, out of what they had sent. In Columns 1 and 2 we report estimates from linear probability regressions where the dependent variable is a dummy equal to 1 if the patient complained. In Column 1, we find that the having any consequences for providers attached to the complaints, as indicated by the Pooled Accountability treatment variable, reduced the probability of complaining by nine percentage points relative to the CB condition (p-value< 0.05). This represents a reduction of just over 30 percent relative to the CB mean. The results in Column 2 are quantitatively similar and show there is little difference in the probability of complaining across the three individual accountability treatments.

In columns 3 and 4 of Table 12 we examine the intensity of the complaining decision. The dependent variable in these OLS regressions is the number of frowning face cards that a patient was willing to send to a provider for any possible amount returned by the provider. Recall that a patient could send up to five cards displaying a frowning face after paying a lump sum cost of 50 KES. The results in Column 3 show that conditional on complaining, patients across all treatments (i.e., the Pooled Accountability variable) sent almost 0.8 more frowning face cards compared to the CB group. This represents close to a 33 percent increase in cards sent relative to the CB mean. The detailed specification in Column 4 allows us to test whether the intensity of complaints differed by our individual treatments. Patients in the PD treatment sent almost one more frowning face card, and patients in the MP treatments sent almost four-fifths more cards

[43] We report summary statistics of patient sending behavior in the Table A.2 of the appendix.
[44] A caveat is the largest effect we find is among patients in the MP treatment in Column 4. With a larger sample, we would expect this coefficient to be statistically significant.

compared to CB. Relative to the Complaint Box mean, this represents a 33 to 44 percent increase in cards sent, respectively. In contrast, the point estimates of the PD-R treatment are smaller and almost half that of the PD treatment. Formal statistical tests show that we cannot reject the equality of all treatments; however, we would likely find a statistically significant difference between PD and PD-R with additional data.

Result 5: *a) Patients are less likely to complain against providers in the Peer Disclosure (PD) and Monetary Penalty (MP) reporting treatments than in the Complaint Box (CB) treatment.*

b) Complaining patients file more complaints against providers in the PD, PD-R, and MP treatments than in the CB treatment.

Overall, the results show that more patients are reticent to complain when their reports lead to actual consequences for providers. However, conditional on filing reports, the interventions lead to greater intensity of complaints (measured by the number of cards sent). The potential for provider retribution seems to lower the intensity of complaints, but as discussed above, not significantly so due to our limited sample size. In exploratory analysis reported in Table A.5 of the appendix, we find some evidence that pre-existing social relationships affect decision making in the reporting games. Specifically, we find that patients are more reluctant to file complaints when they recognized a provider. This pattern was clearest in the MP treatment. This may be due to a fear of retribution outside the lab or unwillingness to take action against an acquaintance, both of which seem to be particularly salient under the MP treatment.

5. Conclusions

Patient trust in health providers and provider reciprocity are key determinants of health-seeking behavior and utilization of services among individuals. Using a modified Trust Game conducted on a sample of actual health providers and patients randomly sampled from public and private health facilities in Nairobi, Kenya, we examined the effectiveness of reporting schemes coupled with either monetary or non-monetary penalties for health-workers' behavior.

Overall, we found that systems that attach (any) tangible consequences to citizen reports improve providers' reciprocity, as measured by the amount of money sent back to patients in a modified Trust Game. In particular, our data show that the threat of social sanctions (in the form of peer shaming) is highly effective in increasing provider reciprocity in the game. In addition, we find that although patients are less likely to complain when reports lead to tangible negative consequences for providers, the decrease in reporting is modest, and the intensity of complaints is higher.

Given the low levels of accountability in public health care systems around the developing world, and the high costs and inefficiencies that come with top-down monitoring and enforcement, our findings suggest that citizen monitoring systems that leverage peer pressure and reputational concerns may be a cost-effective approach to trust-augmenting behavior among providers. One possible caveat of our study is that in the experiment, patients are able to clearly evaluate provider behavior, whereas this may be challenging in the field, especially in contexts of low education. Thus, our results are more relevant for easily observable violations of trust, such as absence from work and bribe requests, which are prevalent in developing settings. Further, trust in service providers is important in the education sector, especially when providers are working with younger children. Thus, our findings could also be relevant in these sectors.

Additional exploratory findings (reported in the appendix) also suggest that improving the anonymity of the reporting system, as suggested by Chassang and Padro i Miguel (2012), is important, as it could significantly increase citizen participation rates. Such a reporting system may require the reporting to be outsourced and managed by a trustworthy third party, potentially in partnership with the government. Further research would be needed to better examine the importance of anonymity in bottom up accountability systems.

REFERENCES

Abbink, K. and Sadrieh, A., 2009. The pleasure of being nasty. Economics letters, 105(3), pp.306-308.

Abbink, K., Dasgupta, U., Gangadharan, L. and Jain, T., 2014. Letting the briber go free: An experiment on mitigating harassment bribes. Journal of Public Economics, 111, pp.17-28.

Aker, Jenny C., Paul Collier, and Pedro C. Vicente. Is information power? Using mobile phones and free newspapers during an election in Mozambique. Review of Economics and Statistics, forthcoming.

Alsan, M., Garrick, O., & Graziani, G. (2019). Does diversity matter for health? Experimental evidence from Oakland. *American Economic Review*, *109*(12), 4071-4111.

Alsan, M., & Wanamaker, M. (2018). Tuskegee and the health of black men. *The Quarterly Journal of Economics*, *133*(1), 407-455.

Anderson, M., and J. Magruder. 2012. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. The Economic Journal, 122(563): 957-989.

Andreoni, J., W. Harbaugh and L. Vesterlund. 2003. The Carrot or the Stick: Rewards, Punishments, and Cooperation. The American Economic Review*,* 93(3): 893-902.

Andreoni, J., & Bernheim, B. D., 2009. Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. Econometrica, 77(5), 1607-1636.

Andreoni, J., & Petrie, R., 2004. Public goods experiments without confidentiality: a glimpse into fund-raising. Journal of public Economics, 88(7), 1605-1623.

Archibong, B. & Annan, F. (2021) "'We Are Not Guinea Pigs': The Effects of Negative News on Vaccine Compliance" Unpublished Working Paper.

Ariely, D., Bracha, A., & Meier, S., 2009. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. The American Economic Review, 99(1), 544-555.

Ashraf, N., Bandiera, O., and Lee, S.S., 2014. Awards unbundled: Evidence from a natural field experiment. Journal of Economic Behavior & Organization, 100: 44_63

Azmat, G. and Iriberri, N., 2010. The importance of relative performance feedback information: Evidence from a natural experiment using high school students. Journal of Public Economics, 94(7): 435-452.

Balafoutas, L., & Nikiforakis, N., 2012. Norm enforcement in the city: a natural field experiment. European Economic Review, 56(8), 1773-1785.

Balafoutas, L., Grechenig, K. and Nikiforakis, N., 2014. Third-party punishment and counter-punishment in one-shot interactions. Economics letters, 122(2), pp.308-310.

Banerjee A, Banerji Rukmini, Duflo E, Glennerster R and Khemani S., 2010. Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. American Economic Journal: Economic Policy 2015:2:1, 1-30.

Banuri, S., & P. Keefer, 2016. Pro-social motivation, effort and the call to public service. European Economic Review, 83, 139-164.

Barr, A., Lindelow, M., & Serneels, 2009. Corruption in public service delivery: An experimental analysis. Journal of Economic Behavior & Organization, 72(1), 225-239.

Barr, A., Packard, T., & D. Serra, 2014. Participatory accountability and collective action: Experimental evidence from Albania. European Economic Review, *68*, 250-269.

Barr, A., Frederick Mugisha, and Pieter Serneels and A. Zeitlin (2012). Information and collective action in community monitoring of schools: Field and lab experimental evidence from Uganda. Working paper.

Bénabou, R., & J. Tirole, 2006. Incentives and prosocial behavior. The American Economic Review, 96(5), 1652-1678.

Berg, J., Dickaut J. and K. McKabe (1995). Trust, reciprocity and social history. Games and Economic Behavior, 10, 122-145.

Besley, T., & M. Ghatak, 2005. Competition and incentives with motivated agents. American Economic Review, 95(3), 616–636.

Blanes i Vidal, J., & M. Nossol, 2011. Tournaments without prizes: Evidence from personnel records. *Management science*, *57*(10), 1721-1736.

Björkman Nyqvist, M., D. de Walque, and J. Svensson. 2017. "Experimental Evidence on the Long-Run Impact of Community-Based Monitoring." American Economic Journal: Applied Economics, 9(1): 33-69.

Bjorkman A and J. Svensson, 2009. Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda. Quarterly Journal of Economics 2009:124 (2): 735-769.

Björkman, M., & J. Svensson, 2010. When is community-based monitoring effective? Evidence from a randomized experiment in primary health in Uganda. Journal of the European Economic Association, 8(2-3), 571-581.

Bold T, Gauthier B., Maestad, Svensson J. and W. Waly, 2011. *Service Delivery Indicators: Pilot in Education and Health Care in Africa.* World Bank.

Brandts, J. and G. Charness, 2000. Hot vs. cold: Sequential experimental games. Experimental Economics 227–38.

Brandts, J. and G. Charness, 2011. The strategy versus the direct-response method: A first survey of experimental comparisons. Experimental Economics 2011, 14(3): 375-398.

Brent, D.A., Gangadharan, L., Mihut, A. and Villeval, M.C., 2019. Taxation, redistribution, and observability in social dilemmas. Journal of Public Economic Theory, 21(5), pp.826-846.

Brock J. M., A. Lang and K. Leonard, 2015. Esteem and social information: On determinants of prosocial behavior of clinicians in Tanzania. Journal of Economic Behavior and Organization 118 (2015) 85–94.

Brock, J. M., Lange, A., & Leonard, K. L., 2016. Generosity and Prosocial Behavior in Healthcare Provision Evidence from the Laboratory and Field. Journal of Human Resources, *51*(1), 133-162.

Buchan, N. R., Croson, R. T. A., and R. M. Dawes, 2002. Swift Neighbors and persistent strangers: A cross-cultural investigation of trust and reciprocity in social exchange. American Journal of Sociology, 108(1), 168–206.

Cabral, L. and Hortacsu, A., 2010. The dynamics of seller reputation: Evidence from eBay. The Journal of Industrial Economics, 58(1), pp.54-78.

Calabuig, V., Fatas E., Olcina G. and I. Rodriguez-Lara, 2013. Carry a big stick or no stick at all. Working Paper ERI-CES, Universitat de València and ERI-CES.

Carpenter, J. and E. Seki, 2011. Do social preferences increase productivity? Field experimental evidence from fishermen in toyama bay. Economic Inquiry 49 (2), 612-630.

Cason, T.N., Friesen, L. and Gangadharan, L., 2016. Regulatory performance of audit tournaments and compliance observability. European Economic Review, 85, pp.288-306.

Chaudhury N, Hammer J, Kremer M., Muralidharan K. and F. Halsey Rogers, 2004. Provider Absence in Schools and Health Clinics. Northeast Universities Development Consortium Conference: October 2004.

Charness, G., Masclet, D. and Villeval, M.C., 2014. The dark side of competition for status. Management Science, 60(1), pp.38-55.

Chassang, Sylvain and Padro-i-Miguel, 2014. Corruption, Intimidation, and Whistle-blowing: A Theory of Inference from Unverifiable Reports. NBER Working paper 20315.

Chevalier, J. A., and D. Mayzlin, 2006. The effect of word of mouth on sales: Online book reviews. Journal of marketing research, 43(3): 345-354.

Cilliers, J., Mbiti, I.M. and Zeitlin, A., Forthcoming. Can public rankings improve school performance? Evidence from a nationwide reform in Tanzania. Journal of Human Resources

Das, J., A. Holl, A. Mohpal and K. Muralidharan, 2016. Quality and Accountability in Healthcare Delivery: Audit-Study Evidence from Primary Care in India. UCSD Working Paper.

Das, J. and J. Hammer, 2014. Quality of primary care in low-income countries: Facts and economics. *Annual. Rev. Econ.* 6.1 (2014): 525-553.

Das, J., A. Holl, A. Mohpal and K. Muralidharan, 2015. The Fiscal Cost of Weak Governance: Evidence from Teacher Absence in India. UCSD Working Paper.

Delfgaauw, J., R. Dur, J. Sol, and W. Verbeke. 2013. Tournament Incentives in the Field: Gender Differences in the Workplace. Forthcoming, Journal of Labor Economics*, 31*(2), 305-326.

Denant-Boemont, L., Masclet, D. and Noussair, C.N., 2007. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. Economic theory, 33(1), pp.145-167.

Duflo E, Dupas P and M. Kremer, 2015. School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. Journal of Public Economics, 123: 92-110.

Dufwenberg, M. and Muren, A., 2006. Generosity, anonymity, gender. Journal of Economic Behavior & Organization, 61(1), pp.42-49.

Dugar, S., 2010. Nonmonetary sanctions and rewards in an experimental coordination game. Journal of Economic Behavior & Organization 73 (3), 377-386.

Eriksson, T., A. Poulsen, M. C. Villeval. 2009. Feedback and incentives: Experimental evidence. Labor Economics. 16(6) 679_688.

Fehr, E., & U. Fischbacher, 2004. Third-party punishment and social norms. Evolution and human behavior, 25(2), 63-87.

Fehr, E., & S. Gächter, 2002. Altruistic punishment in humans. Nature, 415(6868), 137-140.

Fehr, E. and B. Rockenbach, 2003. Detrimental effects of sanctions on human altruism." Nature 422, 137–140.

Fiszbein A, Ringold D and Halsey Rogers F. *Making Services Work: indicators, Assessments, and Benchmarking of the Quality and Governance of Public Service Delivery in the Human Development Sectors.* World Bank 2011.

Francois, P., 2000. Public service motivation as an argument for government provision. Journal of Public Economics, 78(3), 275–299.

Gerber, A. S., Green, D. P., & Larimer, C. W., 2008. Social pressure and voter turnout: Evidence from a large-scale field experiment. American Political Science Review, 102(01), 33-48.

Gill, D., Kissová, Z., Lee, J., & Prowse, V. L., 2015. First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. *Available at SSRN 2641875*.

Goldstein, M., Graff Zivin, J., Habyarimana, J., Pop-Eleches, C., & H. Thirumurthy, H., 2013. The effect of absenteeism and clinic protocol on health outcomes: The case of mother-to-child transmission of HIV in Kenya. American Economic Journal: Applied Economics, 5(2), 58-85.

Gregg, P., Grout, P.A., Ratcliffe, A., Smith, S. and Windmeijer, F., 2011. How important is pro-social behaviour in the delivery of public services? Journal of Public Economics, 95(7), pp.758-766.

Grossman, G, Michelitch K, Santamaria MM. 2016. Texting Complaints to Politicians: Name Personalization and Politicians' Encouragement in Citizen Mobilization. Working paper.

Houser, D. and J. Wooders, 2006. Reputation in auctions: Theory, and evidence from eBay. Journal of Economics & Management Strategy, 15(2): 353-369.

Houser, D., Xiao, E., McCabe, K. and V. Smith, 2008. When punishment fails: Research on sanctions, intentions and non-cooperation. Games and Economic Behavior 62, 509–532.

Kenya National Bureau of Statistics (KNBS), 2010. Kenya Demographic and Health Survey 2008-09. Calverton, Maryland: KNBS.

Kessler, J. and Vesterlund, L. (2015). "The external validity of laboratory experiments: The misleading emphasis on quantitative effects." in Frechette, G.R. and Schotter, A. (Eds.), Handbook of Experimental Economic Methodology, Oxford University Press, Oxford, UK.

Kriss, Peter H., Roberto A. Weber, and Erte Xiao. "Turning a blind eye, but not the other cheek: On the robustness of costly punishment." Journal of Economic Behavior & Organization 128 (2016): 159-177.

Kolstad, J.R. and Lindkvist, I., 2012. Pro-social preferences and self-selection into the public health sector: evidence from an economic experiment. Health policy and planning.

Kovacs, R.J., Lagarde, M. and Cairns, J., 2019. Measuring patient trust: Comparing measures from a survey and an economic experiment. Health economics, *28*(5), pp.641-652.

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

Linardi, S. and J Jones, 2014. Wallflowers: Experimental Evidence of an Aversion to Standing Out. Management Science, 60 (7), 1757-1771.

Linardi, S., & McConnell, M. A., 2011. No excuses for good behavior: Volunteering and the social environment. Journal of Public Economics, 95(5), 445-454.

Lowes, S. R., & Montero, E. (2020). The Legacy of Colonial Medicine in Central Africa. UCSD Working Paper.

Mansuri G. and V. Rao, 2013. *Localizing Development: Does Participation Work?*. Policy Research Report, Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/11859

Martinez-Bravo, M., & Stegmann, A. (2020). In Vaccines We Trust? The Effects of the CIA's Vaccine Ruse on Immunization in Pakistan. CEMFI Working Paper No. 1713

Masclet, D., Noussair, C., Tucker, S., & Villeval, M. C., 2003. Monetary and non-monetary punishment in the voluntary contributions mechanism. The American Economic Review, 93(1), 366-380.

Ministry of Health (2010). Kenya Service Provision Assessment Survey. Ministry of Health, Nairobi, Kenya.

Nikiforakis, N. and Engelmann, D., 2011. Altruistic punishment and the threat of feuds. Journal of Economic Behavior & Organization, 78(3), pp.319-332.

Noussair, C. and Tucker, S., 2007. Public observability of decisions and voluntary contributions in a multiperiod context. Public Finance Review, *35*(2), pp.176-198.

Pradhan M, Suryadarma D, Beatty A, Wong M, Gaduh A and Artha, 2014. Improving Educational Quality through Enhanced Community Participation: Results from a Randomised Field Experiment in Indonesia. American Economic Journal: Applied Economics 2014, 6(2): 105–126.

Prendergast, C., 2007. The motivation and bias of bureaucrats. American Economic Review, 97(1), 180–196.

Resnick, P., R. Zeckhauser, J. Swanson, and K. Lockwood, 2006. The value of reputation on eBay: A controlled experiment. Experimental Economics, 9(2): 79-101.

Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D. (2019). Fast and wild: Bootstrap inference in Stata using boottest. The Stata Journal, 19(1), 4-60.

Rigdon, M., 2009. Trust and reciprocity in incentive contracting. Journal of Economic Behavior and Organization 70, 93–105.

Serra, D., P. Serneels and A. Barr. 2011. Intrinsic motivations and the non-profit health sector: Evidence from Ethiopia. Personality and Individual Differences 51(3):309-314.

Tran, A., Zeckhauser, R., 2012. Rank as an Inherent Incentive: Evidence from a Field Experiment. Journal of Public Economics 96 (9-10), 645-650.

Transparency International, 2011. The Kenya Health Sector Integrity Study Report. Transparency International, Nairobi.

World Bank. Creating Fiscal Space for Poverty Reduction in Ecuador; A Fiscal Management and Public Expenditure Review: Teacher Absence and Incentives in Primary Education. World Bank: 2005.

World Bank. 2013 Kenya Service Delivery Indicators, Online Database, World Bank, Washington DC.

World Bank. 2017 World Development Indicators, Online Database, World Bank, Washington DC.

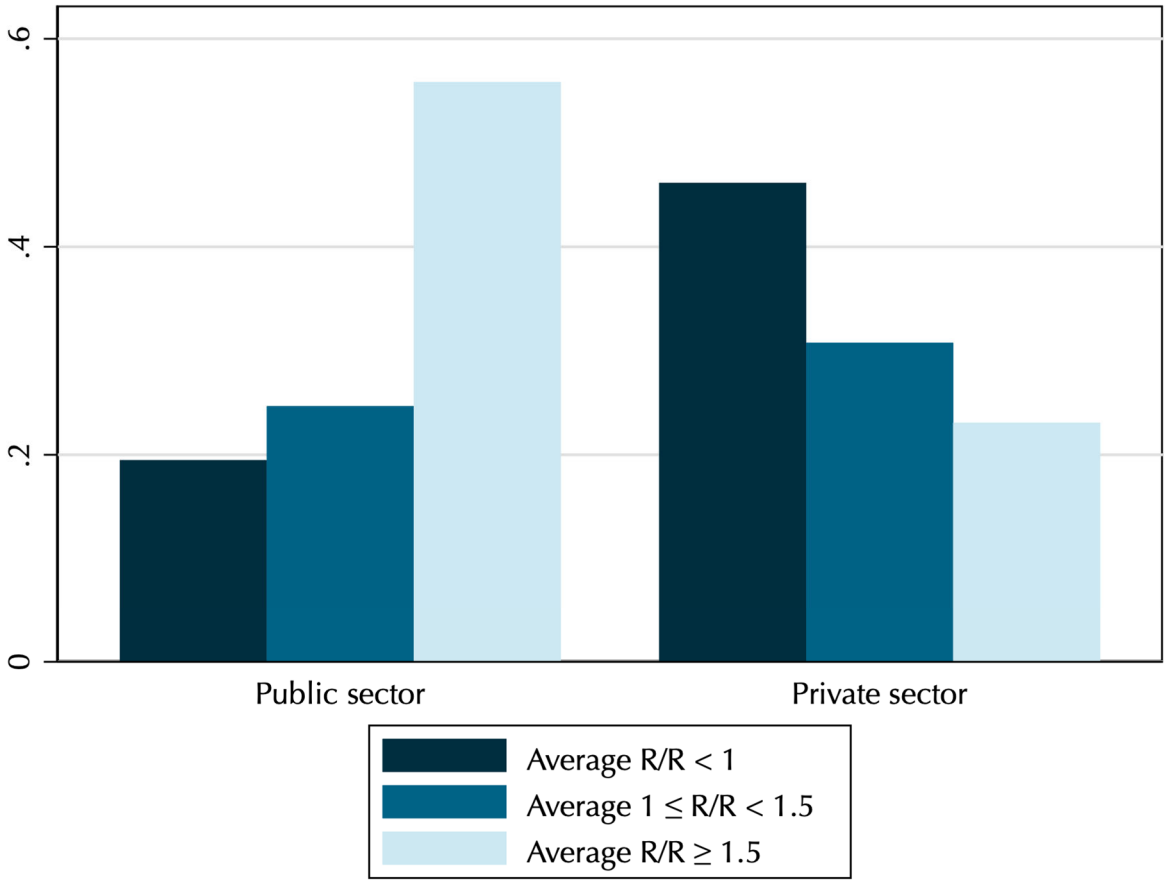
World Bank. 2004. *Making Service Work for Poor People.* World Development Report. Washington, DC: The World Bank.

Xiao, E., & Houser, D., 2005. Emotion expression in human punishment behavior. Proceedings of the National Academy of Sciences of the United States of America, 102(20), 7398-7401.

Xiao, E., & D. Houser, 2011. Punish in public. Journal of Public Economics, 95(7), 1006-1017.

Figures

Figure 1
Providers' behavior in the Baseline Trust Game by sector of employment



Note: The figure shows the percentages of public sector and private sector providers whose average returned-received ratios are lower than 1, between 1 and 1.5, and equal to or greater than 1.5.

Figure 2
Amount returned by providers across treatments

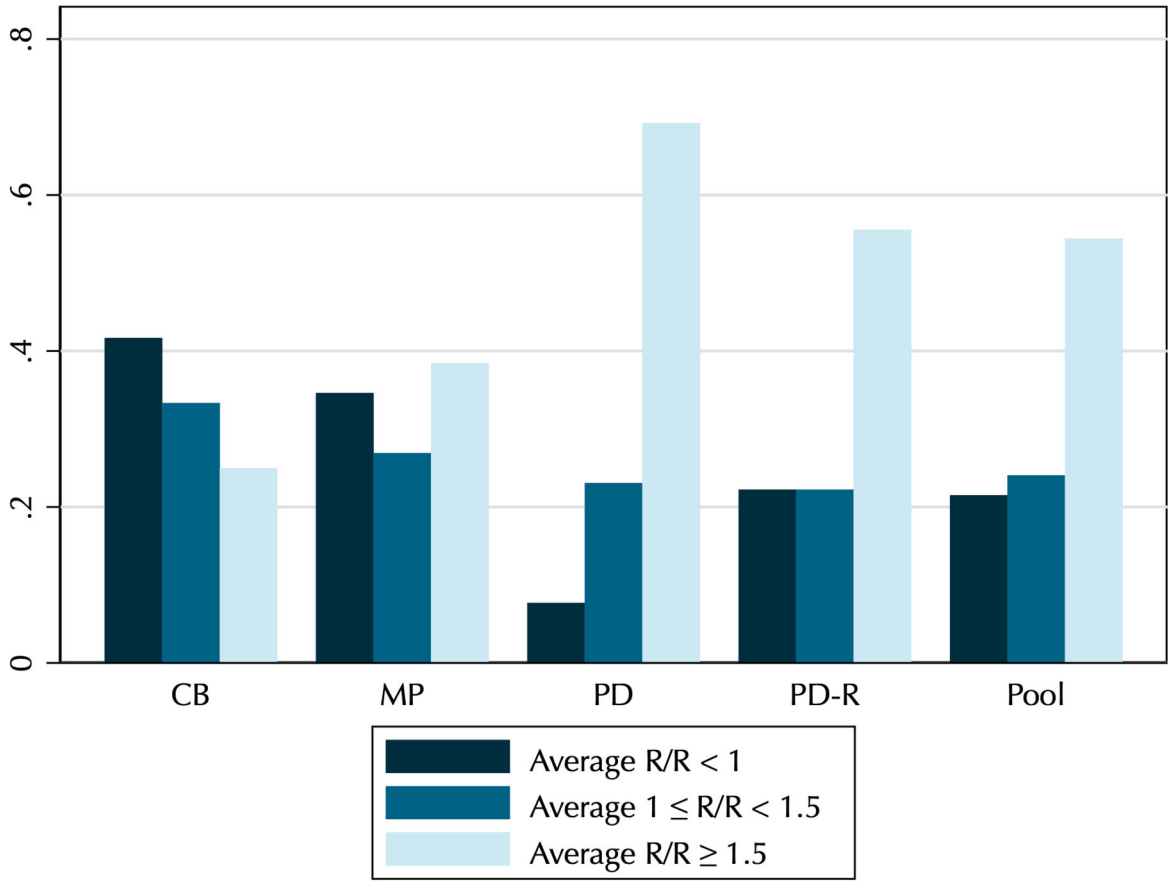
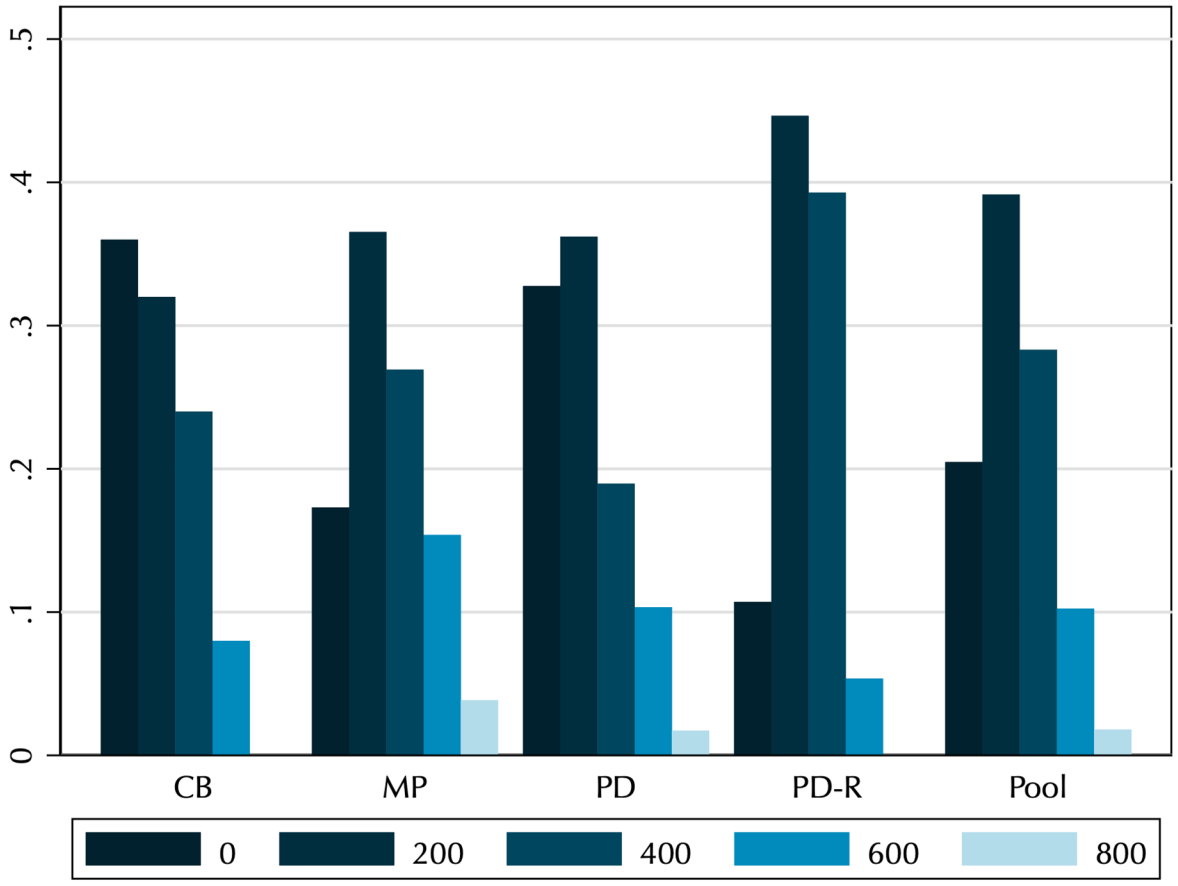


Figure 3
Amount sent to providers by "trusting" patients



Tables

Table 1
Characteristics of the Health Facilities

	Public			Private			Total		
	Mean of percentage	Min	Max	Mean of percentage	Min	Max	Mean of percentage	Min	Max
# of full-time staff	12.5	0.0	33.0	3.7	0.0	13.0	8.7	0.0	33.0
	(9.4)			(2.5)			(8.4)		
# doctors	0.2	0.0	3.0	0.2	0.0	1.0	0.2	0.0	3.0
	(0.6)			(0.4)			(0.5)		
# clinical officers	2.1	0.0	6.0	1.3	0.0	5.0	1.7	0.0	6.0
	(1.4)			(1.0)			(1.3)		
# nurses	10.3	0.0	30.0	2.0	0.0	8.0	6.7	0.0	30.0
	(8.2)			(1.4)			(7.5)		
# part-time staff	0.8	0.0	12.0	1.3	0.0	8.0	1.0	0.0	12.0
	(2.2)			(2.0)			(2.1)		
# patients per day	171.6	6.0	800.0	16.8	1.0	60.0	102.8	1.0	800.0
	(154.6)			(13.4)			(138.7)		
With a lab	0.21	0.0	1.0	0.12	0.0	1.0	0.17	0.0	1.0
With a room for surgeries	0.81	0.0	1.0	0.72	0.0	1.0	0.77	0.0	1.0
With overnight space for patients	0.77	0.0	1.0	0.42	0.0	1.0	0.62	0.0	1.0
With visible complaint box	0.42	0.0	1.0	0.69	0.0	1.0	0.54	0.0	1.0
With exposed list of prices	0.74	0.0	1.0	0.90	0.0	1.0	0.81	0.0	1.0
From Exit Surveys:									
Waiting time	86.1	6.6	183.5	19.4	0.0	105.0	57.3	0.0	330.0
	(44.8)			(23.8)			(57.1)		
Cost of visit	75.2	0.0	380.0	184.0	0.0	550.0	125.6	0.0	550.0
	(116.7)			(149.4)			(136.9)		
Average satisfaction with health worker	4.4	3.7	5.0	4.8	4.1	5.0	4.6	3.7	5.0
	(0.3)			(0.3)			(0.4)		
Average satisfaction with care	4.3	3.5	5.0	4.7	3.0	5.0	4.5	3.0	5.0
	(0.4)			(0.5)			(0.4)		
Patients' years of schooling	10.0	7.7	13.2	11.3	7.7	14.0	10.5	6.5	14.0
	(1.1)			(1.5)			(1.6)		
Patient wealth (asset index)	-0.2	-1.1	0.7	0.6	-1.0	1.7	0.1	-1.6	1.7
	(0.4)			(0.6)			(0.7)		

Note: Facility characteristics, in the upper panel of the table, were registered through facility-level survey. The variables in the bottom panel are generated by surveying patients when exiting the facilities. The asset index that we use as a proxy for patients' wealth was constructed by conducting factor analysis over six household assets: ownership of a TV, a refrigerator, a car, a bank account, good source of fuel, good materials used for house outer walls.

Table 2
Treatments and Sessions

	Sessions	Health professionals	Patients
Treatment			
Complaint Box (CB)	5	24	50
Peer Disclosure (PD)	7	26	58
Peer Disclosure with Retaliation (PD-R)	6	27	56
Monetary Penalty (MP)	6	26	52
Pooled Accountability (Pool)	19	79	166
Total	24	103	216

Note: Each participant (health professional or patient) participated only in one treatment.

Table 3
Patient Characteristics: Full Sample vs. Workshop Sample

	Full Sample	Workshop Participants	P-value
Age	30.52	31.14	0.35
	(10.12)	(14.27)	
Male	0.18	0.23	0.09*
	(0.39)	(0.42)	
Years of education	10.30	10.37	0.73
	(3.23)	(3.43)	
Wealth (asset index)	0.00	-0.07	0.30
	(1.00)	(0.98)	
Married	0.84	0.80	0.18
Works for a wage	0.24	0.23	0.61

Note: Standard deviation in parentheses. P-values are generated from two-sided tests of equality of means across samples.

Table 4
Patients' Balancing Tests

	Age	Female (%)	Years of Schooling	Asset Index	Single (%)	Works for a wage (%)
Complaint Box (CB)	30.49	0.92	9.10	-0.13	0.12	0.22
Peer Disclosure (PD)	30.24	0.70	11.68	0.07	0.25	0.28
Peer Disclosure with Retaliation (PD-R)	31.06	0.68	11.06	0.00	0.23	0.21
Monetary Penalty (MP)	30.10	0.81	9.40	-0.24	0.19	0.17
Pooled Accountability (Pool)	30.47	0.73	10.78	-0.05	0.22	0.22
p-value (MP=CB)	0.82	0.13	0.69	0.57	0.36	0.48
p-value (PD=CB)	0.86	0.01	0.00	0.30	0.11	0.51
p-value (PD-R=CB)	0.70	0.00	0.01	0.48	0.17	0.84
p-value (MP=PD)	0.94	0.19	0.00	0.13	0.51	0.17
p-value (PD-R=PD)	0.60	0.80	0.27	0.75	0.82	0.38
p-value (MP=PD-R)	0.59	0.13	0.01	0.23	0.67	0.60
p-value (Pool=CB)	0.99	0.01	0.00	0.62	0.13	0.97

Note: We report p-values are generated from two-sided tests of equality of means across samples.

Table 5
Characteristics of Health Providers

	Full sample	Public Facility	Private Facility	p-value
Mean or share	N= 103	N= 77	N= 26	
Work in Health Center (vs. Dispensary)	0.86	0.84	0.92	0.32
Clinical Officer (vs. Nurse)	0.25	0.23	0.31	0.46
Female	0.61	0.66	0.46	0.07
Age	36.36 (10.12)	39.18 (9.78)	28 (5.43)	0.00
Joined health profession to "help the poor"	0.24	0.26	0.19	0.49
% absent from work at least 1 day in the last 30 days (self-reported)	0.27	0.30	0.19	0.30
% absent during second visit to the facility	0.44	0.45	0.40	0.67

Note: We report p-values are generated from two-sided tests of equality of means across samples. Standard deviations in parentheses. Our measure of absence during unannounced visits (last row) is on a restricted sample of 94 providers, of which 25 from private facilities and 69 from public facilities.

Table 6
Health Providers' Balancing Tests

	Age	Female (%)	Clinical Officer (%)	Private Sector (%)	Absent during second visit (%)
Complaint Box (CB)	38.88	0.58	0.25	0.25	0.42
Peer Disclosure (PD)	33.27	0.46	0.31	0.31	0.46
Peer Disclosure with Retaliation (PD-R)	37.89	0.78	0.26	0.19	0.37
Monetary Penalty (MP)	35.54	0.62	0.19	0.27	0.50
Pooled Accountability (Pool)	35.59	0.62	0.25	0.25	0.44
p-value (MP=CB)	0.29	0.82	0.63	0.88	0.62
p-value (PD=CB)	0.05	0.40	0.66	0.66	0.81
p-value (PD-R=CB)	0.76	0.14	0.94	0.58	0.74
p-value (MP=PD)	0.34	0.28	0.35	0.77	0.78
p-value (PD-R=PD)	0.07	0.02	0.70	0.31	0.53
p-value (MP=PD-R)	0.40	0.21	0.57	0.47	0.36
p-value (Pool=CB)	0.17	0.75	0.98	0.98	0.88

Note: We report p-values are generated from two-sided tests of equality of means across samples

Table 7
Behavior in the Baseline Trust Game and Absence from Work

	Providers who returned less than received (avg. R/R<1)	**Providers who returned at least as much as received (avg. R/R>=1)**	**Providers who returned at least half of the pie (avg. R/R>=1.5)**	**Providers who always returned at least as much as received (R/R>=1 always)**
Full sample				
% Absent at second visit	42%	44%	38%	41%
Public sector				
% Absent at second visit	57%	42%	39%	38%*

Note: The Returned/Received ratio (R/R) measures the proportion of money providers send back to clients relative to the tripled amount received from patients. This ranges from 0, if providers kept the full tripled amount, to 3, if providers returned the full tripled amount. Since providers played the game using the strategy method, we report categories based on average provider behavior across all possible amounts sent by patients. Asterisks indicate a significance difference (based on Chi-square tests) between absenteeism of providers who belong to the category stated in the top of the column and providers who do not belong to that category. *** p<0.01, ** p<0.05, * p<0.1.

Table 8

Providers' responsiveness to different reporting systems (strategy)

	Returned/Received Ratio (R/R)		Returned/Received Ratio (R/R) ≥ 1		Returned/Received Ratio (R/R) >1	
	[1]	[2]	[3]	[4]	[5]	[6]
Pooled Accountability	0.33***		0.15**		0.13*	
	(0.09)		(0.06)		(0.06)	
	[0.01]		[0.06]		[0.11]	
Peer Disclosure (PD)		0.42**		0.14*		0.16
		(0.17)		(0.08)		(0.10)
		[0.02]		[0.12]		[0.18]
Peer Disclosure with Retaliation (PD-R)		0.26**		0.13**		0.11
		(0.11)		(0.06)		(0.07)
		[0.05]		[0.06]		[0.16]
Monetary Penalty (MP)		0.33**		0.18**		0.12
		(0.12)		(0.07)		(0.09)
		[0.05]		[0.05]		[0.29]
Private	-0.31**	-0.31**	-0.09	-0.09	-0.21**	-0.21*
	(0.15)	(0.14)	(0.09)	(0.09)	(0.10)	(0.10)
	[0.07]	[0.07]	[0.32]	[0.32]	[0.07]	[0.09]
Constant	0.45	0.42	0.36**	0.35**	0.12	0.11
	(0.31)	(0.32)	(0.14)	(0.14)	(0.21)	(0.23)
	[0.18]	[0.23]	[0.02]	[0.03]	[0.59]	[0.64]
Observations	412	412	412	412	412	412
R-squared	0.51	0.51	0.32	0.32	0.33	0.33
Mean of control group (CB)	1.04	1.04	0.62	0.62	0.45	0.45
P-Values for Test of Equality of Coefficients						
MP=PD		0.63		0.68		0.75
MP=PD-R		0.63		0.43		0.91
PD=PD-R		0.30		0.79		0.61

Notes: Clustered standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Wild bootstrap p-values reported in square brackets. Pooled Accountability is binary variable that is equal to one if the participant is assigned to any of the accountability treatments (PD, PD-R, or MP). The Complaint Box (CB) treatment is the omitted category. Returned/Received ratio measures the proportion of money providers send back to clients relative to the amount initially sent by clients. This can range from 0 to 3. Columns 1 and 2 use the continuous variable, while Columns 3 to 6 use binary variables. Basic Controls include dummies for the amount received from clients and the percent returned in the Baseline Trust Game. Additional controls include age, gender, private sector, health worker job title (a clinical officer vs nurse), how many patients the provider knows, and a dummy for knowing at least one other health provider.

Table 9

Treatment Effects by Sector of Employment

	Returned/Received Ratio (R/R)	Returned/Received Ratio (R/R) ≥ 1	Returned/Received Ratio (R/R) >1
	[1]	[2]	[3]
Pooled Accountability	0.42***	0.21***	0.17**
	(0.10)	(0.06)	(0.07)
	[0.01]	[0.00]	[0.10]
Private	-0.05	0.07	-0.09
	(0.13)	(0.10)	(0.09)
	[0.70]	[0.57]	[0.33]
Pooled Accountability*Private	-0.33*	-0.21	-0.15
	(0.17)	(0.13)	(0.11)
	[0.12]	[0.21]	[0.21]
Constant	0.39	0.32**	0.09
	(0.3)	(0.12)	(0.21)
	[0.23]	[0.02]	[0.68]
Observations	412	412	412
R-squared	0.51	0.33	0.34
Mean of control group (CB)	1.04	0.62	0.45
P-Values for Test of Sum of Coefficients			
P-value Pool+Pool*Priv=0	0.52	0.99	0.87

Notes: Clustered standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Wild bootstrap p- values reported in square brackets. Pooled Accountability is binary variable that is equal to one if the participant is assigned to any of the accountability treatments (PD, PD-R, or MP). The Complaint Box (CB) treatment is the omitted category Returned/Received ratio measures the proportion of money providers send back to clients relative to the amount initially sent by clients. This can range from 0 to 3. Columns 1 and 2 use the continuous variable, while Columns 3 to 6 use binary variables. Basic Controls include dummies for the amount received from clients and the percent returned in the Baseline Trust Game. Additional controls include age, gender, private sector, health worker job title (a clinical officer vs nurse), how many patients the provider knows, and a dummy for knowing at least one other health provider.

Table 10

Heterogeneity by Provider Reciprocity in the Baseline Trust Game

	Returned/Received Ratio (R/R)	Returned/Received Ratio (R/R) ≥ 1	Returned/Received Ratio (R/R) >1
	[1]	[2]	[3]
Pooled Accountability	0.56***	0.38***	0.32***
	(0.16)	(0.08)	(0.10)
	[0.02]	[0.01]	[0.02]
Baseline reciprocity	0.88***	0.38***	0.46***
	(0.10)	(0.06)	(0.08)
	[0.00]	[0.00]	[0.00]
Pooled Accountability* Baseline reciprocity	-0.23**	-0.20***	-0.18**
	(0.09)	(0.05)	(0.06)
	[0.05]	[0.01]	[0.03]
Constant	0.27	0.20	-0.02
	(0.25)	(0.12)	(0.18)
	[0.32]	[0.12]	[0.90]
Observations	412	412	412
R-squared	0.58	0.37	0.38
Mean of control group (CB)	1.04	0.62	0.45
P-Values for Test of Sum of Coefficients			
P-value Pool+Pool*Priv=0	0.00	0.01	0.03

Notes: Clustered standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Wild bootstrap p-values reported in square brackets. Pooled Accountability is binary variable that is equal to one if the participant is assigned to any of the accountability treatments (PD, PD-R, or MP). The Complaint Box (CB) treatment is the omitted category. Returned/Received ratio measures the proportion of money providers send back to clients relative to the amount initially sent by clients. This can range from 0 to 3. Columns 1 and 2 use the continuous variable, while Columns 3 to 6 use binary variables. Baseline Reciprocity is the R/R ratio in the Baseline Trust Game played before the Reporting Game. Controls include dummies for the amount received from clients, the percent returned in Baseline Trust Game, age, gender, private sector, health worker job title (a clinical officer vs nurse), how many patients the provider knows, and a dummy for knowing at least one other health provider.

Table 11

Patients Trusting Decisions in the Reporting Game

	Sent Positive Amount to Provider (OLS) Full Sample		Amount Sent to Provider (OLS) (Conditional on sending ≥ 0)	
	[1]	[2]	[3]	[4]
Pooled Accountability	0.13*		29.78	
	(0.07)		(40.79)	
	[0.06]		[0.51]	
Peer Disclosure (PD)		0.00		20.75
		(0.08)		(43.84)
		[1.00]		[0.67]
Peer Disclosure with Retaliation (PD-R)		0.23***		-3.33
		(0.07)		(41.32)
		[0.00]		[0.95]
Monetary Penalty (MP)		0.17**		60.67
		(0.07)		(38.7)
		[0.04]		[0.19]
Constant	0.64***	0.59***	193.65*	182.28*
	(0.21)	(0.20)	(94.52)	(94.84)
	[0.00]	[0.01]	[0.05]	[0.07]
Observations	204	204	154	154
R-squared	0.07	0.11	0.06	0.08
Mean of control group (CB)	0.64	0.64	325.00	325.00
P-Values for Test of Equality of Coefficients				
PD=MP		0.02		0.12
PD=PD-R		0.00		0.32
PD-R=MP		0.37		0.01

Notes: Clustered standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Wild bootstrap p-values reported in square brackets. Pooled Accountability is binary variable that is equal to one if the participant is assigned to any of the accountability treatments (PD, PD-R, or MP). The Complaint Box (CB) treatment is the omitted category. Sent Positive Amount is a binary variable that is equal to one if patients sent any money to providers. Amount sent is conditional on sending a positive amount. This can range from 0 to 800, in multiples of 200. Additional controls include age, gender, wealth, education, whether the patient recognizes at least one health worker, satisfaction with the care they received, and the patient's social status ranking of doctors.

Table 12
Willingness to Complain and Intensity of Complaints

	Complain		Number of cards sent	
	[1]	[2]	[3]	[4]
Pooled Accountability	-0.09**		0.80***	
	(0.04)		(0.25)	
	[0.08]		[0.01]	
Peer Disclosure (PD)		-0.11**		1.04***
		(0.04)		(0.32)
		[0.04]		[0.02]
Peer Disclosure with Retaliation (PD-R)		-0.07		0.49*
		(0.05)		(0.26)
		[0.16]		[0.07]
Monetary Penalty (MP)		-0.09*		0.83**
		(0.04)		(0.32)
		[0.11]		[0.04]
Constant	0.48***	0.46***	2.01**	2.33***
	(0.11)	(0.11)	(0.78)	(0.79)
	[0]	[0]	[0.01]	[0.01]
Observations	879	879	194	194
R-squared	0.03	0.03	0.09	0.10
Mean of control group (CB)	0.29	0.29	2.44	2.44
P-Values for Test of Equality of Coefficients				
PD=MP		0.66		0.55
PD=PD-R		0.46		0.07
PD-R=MP		0.74		0.30

Notes: Clustered standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Wild bootstrap p-values reported in square brackets. Pooled Accountability is binary variable that is equal to one if the participant is assigned to any of the accountability treatments (PD, PD-R, or MP). The Complaint Box (CB) treatment is the omitted category. Complain is a binary variable that is equal to one if patients sent at least one frowning face card to a provider. Number of Cards sent measures the number of frowning face (complaint) cards sent to providers. This is conditional on complaining. The specifications include controls for age, gender, wealth, education, whether the patient recognizes at least one health worker, satisfaction with the care they received, and the patient's social status ranking of doctors.

Appendix

Figure A.1
Timing and structure of data collection

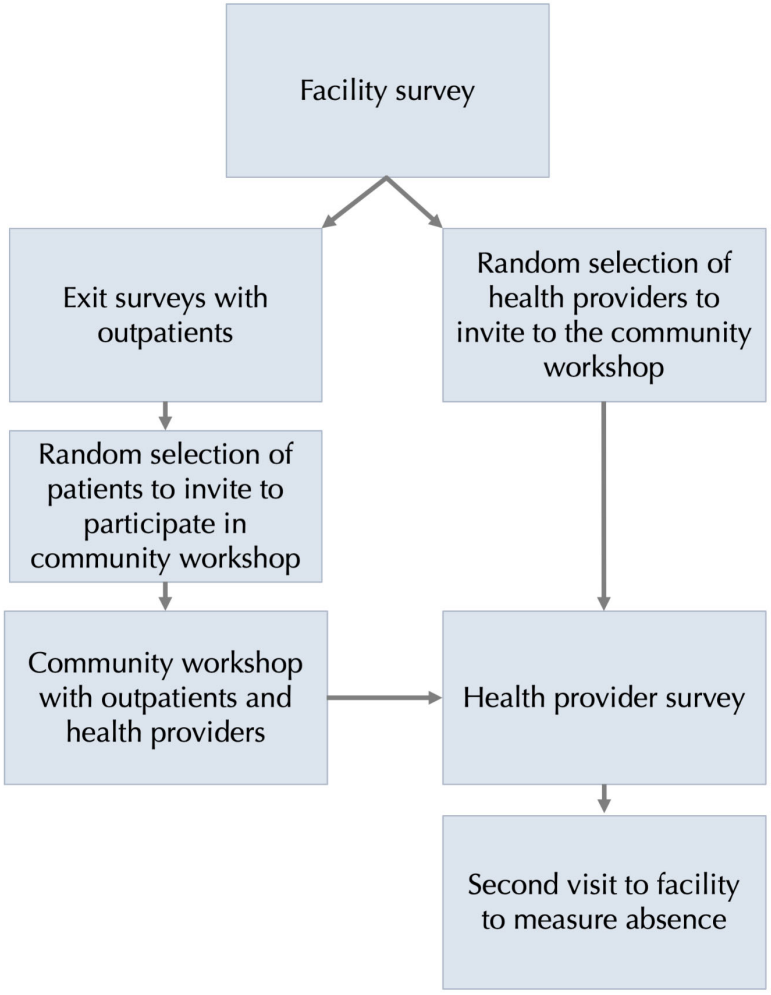


Table A.1
Providers' responsiveness to different reporting systems (strategy) - without additional controls

	Returned/Received Ratio (R/R)		Returned/Received Ratio (R/R) ≥ 1		Returned/Received Ratio (R/R) >1	
	[1]	[2]	[3]	[4]	[5]	[6]
Pooled Accountability	0.30***		0.13**		0.11*	
	(0.07)		(0.06)		(0.06)	
	[0.01]		[0.06]		[0.08]	
Peer Disclosure (PD)		0.39***		0.16**		0.19*
		(0.14)		(0.07)		(0.1)
		[0.02]		[0.07]		[0.13]
Peer Disclosure with Retaliation (PD-R)		0.22***		0.09		0.08
		(0.07)		(0.05)		(0.06)
		[0.02]		[0.15]		[0.18]
Monetary Penalty (MP)		0.29**		0.13		0.07
		(0.12)		(0.08)		(0.08)
		[0.05]		[0.16]		[0.42]
Constant	0.36***	0.36***	0.41***	0.41***	0.04	0.04
	(0.08)	(0.09)	(0.07)	(0.07)	(0.04)	(0.05)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.38]	[0.33]
Observations	412	412	412	412	412	412
R-squared	0.47	0.48	0.28	0.29	0.29	0.29
Mean of control group (CB)	1.04	1.04	0.63	0.63	0.45	0.45
P-Values for Test of Equality of Coefficients						
MP=PD		0.58		0.66		0.38
MP=PD-R		0.60		0.55		0.93
PD=PD-R		0.22		0.20		0.31

Notes: Clustered standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Wild bootstrap p-values reported in square brackets. Pooled Accountability is binary variable that is equal to one if the participant is assigned to any of the accountability treatments (PD, PD-R, or MP). The Complaint Box (CB) treatment is the omitted category. Returned/Received ratio measures the proportion of money providers send back to clients relative to the amount initially sent by clients. This can range from 0 to 3. Columns 1 and 2 use the continuous variable, while Columns 3 to 6 use binary variables. Basic Controls include dummies for the amount received from clients and the percent returned in the Baseline Trust Game. Additional controls include age, gender, private sector, health worker job title (a clinical officer vs nurse), how many patients the provider knows, and a dummy for knowing at least one other health provider.

Table A.2
Amounts sent by patients in the Reporting Game

	Complaint Box (CB)	Monetary Penalty (MP)	Peer Disclosure (PD)	Peer Disclosure with Retaliation (PD-R)	Pooled Accountability
% sending a positive amount	64%	83%**	67%	89%***	80%
Trusting patient sample					
% sending 200	50%	51%	44%	44%	46%
% sending 400	25%	28%	38%	40%	36%
% sending 600	19%	12%	10%	10%	11%
% sending 800	3%	0%	3%	2%	2%

Note: The asterisks refer to P-values are generated from two-sided tests of equality of means across samples, where CB is the comparison treatment. *** p<0.01, ** p<0.05, * p<0.10. Pooled Accountability is binary variable that is equal to one if the participant is assigned to any of the accountability treatments (PD, PD-R, or MP).

Table A.3

Patients Trusting Decisions in the Reporting Game - without additional controls

	Sent Positive Amount to Provider (OLS)		Amount Sent to Provider (OLS)	
	Full Sample		(Conditional on sending ≥ 0)	
	[1]	[2]	[3]	[4]
Pooled Accountability	0.16***		11.36	
	(0.05)		(39.18)	
	[0.01]		[0.78]	
Peer Disclosure (PD)		0.03		8.33
		(0.06)		(42.68)
		[0.62]		[0.85]
Peer Disclosure with Retaliation (PD-R)		0.25***		-13.00
		(0.06)		(40.80)
		[0.00]		[0.78]
Monetary Penalty (MP)		0.19***		42.44
		(0.06)		(39.83)
		[0.01]		[0.33]
Constant	0.64***	0.64***	325***	325***
	(0.04)	(0.04)	(37.75)	(37.99)
	[0.00]	[0.00]	[0.00]	[0.00]
Observations	216	216	164	164
R-squared	0.02	0.06	0.00	0.02
Mean of control group (CB)	0.64	0.64	325.00	325.00
P-Values for Test of Equality of Coefficients				
PD=MP		0.03		0.15
PD=PD-R		0.00		0.39
PD-R=MP		0.32		0.01

Notes: Clustered standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Wild bootstrap p-values reported in square brackets. Pooled Accountability is binary variable that is equal to one if the participant is assigned to any of the accountability treatments (PD, PD-R, or MP). The Complaint Box (CB) treatment is the omitted category. Sent Positive Amount is a binary variable that is equal to one if patients sent any money to providers. Amount sent is conditional on sending a positive amount. This can range from 0 to 800, in multiples of 200. Additional controls include age, gender, wealth, education, whether the patient recognizes at least one health worker, satisfaction with the care they received, and the patient's social status ranking of doctors.

Table A.4

Willingness to Complain and Intensity of Complaints - without additional controls

	Complain		Number of cards sent	
	[1]	[2]	[3]	[4]
Pooled Accountability	-0.09***		0.55**	
	(0.03)		(0.20)	
	[0.06]		[0.02]	
Peer Disclosure (PD)		-0.10**		1.05***
		(0.04)		(0.30)
		[0.04]		[0.02]
Peer Disclosure with Retaliation (PD-R)		-0.10**		0.09
		(0.03)		(0.18)
		[0.04]		[0.63]
Monetary Penalty (MP)		-0.08**		0.66**
		(0.03)		(0.27)
		[0.05]		[0.06]
Constant	0.29***	0.29***	2.44***	2.44***
	(0.03)	(0.03)	(0.11)	(0.11)
	[0.00]	[0.00]	[0.00]	[0.00]
Observations	931	931	203	203
R-squared	0.01	0.01	0.03	0.07
Mean of control group (CB)	0.29	0.29	2.44	2.44
P-Values for Test of Equality of Coefficients				
PD=MP		0.60		0.30
PD=PD-R		0.85		0.01
PD-R=MP		0.65		0.06

Notes: Clustered standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Wild bootstrap p-values reported in square brackets. Pooled Accountability is binary variable that is equal to one if the participant is assigned to any of the accountability treatments (PD, PD-R, or MP). The Complaint Box (CB) treatment is the omitted category. Complain is a binary variable that is equal to one if patients sent at least one frowning face card to a provider. Number of Cards sent measures the number of frowning face (complaint) cards sent to providers. This is conditional on complaining. Basic controls include broad controls for amount sent to providers and the amount received back.

Table A.5
Treatment effects by knowledge of provider

	Complain	
	[1]	[2]
Pooled Accountability	-0.06**	
	(0.03)	
	[0.07]	
Knowledge of provider	0.03	0.03
	(0.05)	(0.04)
	[0.59]	[0.53]
Pooled Accountability* Knowledge of provider	-0.07	
	(0.05)	
	[0.28]	
PD		-0.06
		(0.05)
		[0.25]
PD-R		-0.06*
		(0.04)
		[0.13]
MP		-0.04
		(0.05)
		[0.46]
PD*Knowledge of provider		-0.08
		(0.10)
		[0.50]
PD-R*Knowledge of provider		-0.03
		(0.06)
		[0.62]
MP*Knowledge of provider		-0.12**
		(0.06)
		[0.10]
Constant	0.21*	0.20*
	(0.11)	(0.11)
	[0.09]	[0.10]
Observations	879	879
R-squared	0.43	0.43
Mean of control group (CB)	0.29	0.29

Notes: Clustered standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Wild bootstrap p-values reported in square brackets. Pooled Accountability is binary variable that is equal to one if the participant is assigned to any of the accountability treatments (PD, PD-R, or MP). The Complaint Box (CB) treatment is the omitted category. Complain is a binary variable that is equal to one if patients sent at least one frowning face card to a provider. Basic controls include broad controls for amount sent to providers and the amount received back. Additional controls include age, gender, wealth, education, whether the patient recognizes at least one health worker, satisfaction with the care they received, and the patient's social status ranking of doctors.