Published in final edited form as:

J Health Econ. 2018 March ; 58: 188–201. doi:10.1016/j.jhealeco.2018.02.007.

Giving and Promising Gifts: Experimental Evidence on Reciprocity from the Field

J. Michelle Brock,

European Bank for Reconstruction and Development and CEPR, One Exchange Square, London, UK EC2A 2JN (brockm@ebrd.com)

Andreas Lange, and

University of Hamburg, Department of Economics, Von Melle Park 5, 20146 Hamburg, Germany (Andreas.Lange@wiso.uni-hamburg.de)

Kenneth L. Leonard

Department of Agricultural and Resource Economics, University of Maryland, 2200 Symons Hall, University of Maryland, College Park 20742 (kleonard@arec.umd.edu)

Abstract

We test the value of unconditional non-monetary gifts as a way to improve health worker performance in a low income country health setting. We randomly assigned health workers to different gift treatments within a program that visited health workers, measured performance and encouraged them to provide high quality care for their patients. We show that unconditional non-monetary gifts improve performance by 20 percent over a six-week period, compared to the control group. We compare the impact of the unconditional gift to one in which a gift is offered conditional on meeting a performance target and show that only the unconditional gift results in a statistically significant improvement. This demonstrates that organizations can improve the performance of health workers in the medium term without using financial incentives.

Keywords

gift exchange; reciprocity; health care; field experiment; Tanzania

There is significant empirical evidence that unconditional payments can improve employee performance (Gneezy & List, 2006; Rigdon, 2002). As suggested by Akerlof (1982), gift exchange is one way to understand the relationship between wages and effort: employees may respond to a “gift” of unconditionally higher wages with a more than reciprocal level of effort. In addition, the gift exchange framework can be applied to non-monetary incentives that may also lead to significant improvements in performance, as has been shown in experimental studies. Importantly, non-monetary incentives might work better than money in

Correspondence to: Kenneth L. Leonard.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

signaling when effort should be independent of compensation (Gneezy & Rustichini, 2000; Heyman & Ariely, 2004). For example, subjects paid in candy (compared to cash) provide effort that is invariant to the rate (Heyman & Ariely, 2004) and subjects given a water bottle as a gift outperform those given cash of equal value (Kube et al., 2012). Even folding cash into origami outperforms pure cash in this setting. The fact that subjects in experimental settings might see non-monetary gifts as a signal that their effort is valued differently ties in with the literature on intrinsic motivation. Indeed, switching from non-monetary to monetary payments can decrease performance in some contexts (Gneezy & Rustichini, 2000) as compensation can crowd out intrinsic motivation.

In this paper we test the value of non-monetary unconditional gifts as a way to improve performance in a health care setting. We gave books to randomly selected clinicians working in outpatient settings in urban and peri-urban Tanzania and asked them to work harder. Their performance is compared to clinicians in the control group who were also asked to work harder, but were not given any compensation. Performance was evaluated over a period of approximately 10 weeks. We show significant improvements in performance for the clinicians who received the gift compared to clinicians who did not. Importantly, the gains we observe are still present after 10 weeks, demonstrating an important medium term effect from this simple intervention.

Our choice of health workers in a developing-country context was deliberate as this is a context in which gifts might serve a particularly important policy role. Health care in general is a setting in which effort is difficult for employers to observe (or for patients to evaluate) and almost impossible to verify. In Tanzania, as in most developing country settings, a significant gap exists between effort provided by health workers and their capacity (Das & Hammer, 2007; Das et al., 2008; Maestad & Torsvik, 2008). Nonetheless, health workers in these settings are commonly described as being motivated by intrinsic rewards. The literature on health care is full of references to terms such as *professionalism*, *esteem*, and *caring* (Freidson, 1970; Lindelow & Serneels, 2006; Mathauer & Imhoff, 2006; Serra et al., 2011). Given that reliance on the prosocial instincts of health care workers has failed to assure quality and that most developing countries lack the institutional infrastructure to effectively regulate quality, attention has turned to other forms of motivation, particularly monetary incentives to provide specific inputs.¹ However, paying health workers to increase their workload is not the same thing as paying them to increase quality. Writing contracts based on quality is likely to be much more difficult. Thus, in such settings, gifts and bonuses may help to solve incentive problems that have otherwise proven difficult to address.

The main treatment in our study is giving subjects an unconditional gift. In this treatment, the gift was given at the same time as they were asked to work harder. In order to better understand the way the gift was received we used two additional types of gifts, randomly assigned among two additional treatment groups. A second group was told that they would be given a “gift” later, if their performance on the mentioned tasks improved. This treatment

¹There is evidence that direct incentives (pay for performance) and organizational incentives (supervision combined with institutionalized rewards or punishments) do lead to improved quantity of care. See Eichler & Levine (2009), for an extended discussion of pay for performance; and Basinga et al. (2011); Meessen et al. (2006), for early evidence of success.

(which we call the prize) was not designed to test whether conditional prizes can work (there is significant literature to show that they can) but rather to see if the exact same “gift” worked better in a conditional or an unconditional setting. Because a conditional prize implies a follow up visit (to award the prize) and some feedback on performance (receipt of the prize would signal improved performance) we also used a follow up visit for the unconditional gift treatment. Thus the only differences between these two treatments were the timing of the gift (immediate or follow-up) and the conditionality of the gift. To further explore the role of timing, we also introduced a treatment in which the gift was given at the same time as the follow-up visit. This treatment did not include any feedback at the follow-up visit to avoid making the gift appear as if it was awarded as a prize. This treatment allows us to see if gifts are valuable because they are immediate, or if they are valuable because they are gifts. Importantly, we measured performance after the follow-up visit, which allows us to measure the impact on the delayed treatment category of having received the gift. In contrast, no explicit incentives were offered to the control group. Importantly, our study could still affect the control group as all subjects in the research study were enrolled, visited once by a clinician who observed their practice, visited later by a doctor who encouraged them to improve their performance and specifically asked to improve performance on a particular list of tasks. Each clinician was told that the research team would interview their patients over an extended period of time.²

Our results show significant improvements in performance compared to the control after receiving a gift, but no significant improvement in this first period when the gift was offered as a prize or when the clinician was told they would later receive the gift. For the delayed gift treatment, a significant improvement in performance occurs after the gift is finally awarded and is essentially identical in size to the improvement seen in the first period for the immediate gift treatment. By comparing performance on tasks that were part of the encouragement script (primed tasks) to performance on other tasks which were measured but never mentioned (un-primed tasks), we can show that there was no task shifting in any of the treatments: we observed improvements in both specified and unspecified tasks. Importantly, by using patient exit interviews to measure adherence—a measure we explicitly validate—we are able to observe clinician performance when the clinician does not know he or she is being observed.

Our work builds on the strands of literature that combine gift exchange, non-monetary incentives and duration effects. There are a few other studies that test the effectiveness of non-monetary incentives for motivating performance in the field. Kosfeld & Neckermann (2011) and Bradler et al. (2013) look at whether students hired to do a one-time data entry job perform better when put in a tournament situation, where winners get a non-pecuniary, publicly announced award (a card of recognition signed by a prestigious figure). Their work is based on the idea that awards are valuable to workers because they contribute to increased self-esteem and they distinguish the winner’s status among his or her peers. In this one-shot

²These activities are similar to activities carried out in hundreds of research projects that have been conducted in health care (Jamtvedt et al., 2003, 2006, 2013; Rowe et al., 2005). These meta-studies examine interventions that provide information to clinicians about better practices as well as varying degrees of follow-up, feedback and contact. They find that information alone does not improve performance, but that information combined with subsequent attention—similar to the intervention received by the control group—does improve performance in many studies.

setting, they do find positive and significant effects from symbolic awards. Bradler et al. (2013) even find increases in effort from an unconditional prize, though the response is less substantial. In both studies, the public nature of the award matters. However, these results are short term and it is not clear that this kind of incentive structure is sustainable or repeatable in a real workplace.

Ashraf et al. (2012) also study awards as non-monetary incentives in the health setting. Their field experiment in Zambia compares trainees' sales of condoms under monetary and non-monetary incentives. As in Kosfeld & Neckermann (2011) and Bradler et al. (2013), the non-monetary incentive used is an award that is publicly given out according to a tournament and conditional on performance. They find that only the subjects in the award treatment group perform significantly better than the baseline (where trainees are not paid). But while they can juxtapose the impact of monetary and non-monetary incentives between subjects, the non-monetary incentive involves at least 3 levels of potential motivation: social comparison and status value, the satisfaction of winning itself, and utility from competition. In our work, we attempt to more precisely identify the value of a gift by removing the social recognition and competition dimensions, which we do by offering an unconditional gift in two of the treatments, and awarding each participant's gift in private.

In a setting similar to ours, Currie et al. (2013) test whether a patient receives better care after giving a token (non-monetary) gift to a medical clinician. They also find that gifts increase effort and, interestingly, show that gifts can have implications for both the giver and for others as well. In their setting, gifts from one patient decreased the quality of care provided to other patients if the two patients were perceived by the clinician as unrelated. In our setting, the gift from the our team is explicitly intended to increase effort for others: the patients of the clinician.

The remainder of this paper is structured as follows. In section 1 we discuss the data, our experimental design, and our estimation strategy. In section 2 we review the results of the experiment, and in section 3 we conclude.

1 Methodology

Our sample includes 103 clinicians in the urban and suburban area of Arusha in northeastern Tanzania.³ The sample is representative of the clinicians who see reasonable numbers of patients in the study region and is therefore a policy-relevant population. The sample includes health workers (clinicians) who are trained to provide outpatient care in public, private, and non-profit/charitable facilities in the area.⁴ Of the 103 clinicians we initially

³We initially estimated that there were about 200 clinicians in the sample area and planned to randomly sample clinicians from this population, estimating that with a sample of 100 clinicians (25 in each treatment), we could measure policy-relevant changes in effort at the 10% significance level. However, once we were in the field, we discovered that many of these clinicians did not see large numbers of patients on a regular basis and others were difficult to reach to enroll in the study. As a result, we switched to a convenience sample in which we studied all clinicians who were present at any of a series of visits to facilities that we could easily reach in the sample area.

⁴The 4 cadres of clinicians include assistant clinical officer (ACO), clinical officer (CO), assistant medical officer (AMO), and medical officer (MO). The medical training required for each depends on the degrees an individual already has. Typically, ACOs have the least amount of training, essentially specialized secondary schooling. With no other degrees and 4 years of secondary school, it requires 3 years of training to become a CO. AMOs have on average 3.5 years of medical schooling, and MOs have the equivalent of a United States MD degree.

enrolled, 12 dropped out before they were assigned to treatment and an additional 3 dropped out after assignment.⁵ Importantly, however, there were no differences in characteristics at the baseline for those who attrited.

The field data collection ran from November 2008 until August 2010. We collected data on the clinicians' performance using 4,379 post consultation interviews with their patients. To collect these data, we visited each clinicians' practice on at least 5 separate occasions (once in the baseline and two each for Period 1 and Period 2) over approximately two months. All data collection visits were unannounced.

1.1 Measuring Performance

We measure quality of care by comparing clinician activities to that which is required by protocol for the presenting condition. In other words, according to national guidelines there are certain diagnostic steps that should be taken for every patient who presents with a particular symptom (for example a fever). These activities, referred to here as tasks, include asking certain questions ("how long have you had a fever?"), performing particular physical examinations (checking the patients temperature) and educating the patient (explaining if and when they should return for a follow up visit).

In order to measure the activities of each clinician without affecting their effort (they would exert more effort if they knew a researcher was watching), we used the Retrospective Consultation Review (RCR) instrument (Leonard & Masatu, 2006) in which patients are asked to recall the activities of the clinicians soon after their consultation is complete. Each patient is asked a series of questions such as "did the clinician ask you how long you had had a fever?" The discrete tasks required by protocol differ according to the presenting symptoms and age of the patient. We use the protocols for four types of symptoms (fever, cough, diarrhea and general) and two types of patients (older than or younger than 5 years). During the RCR interview, patients are only asked about tasks that apply to their symptoms and age category. Importantly, we also have data on adherence to protocol collected by clinicians during the scrutiny visit and can therefore show that patient recall matches clinician observation reasonably well (see Appendix A.2). Across these symptoms, there are 74 possible tasks, however only a subset will apply to a particular type of patient. We use the performance on these items to investigate the impact of our treatments.

1.2 Experimental Design

The value of gifts is measured in the context of a standard health quality improvement intervention, one in which clinicians are enrolled in the study, observed by researchers and then provided information on how to improve their practice. Thus, every clinician (including the control) received an intervention which we refer to as the encouragement intervention. Within this intervention we test the value of three additional treatments which used gifts to improve performance. We outline the encouragement intervention first and then describe these additional treatments.

⁵All but two dropped out because they were reassigned to different posts.

1.2.1 The Encouragement Intervention—The order of the overall intervention is shown in Figure 1. In the course of the intervention we met with each clinician three times: once to obtain consent and enroll subjects, once to encourage improvements in performance and once to provide feedback. To measure the impact of these elements, we collected data after each of the three visits: the baseline, Period 1 and Period 2.

Our first meeting with clinicians was to enroll them in the study and then, about a week later (unannounced) we returned to collect data from their patients to measure performance. This baseline data collection had three parts, measuring effort in (i) pre peer scrutiny, (ii) peer scrutiny, and (iii) post peer scrutiny. Usually all three of these parts took place on the same day. The purpose of the baseline data collection was to get a measure of clinician performance before the randomization into treatments. By interviewing patients who had already visited with the clinician, the pre peer scrutiny part of the data collection measures performance before we arrived and therefore is not impacted by our presence and can serve as an estimate of performance in the absence of any research. In the next phase—peer scrutiny—a member of our team sat in the examination room observing the consultations, invoking the well-documented Hawthorne effect in which the health worker significantly increases his or her effort when faced with outside scrutiny (Leonard & Masatu, 2006). For the third portion of the baseline visit, we measured effort after the peer had left the room: post peer scrutiny. Note that, from the perspective of the clinician, they received a visit from the research team; the pre and post peer scrutiny visits are not observable to them.

Encouragement: After the baseline data collection, clinicians were visited by Dr. Beatus, a Tanzanian M.D. and lecturer at a health research institution,⁶ who met with them and read the following encouragement script:

We appreciate your participation in this research study. The work that you do as a doctor is important. Quality health care makes a difference in the lives of many people. Dedicated, hard working doctors can help us all achieve a better life for ourselves and our families.

One important guideline for providing quality care is the national protocol for specific presenting symptoms. While following this guideline is not the only way to provide quality, we have observed that better doctors follow these guidelines more carefully. Some of the protocol tasks that we have noticed to be particularly important are telling the patient their diagnosis, explaining the diagnosis in plain language, and explaining whether or not the patient needs to return for further treatment. In addition it is important to determine if the patient has received treatment elsewhere or taken any medication before seeing you, and to check the patient's temperature, and check their ears and/or throat when indicated by the symptom. For this research, we look at clinician adherence to these specific protocol tasks.

⁶Dr. Beatus represented CEDHA, a quasi-public research centre whose organizational goal is “to contribute to quality health care delivery through human resource development, conducting operational research, offering consultancy services and networking.” His request that health workers provide important inputs is credible in the sense that health workers would be likely to believe this request aligns with organizational goals. Furthermore, it is not likely to be in conflict with any pre-existing organizational expectations; employers may not care as much about quality as Dr. Beatus does, but they are not opposed to quality.

The purpose of the encouragement message was to create a consistent level of implied scrutiny and information similar to most standard quality improvement interventions common in health care. Further, it allowed us to prime five tasks—in a salient and natural way—which we could then use to measure task shifting. We chose these 5 tasks because previous work with these protocols shows that 1) good doctors are much more likely to do these things than poor doctors; 2) average adherence for these tasks is low; 3) when observed by peers, most doctors significantly increase their adherence to these tasks (indicating that they know how and when to do these things, but choose not to); 4) they apply to most patients and symptoms (so that it is easier to collect the required data during the data-collection visits); and 5) patients have a relatively accurate recall of whether or not these things were done (patient reports agree with the reports of research team observers). Therefore, we can collect relatively accurate data on tasks that are important and that almost any clinician can adhere to if he or she so chooses.

Period 1: Following the encouragement visit, our enumerators returned on two separate days to collect more performance data, referred to as Period 1. Unlike the peer-scrutiny visit, the enumerators did not directly observe the clinicians in their practice, did not announce their presence and had no contact with the staff during these visits. This does not mean that the visits had no impact: after the team had left, we suspect most clinicians realized they had been visited by the team. Thus, although the data collected are an accurate representation of the quality of care that would have been provided on that day, our team's presence should have served as a reminder that we were still conducting research and had knock-on effects on quality subsequent to the visit, which would be reflected in later visits.

Follow-Up: After the Period 1 data collection, clinicians received a follow-up visit from Dr. Beatus (outlined below). Clinicians had no further meetings with the research team after the follow-up visit and no data was collected during this visit.

Period 2: After the follow-up visit enumerators returned to collect data on clinicians' effort (again on two separate days). These are referred to as Period 2.

In all, we collected data from 1,496 unique patients in the baseline, 1,557 unique patients during the Period 1, and 1,220 unique patients during Period 2. Most of the Period 1 data collection took place within three weeks of the initial quality-assessment visit for each clinician, though some were as late as 2 months after the initial visit. The Period 2 visits occurred much later, with some as early as 4 weeks after the initial visit but most about 10 weeks after that visit.

1.2.2 The Experimental Treatments—Within the context of this standard encouragement intervention we created three gift treatments to compare to the control. Clinicians were randomly assigned to one of these four groups after the baseline data collection and informed of the details of the treatment during the encouragement visit. The study was blind, meaning that neither patients nor the enumerators collecting the data from them knew the treatment assignment of clinicians. Also, clinicians were not told details about other treatments.⁷ Treated clinicians were given—or offered the opportunity to earn—a book about doctors in the developing world titled “Mountains Beyond Mountains” by

Tracy Kidder and inscribed with a thank you message from the research team. This gift is relevant to the subjects as clinicians and may serve to remind or inspire them to increase effort, but it is not substantial enough to be considered a financial (or material) incentive for increasing effort.⁸ In all cases, the book was given to each participant in a private setting to avoid social recognition and competition among participants. In order to control for the fact that the participants might read and be inspired by the contents of the book, we read two short passages from the book to every participant in the experiment as part of the encouragement script (for all clinicians including the control). In addition, after the study was over, we verified that only one participant in the study had read any part of the book. Thus, the book is salient and inspiring because it was a gift, not because of its contents.

The control, gift, delayed gift, and prize treatments: The first of the 3 treatment groups (*gift*) was given a gift immediately and unconditionally, the second group (*delayed gift*) was promised the gift at the follow-up visit (also unconditionally), and the third (*prize*) was told they would be given the gift if they demonstrated adequate adherence to protocol for the five specific tasks discussed in the encouragement script. Participants in the control group did not receive any gift and did not receive the follow-up visit. Note that, although they did not receive a visit from the team during what would have been feedback, the control group did know that data would be collected on their performance in Periods 1 and 2.

The first treatment (*gift*) is the unconditional gift, given at the same time as the encouragement script was read. Clinicians were told that there would be an additional visit later where we would share their performance results with them. During this later visit, we provided some feedback on their performance but there was no further encouragement or gift. The second treatment is the delayed, unconditional gift treatment (*delayed gift*). After listening to the encouragement script, these clinicians were told that we were going to give them a gift at a future visit. They received that gift at the follow-up visit, but did not receive further feedback. The third treatment is the conditional gift (*prize*). Clinicians in this group were told (after the encouragement script): “We will present the gift to you if you perform these protocols 70% of the time when it is appropriate to perform them, given the symptoms.” The gift was then awarded, if it was earned, at the follow-up visit (the same timing as *delayed gift*). The prize treatment was not a competition among the clinicians, and clinicians did not know whether others had earned the prize. The 70% threshold level was chosen so that most health workers would be able to earn the prize: in fact, 90% of the clinicians in the *prize* treatment did earn the prize. Since earning the prize is endogenous, we do not control for this in any of our empirical work below. Note that, to award a prize, some feedback must be implied and we therefore included the same feedback to the *gift* treatment, allowing a better comparison. The feedback in each treatment was neutral and was not intended to impact effort directly.⁹

⁷There is some suggestive evidence of spillovers, where clinicians reacted to the treatment differently if someone else in their facility was in a different treatment category. Controlling for all possible spillovers does not change the results shown in this paper significantly. Results available on request.

⁸There is a used book market in Tanzania and this book could potentially be sold for between 1 and 3 USD.

⁹We provided information on the frequency with which they performed the five specified tasks. The feedback was not expected to result in any independent performance increase in Period 1, before the feedback is even received. Indeed, Jamtvedt et al. (2003, 2006, 2013), meta-studies of audit and feedback studies show small effects after feedback is received but not before it is received.

Comparing the elements of the treatments: Table 1 shows the combination of types of gift (immediate, delayed and prize) and the use of feedback for the control and three treatments. Our intention in choosing the treatments was to test whether gifts could work to improve quality. We implemented the *prize* treatment as a comparison to the *gift* treatment to test if clinicians would respond better to the same gift if it was conditional. If the prize had a significantly better impact than the gift, that would challenge the idea that gifts are effective. Since a prize requires feedback—someone has to come and award the prize and the receipt of the prize indicates something about performance—it was clear that the best way to compare a gift and a prize was to include feedback.¹⁰ Thus the comparison between gift and prize is the most direct test of the role of conditionality in this experiment. However, due to the fact that a prize cannot be awarded immediately, we also introduced the *delayed gift* treatment with the same timing as the *prize* but without conditionality.¹¹

Our design also allows us to test for task shifting and crowding out because we collected performance data on more activities than were used for determining protocol adherence reported in the follow-up visits. Note that all participants, even the control group, were told about the five tasks important for providing quality care that we highlighted in the study. Thus, for all treatments, but especially for conditional gift (prize) treatment, one may expect some substitution of effort from unmentioned to mentioned tasks, particularly as the latter are used as the announced criterion of quality care.

1.3 Empirical Specification

Our outcome measure is the performance on tasks required by protocol for the patient's condition. We thus have a series of discrete events for each patient, and this patient is assigned to only one clinician (and therefore treatment). To estimate the impact of each treatment we use a multi-level model with nested random effects at the patient and clinician level and include both a task fixed effects and patient characteristics as controls. The multi-level model (Goldstein et al., 2002; Leyland & Goldstein, 2001; Turner et al., 2000) is commonly used to measure treatment effects in medical trials with multiple observations per patient when patients are not randomly assigned to doctors but treatments are randomly assigned across doctors, a structure which fits our data generation process. In this particular case, we use a multi-level model with the binary outcome variable indicating whether the clinician performed each required task, the mixed effect logit model.

¹⁰The multi-visit design may also trigger strategic behavior that goes beyond the actual experimental intervention. Subjects, for example, may reciprocate not in response to the one-off interventions, but because they want more gifts in the future or because they think that improved performance may lead to advantages beyond the experiment. While such strategic behavior might be limited in one-shot games in the lab, we perceive the dynamic interaction as an integral part of gift-exchange in the real world. In this sense, our intervention is designed to measure the strategic response to receiving a gift and to compare it to the strategic response of the control group.

¹¹We note that one could perceive four additional combinations of conditionality and timing that we did not implement: control with feedback, gift without feedback, delayed gift with feedback and prize without feedback. The latter two are unrealistic: (i) it is not possible to award a prize without returning to award it and without implying some level of performance was achieved; (ii) announcing a delayed gift and that feedback on performance will be given would likely appear to clinicians as if it really were a conditional gift. The remaining two treatments were feasible, but our sample size was limited and we decided to focus on the treatments that had the most direct comparison to the gift treatment. One may suspect, though, that control with feedback would not result in a significantly different Period 1 performance compared to the control: control with feedback resembles studies that combine audit and feedback and that have found only small effects after the feedback is received, i.e. not between when it is promised and when it was received (Jamtvedt et al., 2003, 2006, 2013). In addition, we can measure the impact of a gift without feedback using the *delayed gift* treatment in Period 2: although clinicians in the *gift* treatment expected feedback after Period 1, clinicians in the *delayed gift* treatment did not expect feedback after Period 2.

$$\text{prob}(x_{ijk} = 1) = T_i \vec{t} + \beta_k + \varepsilon_i + \varepsilon_j + e_{ijk} \quad (1)$$

The probability of a successful outcome is a function of treatment status, T_i , the task, k (some tasks require more effort), the patient, i (some patients receive more effort than others across all tasks) and the clinician that patients choose to visit, j (some doctors always do more for their patients). Since the treatment status is exogenously assigned at the clinician level, we control for task, patient and clinician level effects with fixed effects for the task (β_k) and nested random effects for patients (ε_i) and clinicians (ε_j). The response to each treatment is observed in two different periods (Period 1 and 2) resulting in 8 treatment status coefficients. In addition, before randomization each clinician is observed in the pre peer scrutiny, peer scrutiny and post peer scrutiny periods resulting in an additional three coefficients. To utilize clinician effects in this model, the pre peer scrutiny period is considered the comparison period and that coefficient is therefore omitted. In addition, because some of the tasks were primed and others were not we can test for the differential impact of treatment on primed and un-primed tasks, to look for evidence of task shifting.

A key assumption of random effects models is that the individual-specific effects are not correlated with any of the other regressors. By design, they cannot be correlated with any of the treatment variables (since these are randomly assigned), but we might be worried that they are correlated with our standard set of patient controls. In the regressions shown in the tables below we control for the type of task and the order in which patients were seen at the facility, but we have also run the same models excluding all controls except the treatment effects. The results for these models are virtually identical in the magnitude and significance of the coefficients suggesting that the random effects are appropriately used in this multi-level model.¹²

Finally, to verify the basic results of this model, we also use four additional econometric specifications as robustness checks: 1) Item Response Theory (also used in similar settings by Das & Hammer, 2005; Leonard et al., 2007) 2) logistic regression with errors clustered at the clinician level, 3) logistic regression with errors clustered at the facility level, and 4) OLS regression of the patient-level performance score with standard errors clustered at the clinician level. Note that by estimating a patient level score, we cannot test the difference between primed and un-primed tasks, since these are combined into one score. The additional specifications are discussed in more detail in Appendix A, along with the results and provide virtually the same interpretation as the main regression shown here.

2 Results

Table 2 shows the distribution of clinician characteristics across treatment groups to validate the balance of randomization. Clinicians in the *Prize* treatment are younger and less likely to practice in the public sector, but otherwise, there are not significant differences among the treatment groups. Importantly, average adherence and clinician quality (which controls for

¹²Results are available on request.

case mix)¹³ are not different in the baseline. Table 3 shows the balance in patient characteristics in the baseline including the adherence to protocol measured at the patient level in the baseline. Table 3 also shows balance in tasks showing probability of performing a task (unconditionally) in the Baseline, and the change in this probability in Period 1 and Period 2.

Patients in *control* treatment are less likely to be in the 5–15 year old age bracket, but otherwise there are no differences in patient characteristics across treatment categories in the baseline. The average adherence to protocol is also not different in the baseline. Importantly, if we look at the change in adherence to protocol we can see a preview of the results we show below: patients in the *gift* treatment experience a greater increase in protocol adherence in Period 1 and patients in the *delayed gift* treatment experience a greater increase in protocol adherence in Period 2. Because there is some variation in clinician and patient characteristics across the treatments we are careful to control for clinician and patient effects in the analyses that follow.

Recall that the dependent variable is whether the clinician completed any number of symptom-specific tasks required by protocol during a consultation. Table 4 shows the impact of treatments on the likelihood of a task being performed in Period 1 and Period 2 (shown in column 1 and column 2 respectively). Since all clinicians are examined in the baseline, we can estimate a participant effect and therefore the baseline is the omitted coefficient in our specification. To give perspective to the changes we examine in this table, we note that studies of intensive interventions in outpatient settings find improvements of between 2 and 5 percentage points in protocol adherence (Jamtvedt et al., 2003; Leonard & Masatu, 2006).

Overall intervention impacts

The overall short-term response after the encouragement visit is small and insignificant: being asked to work harder does not, by itself, lead to any wholesale increase in effort. Interestingly, the long run impact of encouragement combined with repeated contact with the research team leads to a large and significant increase in effort of at least 6 percentage points in period 2. This effect, showing that clinicians in our study responded to the fact that they were being studied, is in the spirit of a Hawthorne effect: even though not present in the room, the research team's occasional presence at the facility may have created a feeling of being scrutinized.¹⁴ Gosnell et al. (2016) find similar results in a very different setting: being part of a study has important implications for performance independent of treatment assignment. Importantly, this effect, if also present in other field settings, has methodological consequences: to investigate behavioral changes among treatment interventions, a properly defined control is necessary. Controls created based on performance levels *before* the start of interventions or drawn from groups that are not equally scrutinized after assignment are not valid and would lead to biased estimates of treatment effects.

¹³This score is the clinician effect θ_j from the IRT model discussed in Appendix A.1.

¹⁴Due to the short time frame of the study, it is unlikely that an effect of this size would be the result of generic quality increases over time, especially relative to the Hawthorne effect seen in the peer scrutiny coefficient, which we know is causal. Further, there were no additional programs or interventions at these facilities that coincided with our intervention, so we know that the increase cannot be due to some other factor consistent across all participants in the study.

Treatment assignment impacts

While clinicians react differently in the post-encouragement (Period 1) and post-study (Period 2) phases, our study was primarily designed to measure the different responses to the treatment across these phases. In Period 1, we find that clinicians in the *gift* treatment significantly increase their effort by over four percentage points above the baseline treatment. The effects for the *delayed gift* and *prize* treatments are smaller and insignificant (although the differences across the treatments are not significant.)

In Period 2, although adherence to protocol increases for all treatments, only the *delayed gift* treatment exhibits a significant increase in effort above that observed by the control group (4.8 percentage points). The response to receiving the unconditional gift during the encouragement visit for *gift* is the same as the response to receiving the unconditional gift at the follow-up visit for *delayed gift*. This shows that it is the receipt of the gift that matters for effort.

We interpret the performance response to gifts, when they are actually received, as reciprocation on the part of the clinician. The encouragement visit essentially invited the clinician to reciprocate by providing effort to a third party. A priori, we did not know whether being promised a gift or receiving it was more important. Promising a gift at a future date may induce clinicians to reciprocate with better performance. Due to discounting and uncertainty of actually receiving the gift, such an effect may be smaller the further in the future the actual gift is to be handed out. Moreover, if an initial response to a promised future gift occurs, the reciprocal action might then be smaller when the gift is actually received. There may also be an annoyance effect, if clinicians are very impatient, resulting in a smaller response in the first period and then no response in the second. While we do not see evidence of annoyance, our data do suggest the possibility of a mild anticipation effect occurs in the time following the promise of the delayed gift: the coefficient for Period 1 for delayed gift in Table 4 is positive.

Task shifting

Table 5 shows the results of the experiment differentiating between those tasks that were primed in the encouragement and those which are important but were not mentioned in the script. Task shifting would manifest as a decrease in the use of non-primed tasks for clinicians who demonstrated an increase in primed tasks. There is no evidence of task shifting. All clinicians appear to increase the use of primed tasks compared to un-primed tasks (both in Period 1 and Period 2), but there is no decrease in un-primed tasks. Note that both *delayed gift* and *prize* show positive responses to non-primed items in this specification. Furthermore all of the differences among the treatments are observed for un-primed tasks. In other words, clinicians do more of the items they are asked to do more of, but they do more of all items when they are given a gift. Professionals therefore appear to prefer the treatment aspects that preserve their autonomy, consistent with cost of control (Falk & Kosfeld, 2006).

Note that, by looking separately at the impact of performance on primed and un-primed items, announcing a gift appears to have a double impact on un-primed items: clinicians respond by increasing effort a little immediately and then increase effort again when the gift

is actually received. Again, this suggests that clinicians respond both to being promised a gift and receiving it, and offers the possibility that the sum of these two effects might be greater when they are separated in time, than when they occur simultaneously. Note, however, that the differences in the sums of the two sets of coefficients (across Period 1 and 2) are not significantly different and that we observe small (and insignificant) decreases for unprimed tasks.

Impacts controlling for differential timing of treatment implementation

Although the randomization of treatment categories created ex-ante comparable groups, there are differences across the treatment categories in the timing of the visits as seen in Table 6. These are driven by the inherent difficulty of making follow-up visits in an outpatient setting characterized by frequent changes in clinician schedules. Frequently we had to visit multiple times to find the clinician present. For the treatments that did not require feedback visits, keeping to the schedule was easier. Table 7 repeats the analysis above but with the addition of control variables for the duration of the experiment. In this regression, the coefficients for the overall treatment effect on the control group (Period 1 and Period 2 effects) change because these treatments are collinear with the variable measuring the duration of the experiment. However, the differences within the four treatment categories remain similar to the previous analysis and remain statistically significant. Thus, the results we obtain are driven by the treatment design, not the differences in the timing across treatments.

Baseline effects and consultation timing effects

Table 8 looks at the impact of the study on the behavior of clinicians during the baseline (before they were randomized into treatment categories) and the dynamics of protocol adherence within each visit. The coefficients in this regression come from the same regression that examined the impacts of treatment overall, Table 4. Holding the pre-peer scrutiny adherence to protocol as the base, the coefficients show that clinicians increase their effort when there is a peer observer in the room, but return to the original level of quality immediately after the peer leaves. This result is important in establishing that the peer scrutiny visit had no long term effects on protocol adherence.

To verify the study methodology, we also examine the possibility that clinicians might hear (from patients or nurses) that the research team had arrived and then start changing their behaviour while data was being collected. If this were so, patients might report higher quality care than what would have happened in the absence of the data collection team. We can look for evidence of exactly this response by looking at how the quality of care varies with the order of patients on the same day. Since the first few patients we interviewed would have consulted with the clinician before the team arrived, it is not possible to alter the quality for these patients, but subsequent patients might see better quality if our presence was detected. Thus, strategic behavior would manifest as increases in care relative to the first few observations from each visit. To test for this we included a variable for the order of patients by day for each clinician after the encouragement visit. Over the entire period of the study, quality increases slightly (although insignificantly) with the order of the visit. After randomization the marginal effect of order on the day of the visit is negative, meaning that

the first patient seen received slightly better care. This is not compatible with any reaction to our team being present. Thus, this coefficient shows that clinicians are unlikely to have known the team was present during the data collection effort, even if they did eventually learn that the team had been present.

Alternative empirical specifications

In the Appendix and Table 9 we report the results of the additional empirical specifications. These tables show that the basic findings shown in Table 4 are the same across all specifications. Across the main empirical model and the robustness checks, we find a consistent effect (about 4 percentage points) across the two periods. Clinicians increase their effort upon *receiving* an unconditional gift, whether if it is immediately given or has been announced before. The announcement of the gift and prize also triggers some (small) performance increases. The receipt of the prize, not surprisingly has a small and insignificant impact on effort, as shown by the Period 2 coefficient for *prize*. In Period 2, the clinician has already earned his prize and can return to previous levels of effort. Examining all specifications, we find that although we cannot dismiss the possibility that announcing a gift or a prize has important impacts, we can reject the hypothesis that giving a gift has no impact.

3 Discussion and Conclusion

In this study, we observed health workers in their natural work setting and investigated their effort response (as measured by protocol adherence) to non-monetary incentives. In addition to encouraging them to work harder, we tested the effect of small non-monetary gifts that are given unconditionally or are conditional on reaching a specific performance level. We also studied the impact of the timing of the gift, i.e., whether it is given immediately or promised and then given at a later date. Our study had three treatment groups: one received an unconditional gift immediately, one earned the gift (prize) through effort over time, and one received an unconditional gift with the receipt timed to match that of the conditional gift. The control group received no gift but was subject to the same implied scrutiny as the treatment groups. In essence, the study examines the value of unconditional non-monetary gifts as part of a broader, independently operated performance improving intervention.

Our study shows that gifts trigger reciprocal effort: health workers reciprocated after receiving an unconditional gift by immediately increasing their effort. Importantly, taking the same gift and adding the element of conditionality (the *prize* treatment) does not improve performance and in fact is not statistically better than doing nothing. This does not mean that no conditional treatment would work but it does mean that our gift worked *because of*, not despite its unconditionality. Strictly speaking we tested the impact of a gift when clinicians knew they would later receive feedback. However, since the impact of the gift itself is strong and almost exactly the same in both the *gift* and *delayed gift* treatment after it is received, and the *delayed gift* treatment has no feedback it seems likely a gift without any feedback is likely to produce the same effect as a gift with feedback: the important element is the gift, not the feedback.

We also found that even participants in the control group increased their performance over the course of the study, thereby suggesting that the continued presence of research teams alone may have generated a significant and positive response from subjects. This persistent effect outperforms the standard Hawthorne effect both in size and duration. While most health care research focuses on encouragement, information, and performance feedback, our design does not seek to identify the impact of these elements. Rather we find that the response to receiving gifts (short-term effect) and the response to the sense of being studied (long-term effect) are important channels through which performance increases of clinicians can be achieved.

The gains seen in this study are economically significant. The 4.5 percentage point response to receiving an immediate gift is about one quarter of a standard deviation of the observed differences in quality among clinicians in the sample, representing about a third of the difference between average protocol adherence at effective and ineffective organizations in a similar setting (Leonard et al., 2007). In a systematic review of the impact of audit and feedback, Jamtvedt et al. (2003) find an average reduction in non-compliant behaviour of 7%, whereas our gains translate to approximately a 20% reduction. In this setting, gifts increase effort without task shifting and despite no investment in training or medicines.

Acknowledgments

This work was funded by a Maryland Agricultural Extension Station seed grant, a contract from the African Region HRH of the World Bank in part funded by the Government of Norway, and the Eunice Kennedy Shriver National Center for Child Health and Human Development grant R24-HD041041, Maryland Population Research Center. We are grateful for the support of the Center for Educational Health, Arusha (CEDHA), specifically Dr. Melkiory Masatu and Dr. Beatus Leon. We thank Ottar Maestad for feedback on the design of the experiment and CMI (Bergen), Dr. Emmanuel Maliti (REPOA) and seminar participants from several universities for feedback on early versions of this paper.

References

- Akerlof GA. Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*. 1982; 97(4):543–69.
- Ashraf, N., Bandiera, O., Jack, BK. No Margin, No Mission? A Field Experiment on Incentives for Pro-Social Tasks. Technical report, CEPR; 2012.
- Basinga P, Gertler PJ, Agnes S, Sturdy J. Effect on maternal and child health services in rwanda of payment to primary health-care providers for performance: an impact evaluation. *Lancet*. 2011; 377(9775):1421–1428. [PubMed: 21515164]
- Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In: Lord, FM., Novick, MR., editors. *Statistical Theories of Mental Test Score*. London: Addison-Wesley; 1967.
- Bock RD, Lieberman M. Fitting a response curve model for dichotomously scored items. *Psychometrika*. 1970; 35(2):179–198.
- Bradler C, Dur R, Neckermann S, Non A. Employee Recognition and Performance-A Field Experiment. 2013
- Cameron AC, Gelbach JB, Miller DL. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*. 2008; 3(90):414–427.
- Currie J, Lin W, Meng J. Social networks and externalities from gift exchange: Evidence from a field experiment. *Journal of Public Economics*. 2013; 107:1930.
- Das J, Hammer JS. Which doctor?: Combining vignettes and item-response to measure doctor quality. *Journal of Development Economics*. 2005; 78:348–383.

- Das J, Hammer JS. Money for nothing, the dire straits of medical practice in delhi, india. *Journal of Development Economics*. 2007; 83(1):1–36.
- Das J, Hammer JS, Leonard KL. The quality of medical advice in low-income countries. *Journal of Economic Perspectives*. 2008; 22(2):93–114.
- Eichler, R., Levine, R., editors. *Performance Incentives for Global Health: Potential and Pitfalls*. Baltimore, MD: Center for Global Development, Brooking Institution Press; 2009.
- Falk A, Kosfeld M. The Hidden Costs of Control. *American Economic Review*. 2006; 96(5):1611–1630.
- Freidson, E. *Profession of Medicine: A Study of the Sociology of Applied Knowledge*. New York: Harper and Row; 1970.
- Gneezy U, List J. Putting behavioral economics to work: testing gift exchange in labor markets using field experiments. *Econometrica*. 2006; 74(5):1365–1384.
- Gneezy U, Rustichini A. Pay enough or don't pay at all. *Quarterly Journal of Economics*. 2000; 115(3):791–810.
- Goldstein H, Browne W, Rasbash J. Tutorial in biostatistics: Multilevel modelling of medical data. *Statistics in Medicine*. 2002; 21:3291–3315. [PubMed: 12375305]
- Gosnell, GK., List, JA., Metcalfe, R. A New Approach to an Age-Old Problem: Solving Externalities by Incenting Workers Directly. Working Paper 22316, National Bureau of Economic Research; 2016.
- Heyman J, Ariely D. Effort for payment a tale of two markets. *Psychological Science*. 2004; 15(11): 787–793. [PubMed: 15482452]
- Jamtvedt G, Young J, Kristoffersen D, Thomson O'Brien M, Oxman A. Audit and feedback: effects on professional practice and health care outcomes (review). *The Cochrane Database of Systematic Reviews*. 2003; (3)
- Jamtvedt G, Young J, Kristoffersen D, Thomson O'Brien M, Oxman A. Audit and feedback: effects on professional practice and health care outcomes (review). *The Cochrane Database of Systematic Reviews*. 2006; (2)
- Jamtvedt G, Young J, Kristoffersen D, Thomson O'Brien M, Oxman A. Audit and feedback: effects on professional practice and health care outcomes (review). *The Cochrane Database of Systematic Reviews*. 2013; (6)
- Kosfeld M, Neckermann S. Getting more work for nothing? symbolic awards and worker performance. *American Economic Journal: Microeconomics*. 2011; 3(3):86–99.
- Kube S, Marechal MA, Puppe C. The currency of reciprocity: Gift exchange in the workplace. *American Economic Review*. 2012; 102(4):1644–62.
- Leonard KL, Masatu MC. Outpatient process quality evaluation and the hawthorne effect. *Social Science and Medicine*. 2006; 63(9):2330–2340. [PubMed: 16887245]
- Leonard KL, Masatu MC, Vialou A. Getting doctors to do their best: the roles of ability and motivation in health care. *Journal of Human Resources*. 2007; 42(3):682–700.
- Leyland, A., Goldstein, H., editors. *Multi-Level Modeling of Health Statistics*. Chichester: Wiley; 2001.
- Lindelow M, Serneels P. The performance of health workers in ethiopia: Results from qualitative research. *Social Science & Medicine*. 2006; 62(9):2225–2235. [PubMed: 16309805]
- Maestad, O., Torsvik, G. Improving the Quality of Health Care when Health Workers are in Short Supply. mimeo, Chr. Michelsen Institute; 2008.
- Mathauer I, Imhoff I. Health worker motivation in africa: the role of non-financial incentives and human resource management tools. *Human Resources for Health*. 2006; 4(24)
- Meessen B, Musango L, Kashala J-PI, Lemlin J. Reviewing institutions of rural health centres: the performance initiative in butare, rwanda. *Tropical Medicine and International Health*. 2006; 11(8): 1303–1317. [PubMed: 16903893]
- Rigdon ML. Efficiency wages in an experimental labor market. *Proceedings of the National Academy of Sciences of the United States of America*. 2002; 99(20):13348–13351. [PubMed: 16578880]

- Rowe AK, de Savigny D, Lanata CF, Victora CG. How can we achieve and maintain high-quality performance of health workers in low-resource settings? *Lancet*. 2005; 366:1026–1035. [PubMed: 16168785]
- Serra D, Serneels P, Barr A. Intrinsic motivations and the nonprofit health sector. *Personality and Individual Differences*. 2011; 51(3):309–314.
- Turner R, Omar RZ, Yang M, Goldstein H, Thompson S. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*. 2000; 19:3417–3432. [PubMed: 11122505]

A Appendix

A.1 Other specifications

To verify the robustness of our estimates we examine results with four other approaches. The first alternative is to model the problem as if clinicians were repeatedly taking a test in which their effort increases the probability of performing the correct task. The empirical specification that reflects this perspective is from task Response Theory (IRT). IRT is a common technique used in the education literature (Birnbaum, 1967; Bock & Lieberman, 1970) and has been used to evaluate protocol adherence in health care (Das & Hammer, 2005; Leonard et al., 2007). It is designed to analyze a series of right/wrong observations where specific tasks may vary in both overall difficulty (β_k) and the return to ability, or “discrimination” (α_k).

$$\frac{\text{prob}(x_{ijk} = 1)}{1 - \text{prob}(x_{ijk} = 1)} = \exp(\alpha_k \cdot (\theta_j + T_i \vec{t}) + \beta_k + Z_i \vec{z}) + \varepsilon_{ijk} \quad (2)$$

The difficulty score (β_k) is similar to an task-fixed effect, and the discrimination score (α_k) measures the importance of clinician effort (θ_j) in providing the specific task. A high discrimination score means that effort plays an important role in getting the right answer and discrimination coefficient near zero means that all clinicians are equally likely to get a right answer. (A negative score means the worst clinicians are more likely to get it right.) We include a vector of patient characteristics (Z_i): the gender of the patient; whether the patient is an infant, child, or adult; gender and age of the caregiver (if applicable); number of symptoms the patient reports; and the patient’s place in line relative to all the patients seen over the course of the day (i.e. whether the patient is seen by the clinician first, second, third, etc.). Treatment group categories (T_i) are modeled as increasing clinician effort directly ($\theta_j + T_i \vec{t}$).

In essence, IRT develops a view of quality that is defined by the observations within the data set. Some tasks are done by everyone; some are done by very few. Of the tasks that very few clinicians do, tasks that are more likely to be done by the best clinicians are given a high discrimination score, tasks that are more likely to be done by the worst clinicians are given a negative discrimination score, and tasks that appear to have no association with quality, are given a low (close to zero) discrimination score. Thus, by using IRT we can weight each task according to its importance in measuring quality.

Rather than standard errors, we report (in brackets) bootstrapped p-values, the proportion of samples (from a sample of 500 draws) for which the coefficient is the opposite sign of the estimated coefficient. We draw samples at the clinician level; clinicians are sampled with replacement from the full sample and all observations associated with a sampled clinician are used in the bootstrap (Cameron et al., 2008).

As a further robustness check, we provide results from the standard logit model with fixed effects for tasks (β_k) and clinicians (Γ_j), and clustered standard errors to take into account the correlation within patients and across patients within clinicians. We show two sets of standard errors for this regression: clustering at the clinician level (treatment assignment) or at the facility level (the level at which patients select care).

$$\text{prob}(x_{ijk} = 1) = \Phi(T_i \vec{t} + \beta_k + \Gamma_j) + e_{ijk} \quad (3)$$

A final specification looks at the quality of care provided at the patient level, not the task level. This does not allow us to look for task shifting within patients (doing more or less of some tasks compared to others after treatment) but it deals with the potential correlation within patients by having only one observation per patient. The average number of tasks performed per patient does not have an intuitive relationship to effort since the difficulty of tasks is very different. To generate a quality score for each patient we follow a similar procedure to the IRT analysis above, except at the patient, rather than clinician level.¹⁵

$$\frac{\text{prob}(x_{ijk} = 1)}{1 - \text{prob}(x_{ijk} = 1)} = \exp(\alpha_k \theta_i + \beta_k) + \varepsilon_{ijk} \quad (4)$$

Then θ_i is the patient level score (there is only one θ per patient, but there are multiple θ per doctor.) We estimate the following regression:

$$\theta_i = \alpha + T_i \vec{t} + \beta_k + \Gamma_j + e_{ijk} \quad (5)$$

Table 9

Experimental Results, Alternative Specifications

	whether clinician provided specific task as reported by patient in exit interview (0/1)				
	(1)	(2)	(3)	(4)	(5)
Period 1					
overall (control is omitted cat.)	0.012 (0.014)	0.025 [0.182]	0.023 (0.025)	0.023 (0.025)	0.018 (0.019)

¹⁵Alternative ways to generate patient level scores, such as principle component analysis are not feasible with this data set because each patient has a limited number of potential tasks and therefore significant portions of the full matrix are missing.

whether clinician provided specific task as reported by patient in exit interview (0/1)					
	(1)	(2)	(3)	(4)	(5)
<i>gift</i>	0.046 *** (0.015)	0.046 * [0.096]	0.046 * (0.028)	0.046 * (0.025)	0.039 + (0.025)
<i>delayed gift</i>	0.021 (0.015)	0.025 [0.158]	0.017 (0.027)	0.017 (0.022)	0.017 (0.021)
<i>prize</i>	0.024 (0.015)	0.031 [0.190]	0.017 (0.029)	0.017 (0.023)	0.022 (0.024)
Period 2					
Overall	0.061 *** (0.014)	0.094 *** [0.000]	0.084 *** (0.026)	0.084 *** (0.023)	0.065 *** (0.020)
<i>gift</i> (follow up)	0.017 (0.017)	−0.003 [0.494]	0.007 (0.040)	0.007 (0.039)	0.016 (0.034)
<i>delayed gift</i>	0.048 *** (0.016)	0.054 ** [0.040]	0.038 (0.029)	0.038 * (0.021)	0.039 * (0.023)
<i>prize</i>	0.023 (0.017)	0.006 [0.394]	−0.006 (0.031)	−0.006 (0.035)	0.017 (0.024)
pre peer scrutiny	omitted	omitted	omitted	omitted	omitted
peer scrutiny	0.024 ** (0.010)	0.038 *** [0.004]	0.007 (0.013)	0.007 (0.012)	0.018 (0.012)
post-peer scrutiny	0.006 (0.012)	0.016 [0.228]	0.005 (0.018)	0.005 (0.017)	0.002 (0.016)
patient order	0.001 (0.001)	−0.002 [0.204]	0.001 (0.002)	0.001 (0.001)	0.001 (0.002)
patient order after enc.	−0.003 (0.002)	−0.004 [0.324]	−0.003 (0.002)	−0.003 ** (0.001)	−0.003 * (0.002)
Observations	96964	96964	96964	96964	4416

Marginal Effects reported. Standard errors in parentheses; P-values in brackets: significance at 1% (***), 5% (**), 10% (*), and 15% (+)

(1): Multi-level logit model with random effects at the patient and clinician level

(2): Non-linear logistic model (IRT). Non-parametric P-values are calculated from 500 block bootstrapped samples (sampling with replacement at the clinician level), where the p-value is the proportion of samples with a coefficient of the opposite sign of the estimated coefficient. The coefficients for the difficulty and discrimination parameters for column 2 are reported in Table 11 in Appendix A.3.

(3): Logit model with clinician fixed effects (as dummy variables) and standard errors clustered at the clinician level.

(4): Logit model with clinician fixed effects (as dummy variables) and standard errors clustered at the facility level.

(5): Fixed effect regression of patient level quality score with clinician fixed effects, clustered at the clinician level.

A.2 Validating the RCR Instrument

For part of the baseline evaluation of each clinician, a clinician sat in the consultation room and recorded the activities of the clinician. These same patients were interviewed later with the RCR instrument. The dual methods of recording information allow us to validate whether patient reports track the report of a professional present in the consultation. Table 10 shows the results of three regressions of the RCR observation on the DCO observation, including 1) item fixed effects, 2) clinician fixed effects and 3) item and clinician fixed effects. In each case the correspondence is very strong and, taking into account clinician and item fixed effects results in a relationship which is statistically indistinguishable from 1. Note that in all regressions in this paper, item and clinician effects are controlled for.

Table 10

Validating the relationship between patient and clinician reports of the same event

Direct Consultation Observation	Retrospective Consultation Review
	task performed (0/1)

Direct Consultation Observation	Retrospective Consultation Review		
Task performed (0/1)	1.2*** (0.068)	1.622*** (0.065)	0.996*** (0.080)
task fixed effects	included		included
clinician fixed effects		included	included
Observations	8593	8593	8577

A.3 Task difficulty and discrimination scores

The IRT analysis of dichotomous data measures the ability of each clinician and the difficulty and discrimination levels of each task. Difficulty measures the difficulty of the task for all participants and discrimination measures the degree to which better clinicians are more likely to get a task correct. For example, the question “did the doctor clearly explain the directions for the drugs” has a particularly high discrimination score and the question “did the doctor ask if the child had convulsions?” has a particularly low discrimination score. This suggests that clinicians with a high ability are much more likely to ask the first question, but not much more likely to ask the second. Although both would appear important, doctors explain that (for the question on convulsions) a mother would never fail to mention convulsions and therefore they see this question as irrelevant. Some doctors do ask the question, but they are not necessarily the better doctors. On the other hand, better doctors do tell their patients about the medications that have been prescribed. This question helps to discriminate between the better and worse doctors. The difficulty scores vary from approximately zero to approximately 5. A negative difficulty score simply suggests that most doctors will ask this question. Since the difficulty score has no natural scale, the hypothesis that it is equal to zero has no economic meaning; we report p-value cutoffs to indicate the strength of the higher scores for indicating tasks that are truly more difficult. Note that almost all doctors welcome and greet their patients but many fewer pinch the skin fold for young children with diarrhea. The ability score serves as a fixed effect for each clinician in the sample. The impact of the experiment is seen in the overall increase in the probability of correctly performing an activity, not in an increase in ability.

Table 11

tasks in the RCR list, changes with scrutiny and encouragement and difficulty and discrimination scores

type task	task	Prac. Qual Scores	
		Discrimination	Difficulty
Greeting and Receiving	Did the doctor welcome and greet you?	5.501 (1.128)***	0.034 (0.723)
	Did the doctor listen to your description of the illness?	3.823 (1.225)***	-1.379 (0.808)*
	Did you have a chair to sit in?	3.244 (1.256)***	-1.797 (0.838)**
General, history taking	Did the doctor ask you how long you had been suffering	8.301 (0.825)***	3.139 (0.515)***

type task	task	Prac. Qual Scores	
		Discrimination	Difficulty
	Did the doctor ask you if there were other symptoms different from the main complaint?	8.820 (0.726)***	4.576 (0.456)***
	Did the doctor ask if you already received treatment elsewhere or took medicine?	9.305 (0.722)***	5.351 (0.458)***
Education			
	Did he give you a name for your illness?	7.439 (0.584)***	4.081 (0.384)***
	Did he explain your illness?	9.742 (0.714)***	5.408 (0.461)***
	Did he explain the treatment?	10.281 (0.879)***	4.351 (0.539)***
	Did he give you advice to improve your health?	10.774 (0.812)***	6.266 (0.509)***
	Did he explain if you need to return?	7.162 (0.574)***	3.983 (0.375)***
	Did the doctor explain what the drugs are for?	13.543 (1.133)***	5.976 (0.681)***
	Did the doctor clearly explain instructions for the drugs?	14.845 (1.216)***	7.263 (0.729)***
	If so, did the doctor explain why you would have this test?	14.009 (1.531)***	6.165 (0.903)***
	Did the doctor order a lab test?	2.446 (0.327)***	1.650 (0.241)***
	Did he explain why you were referred?	9.835 (4.194)**	4.631 (2.488)*
	Did he tell you what to do?	12.612 (5.357)**	6.899 (3.135)**

Table 12

RCR questions (II)

type task	task	Prac. Qual Scores	
		Discrimination	Difficulty
Fever, history taking			
	Did the doctor ask you how long you had a fever?	7.333 (0.861)***	4.082 (0.559)***
	Did the doctor ask you if you had chills or sweats?	5.976 (0.718)***	4.101 (0.486)***
	Did the doctor ask you if you had a cough or difficulty breathing?	5.075 (0.658)***	3.585 (0.451)***
	Did the doctor ask you if you had diarrhea or vomiting?	7.112 (0.807)***	4.654 (0.531)***
	Did the doctor ask if you had a runny nose?	7.665 (0.861)***	4.859 (0.565)***
Fever, history taking, under 5			
	Did the doctor ask the child had convulsions?	2.447 (0.979)***	3.488 (0.689)***
	Did the doctor ask about difficulty drinking or breastfeeding?	4.834 (0.979)***	3.998 (0.662)***
	Listen to the child's breathing, or use a stethoscope?	7.523 (1.152)***	4.865 (0.758)***
	Did the doctor check the child's ear?	5.006 (0.976)***	4.394 (0.676)***
	Did the doctor ask questions about the child's vaccinations?	6.080 (1.061)***	5.256 (0.728)***
Cough, history taking			
	Did the doctor ask the duration of the cough?	7.436 (1.173)***	3.476 (0.742)***
	Did the doctor ask if there was sputum?	6.213 (0.832)***	4.046 (0.557)***
	Did the doctor ask if you had blood in your cough?	5.821 (0.754)***	4.795 (0.526)***
	Did the doctor ask if you had difficulty breathing?	7.303 (0.977)***	4.292 (0.635)***
	Did the doctor as if you also have a fever?	6.807 (0.997)***	3.690 (0.647)***

Table 13

RCR questions (III)

type task	task	Prac. Qual Scores	
		Discrimination	Difficulty
Cough, history taking, under 5			
	Did the doctor ask about the history of vaccinations?	5.944 (1.218)***	5.219 (0.848)***
	Did the doctor ask about difficulty drinking or breastfeeding?	5.333 (1.177)***	4.401 (0.805)***
	Did the doctor ask if the child had convulsions?	2.416 (1.296)*	3.800 (0.922)***
	Did the doctor check the child's ear?	5.707 (1.197)***	4.848 (0.829)***
	Did the doctor ask if the child had diarrhea or vomiting?	5.062 (1.141)***	3.668 (0.758)***
Diarrhea, history taking			
	Did the doctor ask how long you have had diarrhea?	2.121 (1.413)	0.382 (0.943)
	Did the doctor ask how often you have a movement?	4.671 (1.412)***	2.651 (0.923)***
	Did the doctor ask about the way the stool looks?	4.968 (1.474)***	2.725 (0.956)***
	Did the doctor ask if there was blood in the stool?	5.766 (1.382)***	3.864 (0.919)***
	Did the doctor ask if you are vomiting?	6.156 (1.590)***	3.489 (1.020)***
	Did the doctor ask if you also have a fever?	7.461 (1.836)***	4.037 (1.154)***
Diarrhea, history taking, under 5			
	Did the doctor ask about difficulty drinking or breastfeeding?	3.001 (2.121)	2.608 (1.391)*
	Did the doctor ask if the child had convulsions?	−2.877 (3.137)	0.287 (1.985)
	Did the doctor check the child's ear?	3.628 (2.028)*	3.626 (1.366)***
	Did the doctor ask if the child had diarrhea or vomiting?	5.030 (2.679)*	2.288 (1.638)
	Did the doctor ask questions about the child's vaccinations?	1.982 (2.201)	2.656 (1.484)*

Table 14

RCR questions (IV)

type task	task	Prac. Qual Scores	
		Discrimination	Difficulty
Fever, diagnostic			
	Did the doctor take your temperature?	9.730 (0.945)***	6.184 (0.619)***
	Did the doctor check for neck stiffness?	4.887 (0.683)***	4.764 (0.487)***
	Did he ask if you felt weakness from lack of blood?	4.218 (0.652)***	4.062 (0.464)***
	Did he look in your ears or throat?	4.918 (0.697)***	4.778 (0.496)***
	Did he check your stomach?	3.253 (0.638)***	3.719 (0.460)***
	Did he ask for a blood slide?	5.888 (0.766)***	3.425 (0.507)***
Fever, diagnostic, under 5			
	Did the doctor check if the child was sleepy, try to wake up the child?	6.206 (1.018)***	5.568 (0.706)***
	Did the doctor pinch the skin fold of the child?	6.429 (1.032)***	5.640 (0.723)***
	Did the doctor check both of the child's feet?	6.824 (1.165)***	6.556 (0.831)***
	Did the doctor check the child's weight against a chart?	3.411 (0.956)***	3.293 (0.663)***
Cough, diagnostic			

type task	task	Prac. Qual Scores	
		Discrimination	Difficulty
	Did he look at your throat?	4.972 (0.728)***	4.311 (0.524)***
	Did he listen to your chest?	6.663 (0.843)***	4.113 (0.573)***
	Did he take your temperature?	6.976 (0.866)***	4.907 (0.596)***

Table 15

RCR questions (IV)

type task	task	Prac. Qual Scores	
		Discrimination	Difficulty
Cough, diagnostic, under 5			
	Did the doctor check if the child was sleepy, try to wake up the child?	5.185 (1.166)***	4.683 (0.804)***
	Did the doctor pinch the skin fold of the child?	6.696 (1.305)***	5.714 (0.903)***
	Did the doctor check the child's eyes, tongue, and palms?	6.992 (1.306)***	5.624 (0.895)***
	Did the doctor check both of the child's feet?	7.175 (1.474)***	6.932 (1.066)***
	Did the doctor check the child's weight against a chart?	4.562 (1.145)***	4.382 (0.799)***
	Did he pinch the skin on the stomach?	5.061 (1.349)***	4.775 (0.955)***
Diarrhea, diagnostic			
	Did he take your temperature?	6.923 (1.306)***	5.109 (0.912)***
	If the child is under two years, Did he look at the child's head?	0.369 (4.247)	3.596 (2.911)
	Did the doctor offer the child a drink of water or observe breastfeeding?	1.506 (3.471)	3.398 (2.347)
Diarrhea, diagnostic, under 5			
	Did the doctor check the child's eyes, tongue, and palms?	5.238 (2.238)**	4.504 (1.520)***
	Did the doctor check both of the child's feet?	2.140 (2.483)	3.533 (1.700)**
	Did the doctor check the child's weight against a chart?	0.190 (2.156)	1.242 (1.427)
General, diagnostic			
	Did the doctor examine you?	7.935 (0.629)***	4.785 (0.414)***

Figure 1. Timing of Intervention

Shaded boxes indicate periods of data collection using patient exit interviews. Arrows indicate one-on-one visits between research team and clinician participants. The randomization took place after the baseline data collection and before the encouragement visit. No data were collected on days when encouragement and follow-up visits occurred.

Table 1

Description of Treatments

label	Description	Timing of gift / prize		
		immediate	delayed	feedback
<i>gift</i>	gift with feedback	✓		✓
<i>prize</i>	prize with feedback		✓	✓
<i>delayed gift</i>	delayed gift without feedback		✓	

Note: The control group received no gift and no feedback.

Table 2

Clinician Summary Statistics and Balance Tests

Clinician Average Characteristics	Treatment group				min comp p-value
	Control	Gift	Delayed Gift	Prize	
Clinician Age	42.542 (2.319)	43.435 (1.876)	42.542 (1.749)	41.000 (1.963)	0.375
Clinician Gender	0.348 (0.102)	0.174 (0.081)	0.250 (0.090)	0.381 (0.109)	0.129
Yrs Medical Education	3.500 (0.492)	3.750 (0.487)	3.619 (0.411)	3.933 (0.300)	0.465
Yrs working as a healthworker	19.474 (2.614)	16.474 (2.070)	15.739 (1.915)	12.647 (1.998)	0.049
Yrs working with current credential	12.715 (3.111)	10.895 (2.108)	10.348 (1.640)	10.208 (1.975)	0.484
Public Sector	0.440 (0.101)	0.391 (0.104)	0.360 (0.098)	0.190 (0.088)	0.075
Private Sector	0.360 (0.098)	0.435 (0.106)	0.440 (0.101)	0.524 (0.112)	0.274
NGO sector	0.200 (0.082)	0.174 (0.081)	0.200 (0.082)	0.286 (0.101)	0.389
popular facility	0.360 (0.098)	0.435 (0.106)	0.440 (0.101)	0.333 (0.105)	0.471
number of clinicians per facility	4.080 (0.643)	4.565 (0.702)	4.320 (0.652)	4.000 (0.743)	0.583
Clinician Quality (controlling for case mix)	0.634 (0.021)	0.625 (0.025)	0.637 (0.021)	0.632 (0.024)	0.704
Baseline Protocol Adherence (average per clinician)	0.754 (0.028)	0.747 (0.029)	0.776 (0.025)	0.763 (0.033)	0.454
N	25	23	25	21	

Note: min comp p-value is the minimum p-value of all 6 pairwise comparisons among the treatments. Average adherence is the average portion of required items completed by the clinician, per patient. Clinician quality is an estimated parameter and captures the portion of required items completed while taking patient characteristics (case mix) into account.

Table 3

Patient Summary Statistics and Balance Tests

Clinician Average Characteristics	Treatment group			min comp p-value
	Control	Gift	Delayed Gift	Prize
Baseline				
Average Protocol Adherence	0.742 (0.029)	0.714 (0.039)	0.757 (0.028)	0.755 (0.033) 0.369
infant (0–5)	0.401 (0.049)	0.328 (0.041)	0.306 (0.039)	0.319 (0.048) 0.131
child (5–15)	0.104 (0.014)	0.164 (0.019)	0.140 (0.019)	0.142 (0.021) 0.013
adult (15–50)	0.440 (0.041)	0.444 (0.036)	0.442 (0.041)	0.480 (0.040) 0.481
female	0.545 (0.025)	0.532 (0.031)	0.561 (0.025)	0.556 (0.035) 0.464
female caregivers	0.750 (0.034)	0.726 (0.022)	0.684 (0.030)	0.720 (0.038) 0.156
# of patients	366	410	370	330
Task average characteristics				
Probability of performing a task in Baseline, Period 1 and Period 2				
Baseline	0.742 (0.029)	0.715 (0.043)	0.754 (0.037)	0.752 (0.040) 0.361
Change in Probability, Period 1	−0.004 (0.022)	0.048 (0.020)	0.026 (0.017)	0.025 (0.015) 0.079
Change in Probability, Period 2	0.061 (0.022)	0.082 (0.035)	0.109 (0.021)	0.078 (0.029) 0.097
Total # of patients across three periods	1176	1155	1167	940

Note: min comp p-value is the minimum p-value of all 6 pairwise comparisons among the treatments. Standard errors are clustered at the clinician level.

Table 4

Experiment Results in Period 1 and Period 2

Dep Var: Whether clinician performed specific task (0/1)		
	Period 1	Period 2
Improvement in performance due to treatment (compared to baseline)		
Trend for all treatments	0.012 (0.014)	0.061 *** (0.014)
+ marginal effects by treatment		
Control	omitted	omitted
Gift	0.046 *** (0.015)	0.017 (0.017)
Delayed Gift	0.021 (0.015)	0.048 *** (0.016)
Prize	0.024 (0.015)	0.023 (0.023)
Baseline effects	included	
Clinician and patient effects	included	
Consultation timing effects	included	

Note: The data include 96,964 clinician-patient-task level observations. Results derived from Multi-level logit model with random effects at the patient and clinician level. Marginal Effects (percentage point increase compared to the omitted baseline) reported.

Impact for *gift*, *delayed gift* and *prize* are additional effects over the trend for all groups and impact for the control group is equal to the trend for all treatments. Standard errors in parentheses: significance at 1% (***), 5% (**), 10% (*). Period 1 refers to data collected after the encouragement visit. Period 2 refers to data collected after the follow-up visit. Coefficients are from a single regression.

Table 5

Experiment Results in Period 1 and Period 2 by un-primed and primed tasks

Dep Var: Whether clinician performed specific task (0/1)		
	Period 1	Period 2
Improvement in performance due to treatment (compared to baseline)		
<i>Un-primed tasks</i>		
Trend for all treatments	0.003 (0.014)	0.047 *** (0.014)
+ marginal effects by treatment		
Control	omitted	omitted
Gift	0.046 *** (0.015)	0.017 (0.017)
Delayed Gift	0.028 * (0.015)	0.041 ** (0.017)
Prize	0.029 * (0.016)	0.027 (0.018)
<i>Primed tasks</i>		
Trend for all treatments × primed	0.022 ** (0.008)	0.045 *** (0.01)
+ marginal effects by treatment		
Control × primed	omitted	omitted
Gift × primed	0.007 (0.011)	0.018 (0.014)
Delayed Gift × primed	−0.015 (0.011)	−0.005 (0.014)
Prize × primed	−0.007 (0.011)	0.002 (0.015)
Baseline effects	included	
Clinician and patient effects	included	
Consultation timing effects	included	

Note: The data include 96,964 clinician-patient-task level observations. Results derived from Multi-level logit model with random effects at the patient and clinician level. Marginal Effects (percentage point increase compared to the omitted baseline) reported.

Impact for *gift*, *delayed gift* and *prize* are additional effects over the trend for all groups and impact for the control group is equal to the trend for all treatments. Standard errors in parentheses: significance at 1% (***), 5% (**), 10% (*). Period 1 refers to data collected after the encouragement visit. Period 2 refers to data collected after the follow-up visit. Coefficients are from a single regression. Primed items are tasks that were specifically mentioned as important, during the encouragement visit.

Table 6

Duration of Experiment by Treatment

	Treatment Group				Avg.	p-val ^d
	Control	delayed gift	gift	prize		
Days between encouragement and start of Period 1 data collection	7.73 (7.16)	6.03 (5.85)	21.39 (40.92)	8.26 (31.45)	10.98 (26.6)	0.31
Days between encouragement and start of Period 2 data collection	34.26 (19.01)	56.02 (22.59)	49.31 (22.51)	47.07 (26.53)	46.32 (23.97)	0.02
Duration of study	57.59 (19.37)	81.44 (33.97)	73.18 (14.29)	67.57 (17.09)	69.54 (24.88)	0.03

Note: Standard deviation in parentheses^a; P-value for test of joint significant from an OLS regression of the observable variable on the treatment groups.

Table 7

Experiment Results with duration of experiment

Dep Var: Whether clinician performed specific task (0/1)		
	Period 1	Period 2
Improvement in performance due to treatment (compared to baseline)		
Trend for all treatments	−0.026 (0.017)	0.000 (0.020)
+ marginal effects by treatment		
Control	omitted	omitted
Gift	0.042 *** (0.015)	0.009 (0.017)
Delayed Gift	0.024 (0.015)	0.037 ** (0.016)
Prize	0.027 * (0.015)	0.018 (0.017)
Change in performance with duration		
<i>Log of days since encouragement</i>	0.018 *** (0.004)	
Baseline effects	included	
Clinician and patient effects	included	
Consultation timing effects	included	

Note: The data include 96,964 clinician-patient-task level observations. Results derived from Multi-level logit model with random effects at the patient and clinician level. Marginal Effects (percentage point increase compared to the omitted baseline) reported.

Impact for *gift*, *delayed gift* and *prize* are additional effects over the trend for all groups and impact for the control group is equal to the trend for all treatments. Standard errors in parentheses: significance at 1% (***), 5% (**), 10% (*). Period 1 refers to data collected after the encouragement visit. Period 2 refers to data collected after the follow-up visit. Coefficients are from a single regression.

Table 8

Baseline and consultation timing effects

Dep Var: Whether clinician performed specific task (0/1)	
<i>Pre Treatment performance</i>	
Pre Peer Scrutiny	omitted
Peer scrutiny	0.024 ^{**} (0.01)
Post-peer scrutiny	0.006 (0.012)
<i>Consultation Timing</i>	
Order, by day	0.001 (0.001)
Order, by day, after enc.	-0.003 [*] (0.002)

Note: These coefficients are drawn from the regression shown in Table 4. The data include 96,964 clinician-patient-task level observations. Results derived from Multi-level logit model with random effects at the patient and clinician level. Marginal Effects (percentage point increase compared to the omitted baseline) reported.

Standard errors in parentheses: significance at 1% (***), 5% (**), 10% (*).