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

J. Michelle Brock† Andreas Lange‡ Kenneth L. Leonard§

February 8, 2014

Abstract

In this study, we consider how gift-exchange and bonus systems function in a natural field setting by measuring the effort response of participants to non-monetary gifts over time. We show that small unconditional gifts can increase effort in the workplace medium-term period (2 to 4 weeks), not just over short periods as has been shown with students in the laboratory. Our field experiment tests the difference in effort response to unconditional gifts delivered immediately, promised unconditional gifts delivered later, and conditional "gifts" linked to reaching a specific performance target. We find important benefits from promising to give an unconditional gift later: participants respond positively to a promised gift twice by increasing effort when the gift is promised and again when it is received. A promised gift outperforms both the unconditional gift delivered immediately, which leads to a single positive response, and the conditional gift based on performance, which does not trigger any significant behavioural change after the gift is delivered.

JEL Classification: C93, I1, J41, O1
Keywords: gift exchange, reciprocity, health care, field experiment, Tanzania

*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.

†Corresponding author: European Bank for Reconstruction and Development, London, UK brockm@ebrd.com

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

§2200 Symons Hall, University of Maryland College Park, MD 20742 kleonard@arec.umd.edu

1 Introduction

It is a common assumption in economics that increased compensation for effort leads to increased effort.[1] However, tying compensation directly to effort, such as paying a piece rate, can sometimes crowd out intrinsic motivation and thus reduce effort (Fehr and Schmidt, 2004; Frey, 1997). In such settings, bonuses can be effective tools for incentivizing effort as they are less rigorously related to performance, require less monitoring, and therefore may preserve the trust that often underpins workers' willingness to exert effort on behalf of the employer.[2]

In this paper, we provide important new insights into the functioning of gift-exchange and bonus systems in a real workplace setting. This is in contrast to most of the existing literature (e.g., Bradler et al., 2013; Gneezy and List, 2006; Kosfeld and Neckermann, 2011), which looks at students' behaviour in a one-shot employment situation or workers in a new employment opportunity. Specifically, we conducted a field experiment with health care workers at outpatient clinics in the urban and suburban areas of the Arusha region in Tanzania, testing how conditional and unconditional gifts impact their performance over long periods, ranging from 6 weeks to almost 6 months. To differentiate between conditional and unconditional gifts in this real-world setting, we needed to control for the fact that conditional gifts often come later than unconditional gifts. We thus randomly assigned health workers to one of 3 treatment groups and the control. Members of one group received an immediate (unconditional) gift, those in the second group were promised an unconditional gift at a later time, and those in the third were promised a gift at a later time if they met a performance target. Importantly, because we could not observe effort without potentially impacting motivation in this setting, the control group received no gifts but was subjected to the same

[1] The evidence of the positive correlation between compensation and effort is mixed (Deci, 1971; Gneezy and Rustichini, 2000; Mas, 2006) This is particularly the case when we expect that some degree of intrinsic motivation drives effort or when worker effort produces multiple outputs (Deci, 1972; Kreps, 1997).

[2] In fact, there is a significant literature on the role of gifts in economic exchanges, including work in laboratory settings and in field experiments (for example, Akerlof, 1982; Gneezy and List, 2006; Rigdon, 2002). We extensively review this literature in the next section.

level of implied scrutiny from the research team over the course of the study. Aside from allowing us to distinguish the effect between conditional and unconditional gifts, comparing these treatments also facilitates understanding of the potential mechanism underlying the effectiveness of our gift.

The study shows that the treatment group given an immediate gift had the largest effort response in the short run and that, relative to the control group, all types of gifts caused workers to increase their effort. However, there are differences between the initial response (after a visit explaining the treatments) and the later response (after a follow-up visit). In the initial period, agents who received the gift up front increased their performance about twice as much as those who were promised the gift at a later stage and those who had to earn the gift by reaching a specific performance target. But later, after the promised gift was delivered or the earned gift was handed over, those agents who had received their gift up front and those who received the gift in return for reaching performance targets returned to a performance level similar to the control group's, behaving as if they believed they had already repaid or earned the value of the gift they received. In contrast, those who later received the unconditional gift that had been previously promised, increased their effort relative to the control group, even though they had also increased their effort somewhat when they were originally promised the gift. Introducing the temporal element of promising but delaying the gift revealed that the impact of gifts can be increased by dividing them into 2 parts: learning about a gift and receiving it. In other words, promising a gift before delivering it led to a "double dividend", with the group receiving this treatment having the largest overall increase in effort over the 2 time periods. This finding yields important insights into the temporal aspects of reciprocity in the workplace.

In order to measure the quality of care that the health workers participating in the study provided to their patients, we interviewed patients as they left their outpatient consultations, about items required by national protocol to be addressed during consultations concerning their presenting symptoms. The existence of these protocols allows us to assume that in-

creases in effort are good for the patient in the sense that they contribute to the probability of being cured, the probability of eliminating extra incorrect diagnoses (and therefore reducing excess use of medicine), and the probability that the patient will understand the treatment prescribed. Our interventions provided a one-off small addition to the normal pay structure for all of the participants; none of our treatments impacted their caseload, hours of work, competencies, wages, or promotion possibilities.

In all treatment groups and the control, clinicians received encouragement in which we explained that we would like them to increase their effort, told them of 5 items that are commonly neglected in the study region, and offered information about these important items. To control for the impact of scrutiny on participants, we maintained the same level of contact and exposure to our research team in the control group as in the gift treatment groups. Importantly, we found that even in the control, the workers showed performance increases over the course of our study. The impacts of gifts in the treatment groups discussed here are compared to this increase in the control group: all participants increased their effort over the course of the study, but those who received gifts increased their effort more. The fact that we observed increases in effort for the control group demonstrates the methodological importance of creating a legitimate control in a long-term randomized controlled trial study. Relying instead on performance levels taken before the interventions as a baseline would have led to a significant overestimation of the treatment effect. Likewise, comparing the treatment groups to a control group that did not receive similar levels of attention—as is often done in RCTs in developing countries—would also have significantly overstated the impact of the intervention: simply being in the control group can change behaviour in this type of a setting. We were also able to test for crowding out or task shifting by examining performance on tasks that were not offered as examples of particularly important tasks during the encouragement visit. Although effort related to mentioned activities increased more than unmentioned activities, we found no evidence of task shifting or crowding out from gift giving.

Finally, we consider our choice of health workers in a developing-country context to be important. This is a setting in which effort is difficult to observe and almost impossible verify. We know that a significant gap exists between effort provided by health workers and their capacity (Das and Hammer, 2007; Das et al., 2008; Maestad and Torsvik, 2008; Rowe et al., 2005). Despite the low quality of health care in most developing countries, 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 and Serneels, 2006; Mathauer and 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.[3] 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 problems that have otherwise proven difficult to address.

The remainder of this paper is structured as follows: in the following section, we discuss the existing evidence on the link between gifts and effort and how it relates to our current experiment. In section 3 we discuss the data, our experimental design, and our estimation strategy. In section 4 we review the results of the experiment, and in section 5 we conclude.

2 Inducing Effort with Gifts

Gifts are important in all cultures and have played an important economic role in most primitive societies (Posner, 1980). A gift is "any exchange of goods and services with no guarantee of recompense in order to create, nourish, or recreate social bonds between peo-

[3]There 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 and Levine, eds (2009), for an extended discussion of pay for performance; and Basinga et al. (2011); Meessen et al. (2006), for early evidence of success.

ple" (Cailleé, 2001). Despite the lack of *guaranteed* recompense, research has focused on the reciprocal obligations *implied* by a gift (Mauss, 1925). Gifts form part of the "ritual practice through which the current value of a relationship may be communicated and maintained" (Berking, 1999). Camerer (1988) suggests that gifts are useful as signals of intentions within relationships. In most traditional societies, gifts are more accurately described as loans (Osteen, ed, 2002). As Posner (1980, pp. 16) notes "gifts in primitive society are explicitly reciprocal: a man is under a strong moral duty to repay a gift, when he can, with a gift of equivalent value." Development economists often encounter gifts as central to economic relationships such as social insurance networks (Fafchamps, 1992) and landlord-tenant relationships (Sadoulet et al., 1994), and even as a basic tool for avoiding theft (Schechter, 2007). Thus, gifts form part of a long-term incomplete contract between the recipient and giver.[4] It is not always clear, however, whether gifts help generate a relationship characterized by mutual exchange (à la Cailleé) or invoke reciprocal responses within an existing relationship (à la Posner).

Gift exchange has been discussed extensively as a way to understand wages and effort in the workplace (Akerlof, 1982; Gneezy and List, 2006; Rigdon, 2002). These authors posit and find that giving above-market wages to employees can result in higher effort, even though the wages are not contingent on effort, pointing to reciprocity as a motive for worker effort choices. To test these implications, Gneezy and Rustichini (2000) conducted experiments with treatments of either getting paid to do a specific task or not getting paid to do the task. They found the expected result that higher compensation induced higher effort within the group that was paid, but total effort was higher in the group that was not paid.[5] Where tested, the experimental evidence highlights the importance of duration, in particular, that gifts lead to short term gains. Gneezy and List (2006) posit that the

[4]Fafchamps (1992) suggests that the recipient of a gift in a social insurance exchange is obligated to give to another person in need at a future date, not the original giver.

[5]The task in their experiment was collecting donations for a charity. Since we can expect that some degree of intrinsic motivation drives effort in this kind of task, this result is a good example of how extrinsic incentives may crowd out intrinsic incentives.

short-term increase in effort from receiving a gift might be caused by a "hot" evaluation of the relative costs and benefits of the task; in the long run, a "cold" evaluation of the same incentives may lead to a different allocation of effort. However, the impact may also come from temporal effort shifting: working harder because you received the gift makes you tired and results in lower levels of effort later. Their design highlights this impact but does not allow further investigation. Whether due to the difference in "hot" and "cold" evaluations or effort shifting, the contrast between short- and long-term impacts of monetary bonuses suggests the importance of better understanding the temporal aspects of gift exchanges in the workplace.

Further experiments have shown that using non-monetary incentives, rather than cash bonuses, can alter the way subjects perceive incentives and thus the nature of their reciprocity in a gift-exchange context. Heyman and Ariely (2004) test the idea that payment types (monetary vs. non-monetary) signal different types of markets in real effort tasks: money markets and social markets, each with its own norms. Framing the task within one of these markets determines the norm for the relationship between payment and effort. They use a between-subjects design to test the differences between candy and money as flat-rate rewards for effort. In the social market (candy), effort was invariant to the rate, but in the money market, the rate impacted effort. Thus, the type of payment itself is a cue as to how effort is perceived. Kube et al. (2012) examine in more detail the value of non-monetary gifts in real effort tasks in the laboratory. They find that subjects who are given a water bottle gift (wrapped up with a bow) out-perform those who receive a monetary bonus, and that subjects who receive a cash bonus folded in origami style outperform those who received the same cash bonus not in the origami shape. In line with Dur's theory of managerial attention (Dur, 2009), Kube et al. suggest that "its the thought that counts." In these studies, offering non-monetary incentives effectively elicited greater effort than monetary incentives, potentially due simply to how participants perceived this alternate frame of the gift-exchange relationship.

If participants in these experiments are purely self-interested, the frame of the effort-compensation link should not matter. However, strong evidence exists that preferences for fairness and reciprocity can directly affect the relationship between effort and compensation. Results from simple games such as the ultimatum game, the trust game, and various public goods games overwhelmingly demonstrate that most people (including workers) are not purely self-interested profit maximizers (Andreoni and Miller, 2002; Forsythe et al., 1994; Palfrey and Prisbrey, 1996, 1997). Using decision making in laboratory experiments as a measure of social preferences, Barr and Serneels (2009) (investment game) and Carpenter and Seki (2010) (public goods game) both find a positive relationship between worker productivity in the field and social preferences. Gächter and Falk (2000) look at overcoming inefficiencies from incomplete contracting over worker performance and find that social embeddedness in non-repeated games can achieve results similar to long-term contracting.[6] As is predicted in Sliwka (2007), trust performs better than monetary incentives as a contract enforcement device. Thus, when workers have social preferences, multiple options are possible for framing the effort-compensation link, and strictly increasing monetary incentives may not be the most effective way to motivate effort.

There are a few studies that test the effectiveness of non-monetary incentives for motivating performance in the field. Kosfeld and 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 setting, they do find positive and significant effects from symbolic awards. The Bradler et al. (2013) study even shows 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

[6]See also Bohnet and Frey (1999); Eckel and Grossman (1996) on the value of embeddedness.

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. Their field experiment in Zambia compares trainees' sales of condoms under monetary and non-monetary incentives. As in Kosfeld and 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 2 of the treatments, and awarding each participant's gift in private.

3 Methodology

Based on previous research measuring the quality of health care among clinicians in Tanzania, we know that effort varies significantly across clinicians as well as across the types of patient a single clinician might see. But there is no evidence in any research in Tanzania of a secular upward trend in effort by the average clinician: in the absence of our research we could expect that there would be no changes in effort. We therefore designed our sampling strategy on the assumption that we would measure *changes* in effort by clinician, relative to a baseline, after controlling for the illness condition of the patient following the methodology used in Leonard et al. (2007).

3.1 The Sample and Data Collection

We collected data on clinician performance for 103 clinicians by collecting 4,379 post consultation interviews with patients in the semi-urban area of Arusha, Northeast Tanzania. We 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 sampled from among clinicians who were ever present at a series of facilities that we could easily reach in the sample area. This sample is representative of the clinicians who see reasonable numbers of patients in the study region and is therefore a policy-relevant population. The field data collection ran from November 2008 until August 2010.

The sample includes public, private, and non-profit/charitable facilities. We restricted our attention to clinicians because they are the primary health workers who provide the outpatient care in the area. They fill the role of "doctor," though the majority of them do not have full medical degrees.[7] Of the 103 clinicians we initially enrolled, 12 dropped out before they were randomized into one of the 4 treatment groups. An additional 3 dropped out after being assigned to treatments.[8] We were able to visit 10 clinicians who practiced in the rural areas only twice for data collection after the encouragement visit due to the expense of returning for the post-study visits. Thus, the sample changed slightly over the course of the study. Importantly, however, there were no differences in characteristics at the

[7]The 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.

[8]Of the 15 who dropped out, 2 revoked their consent and the remainder went on leave or were reposted and therefore unavailable.

baseline for those who attritted.

We collected data on the quality of care provided by each clinician on at least 5 separate occasions occurring over a time span of approximately 2 months per clinician. Start dates were staggered and the days on which we collected data for any given clinician were not announced in advance. On each day of data collection, we did exit interviews with all the patients the clinician saw in a 4-hour window using the Retrospective Consultation Review (RCR) instrument to measure clinician effort. The RCR is an exit interview instrument intended for patients after their visit to a clinician has ended and is a slightly modified version of the instrument used by Leonard and Masatu (2006).[9] We discuss the RCR data in detail in subsection 3.3.

3.2 Experimental Design

Our study consisted of 3 treatments and a control. We designed the various treatments and the control in order to determine the effort pay-off from conditional and unconditional non-monetary gifts. The gift was a book was 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.[10] One of the 3 treatment groups was given the book immediately and unconditionally, the second was promised the book at a follow-up visit (also unconditionally), and the third was told they would be given the book if they demonstrated adequate adherence to protocol for the 5 items we originally mentioned. In all cases, the book was given to each participant in a private setting to avoid social recognition and competition among participants as factors. Due to the dynamic nature of conditional gifts (used in one of our treatments), all treatments were administered across 2 visits, described below, and the marginal effects of conditional versus unconditional gifts are

[9] All of the questions used on the RCR are listed in Table 7.

[10] There is a used book market in Tanzania and this book could be sold for between 1 and 3 USD.

identified using measurements of post-follow-up effort. Measurements of post-encouragement effort allow us to identify the immediate impact of promising an unconditional gift (compared to the control). We explain this in more detail below.

We describe the experimental design in 2 parts, focusing first on the order of research visits, as this was constant across all health workers. Then we discuss the specifics of the experimental treatments.

3.2.1 Order of Research Visits

The order of the research is laid out in Figure 1. In the course of the study, we met with each clinician up to 4 times and collected data to measure their effort on 5 to 7 separate occasions. Visits described above the time line in Figure 1 are those visits in which a member of the research team met with participants. Visits described below the time line are those visits in which enumerators interviewed patients to collect data. The data-collection visits were not announced, and there was no direct contact between the research team and the participants during these visits. Because data on effort were drawn from exit interviews with patients, the clinician did not necessarily know that the research team was present during a data-collection visit. The study was double blind: neither patients nor the enumerators collecting the data from them knew the treatment assignment of clinicians. In fact, patients and enumerators knew nothing about the nature of the experiment.

Our first meeting with clinicians (Visit 1) was to enroll them in the study. This visit occurred before any data were collected.[11] Each clinician was subsequently visited for initial quality assessment (Visit 2). This data-collection visit had 3 parts, measuring effort in (i) the baseline, (ii) under peer scrutiny, and (iii) after peer scrutiny. Usually all 3 of these parts of

[11] We obtained permission first from facility managers and then from individual clinicians. (Sometimes these were the same people.) Each clinician signed 2 consent forms: one for participation in the direct consultation aspect of the study and one for participation in the randomized encouragement-gift part of the study. Clinicians were not told what the 4 treatments were, but were informed that they would be randomly assigned to one of 4 treatment groups and that they may not have the same experience as their peers in this part of the study. We used this "plain language" so as to be as clear as possible given that most of the participants had never been involved in a similar type of research.

Visit 2 took place on the same day. Because data were collected by interviewing patients after they had left the consultation, the clinician did not realize that data were being collected during the baseline portion of Visit 2, so we were able to measure the normal quality of care. During the peer scrutiny portion of Visit 2, 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 and Masatu, 2006). Lastly, for the third portion of Visit 2, we measured effort after the peer had left the room: post-scrutiny. We included the post-scrutiny measure to ensure that scrutiny did not have a lasting impact and that health workers were not forced to reduce effort (below baseline levels) following scrutiny to "catch-up" or smooth out total effort.

After the initial data-collection visits, clinicians were randomized into one of 3 treatments or the control group and then we met with clinicians individually to provide encouragement and to tell them of their treatment status (Visit 3). During this visit Dr. Beatus, a Tanzanian M.D. and lecturer at a health research institution, met each clinician and read the following 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 items 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 items.

We chose these 5 items 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 items is low; 3) when observed by peers, most doctors significantly increase their adherence to these items (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 items that are important and that almost any clinician can adhere to if he or she so chooses.

Following the encouragement visit, our enumerators returned twice to conduct exit interviews (Visit 4 and Visit 5) with each clinician's patients using the RCR instrument. We used 2 visits to increase the number of observations collected per clinician in this period. As discussed above, we did not directly observe the clinicians in their practice and clinicians had no contact with the research team during these data-collection visits. Enumerators did not announce their arrival at the facilities to anyone. After the team had left, however, most clinicians found out that we had visited their facility on that day. Thus, although the data collected are an accurate representation of the quality of care that would have been provided in our absence, our team's presence should have served as a reminder that we were still conducting research. Clinicians' awareness of this fact might therefore have affected quality after the visit, which would be reflected on subsequent visits.

After these 2 data-collection visits, clinicians received a follow-up visit (Visit 6) from Dr. Beatus, with the specific activities of this visit differing by treatment group. Clinicians had no further meetings with the research team after Visit 6. However, enumerators continued

to collect data on clinicians' effort by interviewing patients 2 more times (Visit 7 and Visit 8) during this period.

In the analysis that follows, we combine data on effort from Visits 4 and 5, referring to these as the post-encouragement visits, and from Visits 7 and 8, referring to these as post–follow-up visits. In all, we collected data from 1,496 unique patients over the 3 portions of the first data-collection visit (Visit 2), 1,557 unique patients during Visits 4 and 5, and 1,220 unique patients during Visits 7 and 8.

Although we had originally hoped to follow consistent timing for every clinician, this turned out not to be possible. Health worker schedules changed on a regular basis, making it difficult to ensure that we arrived on a day when the participating clinician was present. Thus, it sometimes took many visits to the same facility to find the clinician delivering care. Figure 2 shows the time distribution of the 4 data visits after the encouragement visit. Most of the post-encouragement visits took place within 3 weeks of the initial quality-assessment visit for each clinician, though some were as late as 2 months after the initial visit. The post-follow-up visits occurred much later, with some as early as 4 weeks after the initial visit but most about 10 weeks after that visit.

3.2.2 The Experimental Treatments

The over-arching goal of the research was to see if unconditional gifts could increase effort for a policy-relevant period of time (2 to 4 weeks). To help us understand how unconditional gifts work, we chose to address a number of items, including 1) unconditionality itself, 2) the possible effects of timing on gift-giving, 3) the implicit encouragement that may be part of any visit from a research team, and 4) the possible effects of being promised feedback after receiving a gift. We cannot control for all combinations of these factors but chose to 1) allow for one group to earn the conditional gift through effort over time, 2) have the conditional prize and the promised, unconditional gift given at the same point in the research, 3) include a treatment that would allow us to separate out impact of feedback on performance from

the impact of receiving a gift but no feedback, 4) include a control with no gift or feedback for comparison with receiving the unconditional gift.

The differences among the control group and the 3 treatment groups, described below, are outlined in Table 1.

Defining the control group After being enrolled during Visit 1 and having the baseline quality of their care measured during Visit 2, all participating clinicians, including those in the control group, received the encouragement visit (Visit 3) in which Dr. Beatus read the script with information about the 5 important tasks for providing quality care. Following this visit, there was no additional direct contact with the control group, but we did continue to collect data for that group in order to control for the impact of being researched and for the information potentially transmitted during the encouragement visit concerning the 5 important items.

Treatments 1, 2, and 3 The treatments are outlined in Table 1. Participants in the 3 treatment groups all received or were given the opportunity to earn the same gift, but the giving of the gift varied in conditionality and timing among groups. T1 is the delayed, unconditional gift treatment (*delayed gift*). After listening to the encouragement script, T1 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. T2 is the early, unconditional gift treatment. T2 (*early gift*) clinicians were given a gift immediately after the encouragement script and 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 to date but there was no further encouragement or gift. The T3 group, (*prize*), was 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 T1). We also provided T3 participants feedback during the follow-up

visit (as we did for T2). Note that the prize treatment was not a competition among the clinicians, and clinicians did not know whether others had earned the prize. Also, the 70% threshold level was chosen so as to ensure that most health workers would be able to earn the prize: in fact, 90% of the clinicians in T3 did earn the prize. Since earning the prize is endogenous, we do not control for this in any of our empirical work below.

We designed T2 so that comparison of T2 post–follow-up performance with T3 post–follow-up performance would provide a clear measure of the impact of the earned gift: both T2 and T3 received feedback, but only T3 received the earned gift after feedback. Since some form of feedback was unavoidable in the prize treatment (earning the prize implied a measured level of performance) we decided to duplicate the feedback given in both treatments. The feedback was neutral and was not intended to impact effort directly.[12] Finally, we designed T1 as an unconditional gift given later—with the same timing as the prize—to allow us to compare gifts and prizes with the same timing structure (T1 vs. T3) and to compare gifts given at different points in time (T1 vs T2). However, this is not strictly identified, as T1 does not have the promise of feedback. (We did not include feedback in T1 because subjects might perceive a future gift as conditional if they knew there would be feedback.)

The primary results of our study are from data comparing effort between treatments after each of the intervention visits as follows. Note that T1e refers to effort after encouragement for T1, T2e after encouragement for T2, etc., and T1f, etc., refers to effort after follow-up:

- The impact of promising a gift: → [a] T1e - Ce
- The relative impact of an unconditional gift over a conditional gift:
  - After encouragement: → [b1] T3e – T2e
  - After follow-up: → [b2] T3f – T2f

[12]We provided information on the frequency with which they performed the 5 specified tasks. The neutral nature of the feedback means that we did not, a priori, expect it to impact effort in and of itself. This is in line with findings in medical interventions with this kind of feedback (Jamtvedt et al., 2003).

- The impact of feedback only: → [c] T2f – Cf
- The impact of an unconditional gift: → [d] T1f – Cf
- The impact of a delayed unconditional gift versus a delayed conditional gift, controlling for the impact of feedback: → [e] [T3f – T1f] – [T2f – Cf]

Results from these comparisons are reported in Table 4. Further, our regression analysis allows us to control for baseline and post–encouragement effort in each treatment.

Despite the fact that the encouragement visit functions primarily as a set-up for the follow-up visit, we also use it to learn something about the value of a promised gift in the future (T1e vs Ce) and the relative value of promising a conditional gift versus giving an unconditional gift without delay (T2e vs T3e). While the latter is a combined effect comparison (unconditional + gift now vs conditional + gift later), it may be useful for policy purposes. Note that with this design, we cannot strictly identify the value of an unconditional gift given without delay (and without feedback) over the promise of a future unconditional gift or over the promise of feedback alone for improving performance.

Our design 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 5 items important for providing quality care that we highlighted in the study, and the *prize* treatment was told their gift was conditional on performance on these items. Thus, for all treatments, but especially for conditional prize treatment, one may expect some substitution of effort from unmentioned to mentioned items, particularly as the latter are used as the announced criterion of quality care.

3.3 Data and Empirical Specification

Our data are taken from the patient exit interviews conducted on data–collection visits (Visits 2, 4, 5, 7 and 8) using the RCR instrument. The RCR instrument captures whether

or not the clinician did the tasks he is required to do by asking patients or their caregivers, shortly after their consultation, if the clinician did those items.

The correlation in reported performance between patient recollection (from the RCR) and peer scrutiny is 76%. Leonard and Masatu (2006) examine the correlation between changes in patient reports and changes in observer reports for these same instruments and show that, even though patients have positive bias overall, patient reports and observer reports are significantly correlated after controlling for both clinician and item effects: patients are too positive, but they are more positive when the clinician does more.

Thus, we have data on a series of items for each patient, and x_{ijk} describes whether a clinician, j, performed task k that is required by protocol for patient i. x_{ijk} is equal to 0 if the clinician did not perform the task and 1 if the clinician did complete the task. These tasks can involve greeting the patient and offering him or her a chair, asking the patient how long he or she has been suffering from particular symptoms, asking about additional symptoms, examining the patient, and explaining the diagnosis properly, for example. The discrete items required by protocol differ according to the presenting symptoms of the patient and the type of patient. We use the protocols for 4 types of symptoms (fever, cough, diarrhea and general) and 2 types of patients (older than or younger than 5 years). During the RCR interview, patients are only asked about items that apply to their symptoms and age category. There are 74 total items with a much smaller subset applying to any given patient.

We estimate the impact of our 3 treatments on effort using a non-linear logistic regression specification that controls for the characteristics of specific items and the clinician's ability. We use interaction terms ($\vec{T_i}$) in our regressions to allow for each treatment group to respond differentially to the experimental interventions (encouragement visit and follow-up visit). Thus, the key explanatory variables of interest are dummy variables indicating the treatment group of each clinician interacted with whether data were collected during one of the post-encouragement visits or during one of the post-study visits.[13]

[13]We include participant fixed effects in our analysis and therefore control for any differences between clinicians and treatment groups before the assignment to treatment category.

The regression is an adaptation of Item Response Theory (IRT)[14], which simultaneously solves for a difficulty (β_k) and discrimination (α_k) score of each item and the baseline ability (θ_j) of each clinician:

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

The difficulty score (β_k) is similar to an item-fixed effect, and the discrimination score (α_k) measures the importance of clinician ability (θ_j) in providing the specific item. 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 are included in $\vec{T_i}$. We model treatment as increasing the probability that clinicians will perform an item multiplied by the discrimination score for that item: more weight is given to more important items.

In essence, IRT develops a view of quality that is defined by the observations within the data set. Some items 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 item according to its importance in measuring quality.

The standard errors for this regression are derived from 500 bootstrapped samples drawn from within the visit group[15] with replacement at the patient level (all items for a particular visit are either sampled or not sampled). By sampling at the patient level, we take into account the dependence (clustering) of the multiple observations for a particular patient

[14]See Birnbaum (1967); Bock and Lieberman (1970); Das and Hammer (2005); Leonard et al. (2007) for application to quality measurement in health care.

[15]Visit 2; Visits 4 and 5; and Visits 7 and 8

(Cameron et al., 2008)

As robustness checks, we include 4 additional specifications in our discussion:

2. Logit model for each item with item-specific dummy variables ($\vec{D}$): $x_{ijk} = \Phi(\vec{D}\vec{d} + \vec{Z}_i\vec{z} + \vec{T}_i\vec{t}) + e_{ijk}$

3. Logit model for each item with item-specific dummy variables and patient random effects: $x_{ijk} = \Phi(\vec{D}\vec{d}+\vec{Z}_i\vec{z}+\vec{T}_i\vec{t})+e_i+e_{ijk}$. We can estimate random effects because there are multiple items for each patient; however, we cannot estimate patient fixed effects because each patient is uniquely assigned to a clinician (and therefore treatment).

4. OLS model for each item with item-specific fixed effects and errors corrected for clustering at the patient level: $x_{ijk} = \vec{Z}_i\vec{z} + \vec{T}_i\vec{t} + e_k + e_{ijk}$

5. OLS regression of patient-level protocol adherence (percentage of all required items actually performed for each patient) with doctor fixed effects: $\bar{x}_{ij} = \vec{Z}_i\vec{z}_i+\vec{T}_i\vec{t}+e_j+e_{ij}$. This specification controls for correlation between item performance at the visit level and clustering at the doctor level, but ignores item difficulty.

In addition to these robustness checks, we were concerned that clinicians might hear (from patients or nurses) that the research team had arrived and then start changing their behaviour so that the patients would report improvements. 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 (false) quality. To examine this possibility, we look for trends in the quality of care with the order of patients. In the absence of any research presence, the quality of care declines slightly over the day; those who are very sick tend to arrive early at the health facility. Indeed, in the baseline, we see that quality does decline slightly over the course of the day. However, if the clinician found out we were present on any given day, effort should increase. To test for this possibility, we include in all regressions a variable to measure the change in quality over

the course of the day after assignment to treatment categories. We find that this variable is always negative and rarely significant: there is no evidence that clinicians noticed we had arrived and then increased their effort.

4 Results

Before discussing the results of the experiment, we confirm that quality, along with other relevant characteristics, is uncorrelated with assignment to treatment group. Table 2 shows the distribution of characteristics across treatment groups. Average clinician ability is nearly identical across treatments. There is some variation across groups for other variables, but a regression of each of these characteristics on treatment dummies shows no significant correlations except in years of experience. Since we are examining changes in effort, this difference will be removed and is not evidence of a failed randomization.

Table 3 shows the results from examining the performance of clinicians across the 3 different treatments (with the control as the omitted category) and in the post–encouragement and post–follow-up periods.[16] Since all clinicians are examined in the baseline, we can estimate a participant fixed effect and therefore the baseline visit is an omitted variable in our specification.

Recall that the dependent variable is whether the clinician completed any number of symptom-specific tasks required of him/her by protocol during a consultation. The first model (column one) corresponds to the results for the full sample of clinicians and all protocol items that were observed using the RCR. The second model (bridging columns 2 and 3) reports the differential impact of whether an item was mentioned during the encouragement visit. The overall impact is reported in column 2 and the additional impact of the intervention for items that were explicitly mentioned is reported in column 3. Model 3 (column 4) is the same regression as model 1, but with the restricted sample of clinicians who finished all stages of the research.

[16]The coefficients for the difficulty and discrimination are reported in Table 7 in Appendix A.1.

Table 4 shows the differences and significance from comparing the treatments to each other, drawn from the results shown in the first column of Table 3. Note that in this table we report differences in percentages, rather than the original coefficients (each number is multiplied by 100). Rather than comparing effort to the baseline, each of the treatments is compared to each other and to the effort shown under scrutiny and after scrutiny. The treatment effects are shown in both the post–encouragement phase (Visits 4 and 5) and the post–follow-up phase (Visits 7 and 8) as well as for the combination of the 4 visits.

We structure the discussion of results as follows: we first discuss the impact of our study itself on performance, i.e., the role of continued contacts between the research team and the health facility on the performance of workers. Here, we concentrate on the temporal features within the full sample as seen in the control. In a second step, we explicitly investigate the short- and long-term effects of the respective treatments by considering the marginal effects of each treatment relative to the baseline and compared to other treatments. Except where otherwise noted, we examine the impact as seen in the first model (our preferred regression).

4.1 The Impact of Being Studied

Column (1) of Table 3 shows that clinicians increased their effort by about 3.5 percentage points when they were under the scrutiny of a peer, but returned to their normal level of effort immediately after the peer left the room (post-peer scrutiny is not significant). Thus, there is no long-term impact of being observed by a peer and, in particular, no temporal effort substitution: clinicians did not substitute their effort over the remaining patients after they had worked harder when under scrutiny. The combination of these 2 effects shows that the average clinician *can* increase his or her effort without any new information and that this increase does not come at a cost to other activities.

Clinicians increased their effort after the encouragement visit by 2.6 percentage points, almost the size of the Hawthorne effect, despite the fact that no peers were watching them provide effort. This increase, however, is not statistically significant, suggesting that the

encouragement script, by itself, does not invoke a Hawthorne effect. However, we see a much larger and significant increase in effort later in the study (9.6 percentage points increase, compared to the baseline). As the control group did not receive a follow-up visit, this cannot be due to either the encouragement script or the follow-up. We attribute this increase in quality to the continued presence of the enumerators conducting exit interviews at the health care facility: each additional data-collection visit may have served to remind health workers that they were part of a study, thereby invoking a cumulative positive response. As noted above, these performance increases are not driven by doctors reacting to the arrival of the research team on the specific day of the data visit, but by the fact that they know research activities are ongoing.

Result 1 *A one-time encouragement in conjunction with a continued presence of a research team at the health facility can generate significant increases in performance of workers over time.*

Result 1, which shows that on average 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 occasional presence of the research team at the facility may have created a feeling of being scrutinized.[17] This is in contrast to the "hidden costs of control" idea, where workers decrease effort in response to being scrutinized (Falk and Kosfeld, 2006). Importantly, this effect, if also present in other field settings, has methodological consequences: to investigate behavioural 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.

[17]We acknowledge that this effect is not strictly identified as causal. However, due 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 observed in Visit 2, 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.

4.2 The Role of Gifts and Incentives

Our findings are reported in Table 3 and 4. We first discuss the short-run findings. Health workers who received the unconditional gift at the encouragement visit increased their effort by 4.4 percentage points post-encouragement, which suggests that workers reciprocated upon receiving the gift. Note that this group is the only group with a statistically significant post-encouragement response when compared to the control, even though both *delayed gift* and *prize* do exhibit a response. Promising the book unconditionally but delaying giving it led to an increase of 2.6 percentage points post-encouragement, and announcing the book as a prize for performance above a threshold level led to a similar post-encouragement increase in effort of 2.8 percentage points. Accordingly, Table 4 shows that, when we examine the non-parametric bootstrap test (which is a one-sided test), *early gift* is significantly different from the control (T2e vs. Ce, p-value <0.00) whereas *delayed gift* and *prize* are only marginally different from the control (T1e vs. Ce, p-value=0.07; T3e vs. Ce, p-value = 0.11). On the other hand, the differences among the post–encouragement treatments are not significant.

We find no support for the hypothesis that conditionality improves performance in the short run (test [b1] in Table 1, p-value = 0.77), or over the long run (test [b2] in Table 1, p-value = 0.36). In addition, because the prize is based on performance—and so is by definition delayed—we can compare a prize to a delayed gift (T3e vs. T1e), ignoring the impact of future feedback: the short-run difference in effort is also not significant (p-value = 0.46). Taken together, these comparisons show that, in the short run, the early unconditional gift performs best and there is no gain from making gifts conditional. This provides some evidence for test [a] in Table 1, that a delayed unconditional gift invokes a response.

These short-run findings are in contrast to the treatment effects that persist after the follow-up visit, i.e., after the books are given in *delayed gift* and *prize* treatments. Here, no marginal effect above the control remains for the *early gift* treatment (test [c] in Table 1, p-value = 0.61). This suggests that workers reciprocate upon receiving the gift only for a limited time and that receiving feedback on performance provides no extra boost to perfor-

mance. While this temporal impact of gifts is consistent with results in the literature, we should note that the positive response in our real workplace setting lasts much longer than just the few hours found in Gneezy and List (2006).

Importantly, the impact of receiving a book in *delayed gift* is large and statistically significant (test [d] in Table 1, p-value < 0.00). In contrast, clinicians do not show any increase in effort upon receiving the book when it is given as a prize: the impact of receiving the promised book is significantly greater than earning the book (p-value $= 0.01$) even after we control for the impact of feedback on performance (test [e] in table 1, p-value $= 0.01$). In fact, the increase in performance in *delayed gift* after receiving the gift (5.1 percentage points greater than the control) is almost identical to the early response of clinicians to receiving the gift in the early gift treatment. However, for the *prize* treatment, workers apparently do not feel the need to reciprocate upon receiving the book; they may see receiving it as a reward for their previous performance. Conditionality, therefore, does not work in the long run. While it leads to an increase in performance similar to that exhibited by those in the promised unconditional gift treatment in the short run before the prize is awarded, it performs significantly worse in the longer run after the prize or gift is handed over. It would be better to simply promise the gift unconditionally at the future visit.

This benefit from promising a delayed gift can also be seen from examining the average response across the post-encouragement and post–follow-up effects (panel 2 in Table 4): being promised a gift and given it at a later date generates a total impact that is significantly greater than the control (p-value=0.01); almost significantly larger than receiving an early gift(p-value=0.13); and significantly larger than being awarded a prize based on performance (p-value=0.10). Note that the total impact of the immediate gift is not significantly larger than the impact of the prize (p-value=0.40), but is almost significantly larger than the control (p-value=0.13). On the other hand, the total impact of the prize is not greater than the control treatment (p-value=0.22).

We can summarize our findings as follows:

Result 2 *The performance of health workers increases upon receiving an unconditional gift. The promise of a later gift additionally triggers an immediate reciprocal action such that the promised and delayed giving of a gift outperforms, in aggregate, the immediate delivery of the gift. Gifts that are conditional on reaching a specific performance level are not better than conditional gifts in this context: they do generate a minor positive response until the conditionality is resolved, but fail to trigger reciprocity from workers upon receiving the gift.*

While Result 2 shows benefits from promising a delayed gift, we should note again that the temporal structure of immediate gifts differs from the delayed gift treatment. If one is not interested in the aggregate performance over a longer horizon but instead is interested in immediate increases in effort, the best option is to give an unconditional gift immediately. There is no evidence of temporal effort shifting in any treatment after the initial increase (i.e., health workers do not "take back" their gift by reducing future effort), though effort does return to the same level as the control.

4.3 Task Shifting

So far, the treatment effects are discussed based on the performance averaged across all protocol items. As the performance contract in the *prize* treatment was based on 5 specific items mentioned to all participants in the encouragement visit, we also consider the potential redirection of effort towards these items. Model 2 (columns 2 and 3) of Table 3 shows the same basic results as discussed above, but additionally distinguishes the marginal effect of items being mentioned in the encouragement visit. The peer scrutiny and post–scrutiny impacts do not change because items were not highlighted before the encouragement. Immediately following the encouragement visit we see a small but not significant differential increase in performance among the items that were explicitly mentioned: there is no evidence that health workers reacted differently to the items we mentioned. However, in the later phase there is a significant increase in effort for these items compared to other items. The impact of the 3 treatments is similar to those seen in the previous model, but in model 2 we see that

the effort changes do not depend on items being announced in the encouragement visit. In particular, the benefits from gift exchange or an incentive contract accrue over all items and do not lead to decreases in performance on unmentioned items.

Result 3 *Announcing specific protocol items as being important or using them as the basis for assessing the performance in an incentive contract increases effort towards these items in the long run, but does not lead to a substitution of effort away from other items and does not significantly interact with gifts, promised gifts, or prizes.*

Recall that the control received no feedback, so the gains in the later phase cannot be attributed to feedback. Neither the encouragement nor the feedback visits led to increases in the performance of mentioned items. Rather, it must be that continued scrutiny causes health workers to increase all types of effort, potentially with a focus on the mentioned items.

4.4 Robustness checks

Model 3 of Table 3 examines the smaller sample that finished the whole study using the same regression model used in model 1 in Table 3. Table 5 and Table 6 examine the results for our additional empirical specifications for all items (Table 5) and items differentiated by whether they are primed (mentioned in the script) (Table 6). The results across these specifications are broadly similar: the control group experiences significant gains in the later rounds; *early gift* has a significant response immediately but not in the long run; *delayed gift* has a significant response after the gift is awarded and neither *prize* nor *early gift* has responses in the long run. Importantly, the restricted sample shown in model three in Table 3 exhibits similar responses to those seen in model one.

5 Discussion and Conclusion

In this study, we observed health workers in their natural work setting and investigated how their effort (as measured by protocol adherence) can be increased. In addition to

encouraging them to work harder, we examined the ways that health workers might react to small non-monetary gifts that are given unconditionally or are conditional on reaching a specific performance level. Particularly, we 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.

We 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 generate a significant and positive response from subjects. This persistent effect outperforms the standard Hawthorne effect both in size and duration. With the Hawthorne effect, health workers return to normal levels of effort immediately after a peer leaves the room and even their initial response is only about a third of their long-term response to the continued study. Intriguingly, given their roles in our study and most health care research, the direct roles of 1) encouragement, 2) information on what items are important to researchers, and 3) feedback on performance are not significant. Instead, what we find matters are the sense that the health worker is being studied (and the relationship to the research team implied by the study) and the gifts that we offered.

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 from 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.

Our study further shows that gifts trigger some reciprocal effort by workers: health workers reciprocated after receiving an unconditional gift by immediately increasing their effort. Providing workers with an incentive contract (i.e., making a gift conditional on reaching a specific performance level) fails along 2 dimensions. First, it performs no better

in the short run than immediately handing over an unconditional gift. Second, it does not trigger any additional response at the time that the gift is finally given. This is different when a promised gift is given unconditionally at a later point in time. In fact, promising the gift for a later date triggers both an immediate response and an additional response once the gift is given. As a consequence, unconditional promises that are fulfilled later work better than either unconditional immediate gifts or conditional promises.

We interpret these results as suggesting that gifts accelerate the formation of relationships and evoke a reciprocal response. When the delayed gift treatment group receives the gift during the follow-up visit, the relationship between research team and participant is already formed: we see this in the fact that the control group exhibits a significant improvement in effort at the same point in time. However, within the context of this relationship, the gift evokes a significant additional effort increase: a reciprocation for the gift. This reciprocation is similar in size to the reciprocation observed by the early gift treatment in the first round. However, these are not simply identical reciprocations. Although the early gift treatment increases effort in the first round, both the delayed gift and prize treatments increase effort as well (in comparison to the control). Thus, the idea of a gift (whether given, promised, or suggested as a prize) appears to evoke a response that is separate from the desire to reciprocate. The fact that the combined effect of the delayed gift is greater than that of the early gift suggests that participants have not simply calculated the value of the gift and reciprocated correspondingly. In this study, gifts help to establish a relationship and to evoke reciprocation, and it appears to be economically advantageous to separate the 2 impacts (by first promising then later giving the gift) rather than to combine them (by giving a gift immediately).

The relationship between researcher and subject is clearly important to this study. In our setting, Dr. 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 as Dr. Beatus does about quality, but they are not opposed to quality. Thus, although our research is outside of the normal workplace relationship, once the relationship is formed, we do not see it as very different from the relationship that exists between employer and employee.

Two caveats on the generalizability of this study are necessary. First, the mediocre performance of the incentive contract (conditional prize) here should not be taken to imply that all such contracts backfire. In particular, significant monetary prizes may work where small prizes fail. Rather, our results show that even a gift with low monetary value may trigger substantial performance increases when given unconditionally. Second, consistent with existing studies on gifts, we find that the response to a gift is not permanent. Additional research may consider the cumulative effect of using unconditional gifts as a repeated incentive mechanism.

References

Akerlof, George A, "Labor Contracts as Partial Gift Exchange," *The Quarterly Journal of Economics*, 1982, *97* (4), 543–69.

Andreoni, James and John Miller, "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica*, 2002, *70* (2), 737–753.

Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack, "No Margin, No Mission? A Field Experiment on Incentives for Pro-Social Tasks," Technical Report, CEPR 2012.

Barr, Abigail and Pieter Serneels, "Reciprocity in the Workplace," *Experimental Economics*, 2009, *12* (1), 99–112.

Basinga, Paulin, P. J. Gertler, Soucat Agnes, and J. Sturdy, "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.

Berking, Helmuth, *The sociology of giving*, London: Sage, 1999. Translated by Patrick Camiller.

Birnbaum, Allan, "Some latent Trait Models and their Use in Inferring an Examinee's Ability," in Frederic M. Lord and M. R. Novick, eds., *Statistical Theories of Mental Test Score*, London: Addison-Wesley, 1967.

Bock, R Darrell and Marcus Lieberman, "Fitting a Response Curve Model for Dichotomously Scored Items," *Psychometrika*, June 1970, *35* (2), 179–198.

Bohnet, Iris and Bruno S. Frey, ""The sound of silence in prisoner's dilemma and dictator games," *Journal of Economic Behavior & Organization*, 1999, *38* (1), 43–57.

Bradler, Christiane, Robert Dur, Susanne Neckermann, and Arjan Non, "Employee Recognition and Performance-A Field Experiment," 2013.

Cailleé, Alain, "The double inconceivability of the pure gift," *Angelaki: Journal of the Theoretical Humanities*, 2001, *6* (2), 23–38.

Camerer, Colin, "Gifts as Economic Signals and Social Symbols," *American Journal of Sociology*, 1988, *94*, S180–S191.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, 2008, *3* (90), 414–427.

Carpenter, Jeffrey and Erika Seki, "Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay," *Economic Inquiry*, 2010, *49* (2), 612–630.

Das, Jishnu and Jeffrey S. Hammer, "Which Doctor?: Combining Vignettes and Item-Response to Measure Doctor Quality," *Journal of Development Economics*, 2005, *78*, 348–383.

_ and _, "Money for Nothing, The Dire Straits of Medical Practice in Delhi, India," *Journal of Development Economics*, 2007, *83* (1), 1–36.

_ , _ , and Kenneth L. Leonard, "The Quality of Medical Advice in Low-Income Countries," *Journal of Economic Perspectives*, 2008, *22* (2), 93–114.

Deci, Edward L., "Effects of externally mediated rewards on intrinsic motivation.," *Journal of Personality and Social Psychology*, April 1971, *18* (1), 105–115.

_ , "Intrinsic motivation, extrinsic reinforcement, and inequity.," *Journal of Personality and Social Psychology*, April 1972, *22* (1), 113–120.

Dur, Robert, "Gift exchange in the workplace: Money or attention?," *Journal of the European Economic Association*, 2009, *7* (2-3), 550–560.

Eckel, Catherine C. and Philip J. Grossman, "Altruism in Anonymous Dictator Games," *Games and Economic Behavior*, 1996, *16* (2), 181–191.

Eichler, Rena and Ruth Levine, eds, *Performance Incentives for Global Health: Potential and Pitfalls*, Baltimore, MD: Center for Global Development, Brooking Institution Press, 2009.

Fafchamps, Marcel, "Solidarity Networks in Preindustrial Societies: Rational Peasants with a Moral Economy," *Economic Development and Cultural Change*, 1992, *41* (1), 147–74.

Falk, Armin and Michael Kosfeld, "The Hidden Costs of Control," *American Economic Review*, 2006, *96* (5), 1611–1630.

Fehr, Ernst and Klaus M. Schmidt, "Fairness and Incentives in a Multi-Task Principal-Agent Model," *Scandinavian Journal of Economics*, 2004, *106* (3), 453–474.

Forsythe, Robert, J. L. Horowitz, N. E. Savin, and M. Sefton, "Fairness in Simple Bargaining Experiments," *Games and Economic Behavior*, May 1994, *6* (3), 347–369.

Freidson, E., *Profession of Medicine: A Study of the Sociology of Applied Knowledge*, New York: Harper and Row, 1970.

Frey, Bruno S., *Not Just for the Money: An Economic Theory of Personal Motivation*, Cheltenham: Edward Elgar, 1997.

Gächter, Simon and Armin Falk, "Work Motivation, Institutions, and Performance," Technical Report Working Paper No. 62, IEER October 2000.

Gneezy, Uri and Aldo Rustichini, "Pay Enough or Don't Pay at All," *Quarterly Journal of Economics*, 2000, *115* (3), 791–810.

_ **and John List**, "Putting behavioral economics to work: testing gift exchange in labor markets using field experiments," *Econometrica*, 2006, *74* (5), 1365–1384.

Heyman, James and Dan Ariely, "Effort for payment a tale of two markets," *Psychological Science*, 2004, *15* (11), 787–793.

Jamtvedt, G, JM Young, DT Kristoffersen, MA Thomson O'Brien, and AD Oxman, "Audit and feedback: effects on professional practice and health care outcomes (Review)," *The Cochrane Database of Systematic Reviews*, 2003, (3).

Kosfeld, Michael and Susanne Neckermann, "Getting More Work for Nothing? Symbolic Awards and Worker Performance," *American Economic Journal: Microeconomics*, August 2011, *3* (3), 86–99.

Kreps, David M., "Intrinsic Motivation and Extrinsic Incentives," *American Economic Review*, May 1997, *87* (2), 359–364.

Kube, Sebastian, Michel Andre Marechal, and Clemens Puppe, "The Currency of Reciprocity: Gift Exchange in the Workplace," *American Economic Review*, June 2012, *102* (4), 1644–62.

Leonard, Kenneth L. and Melkiory C. Masatu, "Outpatient process quality evaluation and the Hawthorne Effect," *Social Science and Medicine*, 2006, *63* (9), 2330–2340.

_ **,** _ **, and Alex Vialou**, "Getting Doctors to do their best: the roles of ability and motivation in health care," *Journal of Human Resources*, 2007, *42* (3), 682–700.

Lindelow, Magnus and Pieter Serneels, "The performance of health workers in Ethiopia: Results from qualitative research," *Social Science & Medicine*, May 2006, *62* (9), 2225–2235.

Maestad, Ottar and Gaute Torsvik, "Improving the Quality of Health Care when Health Workers are in Short Supply," mimeo, Chr. Michelsen Institute 2008.

Mas, Alexandre, "Pay, Reference Points, and Police Performance," *Quarterly Journal of Economics*, 2006, *121* (3), 783–821.

Mathauer, Inke and Ingo Imhoff, "Health worker motivation in Africa: the role of non-financial incentives and human resource management tools," *Human Resources for Health*, 2006, *4* (24).

Mauss, Marcel, *The Gift: The Form and Reasons for Exchange in Archaic Societies*, London and New York: Routledge, 1925. Translated, 1990, by W.D. Halls.

Meessen, Bruno, Laurent Musango, Jean-Pierre I. Kashala, and Jackie Lemlin, "Reviewing Institutions of rural health centres: the Performance Initiative in Butare, Rwanda," *Tropical Medicine and International Health*, 2006, *11* (8), 1303–1317.

Osteen, Mark, ed., *The Question of the Gift: Essays across Disciplines* Routledge Studies in Anthopology, New York, NY: Routledge, 2002.

Palfrey, Thomas R. and Jeffrey E. Prisbrey, "Altruism, Reputation and Noise in Linear Public Goods Experiments," *Journal of Public Economics*, 1996, *61* (3), 409–427.

_ **and** _ , "Anomalous Behavior in Public Goods Experiments: How Much and Why?," *American Economic Review*, December 1997, *87* (5), 829–846.

Posner, Richard A, "A Theory of Primitive Society, with Special Reference to Law," *Journal of Law and Economics*, April 1980, *23* (1), 1–53.

Rigdon, Mary L., "Efficiency wages in an experimental labor market," *Proceedings of the National Academy of Sciences of the United States of America*, October 2002, *99* (20), 13348–13351.

Rowe, A. K., D. de Savigny, C. F. Lanata, and C. G. Victora, "How can we achieve and maintain high-quality performance of health workers in low-resource settings?," *Lancet*, SEP 17 2005, *366*, 1026–1035.

Sadoulet, Elisabeth, Seiichi Fukui, and Alan de Janvry, "Efficient share tenancy contracts under risk: The case of two rice-growing villages in Thailand," *Journal of Development Economics*, December 1994, *45* (2), 225–243.

Schechter, Laura, "Theft, Gift-Giving, and Trustworthiness: Honesty Is Its Own Reward in Rural Paraguay," *American Economic Review*, 2007, *97* (5), 1560–1582.

Serra, D., P. Serneels, and A. Barr, "Intrinsic Motivations and the Nonprofit Health Sector," *Personality and Individual Differences*, 2011, *51* (3), 309–314.

Sliwka, Dirk, "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes," *American Economic Review*, June 2007, *97* (3), 999–1012.

Figure 1: Timing of Intervention

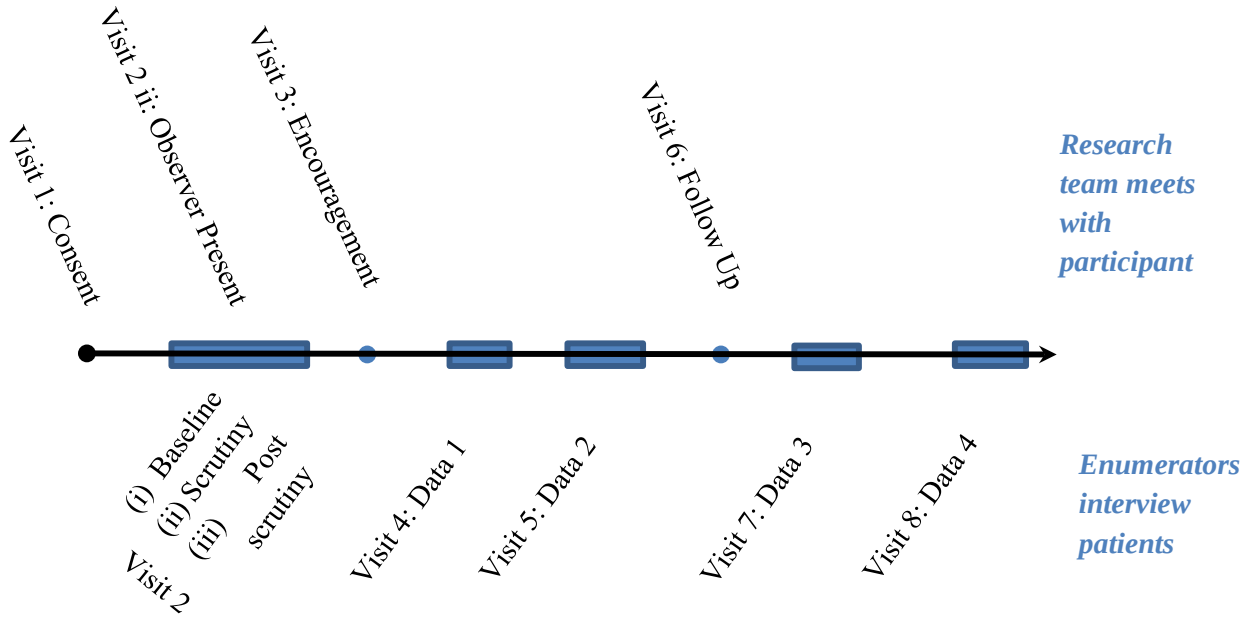

Table 1: Description of Treatments

	Encouragement visit (e)			Follow-up visit(f)	
Group	Enc. Script	Gift	Promised Feedback	Gift	Feedback
Control	yes	no	no ⟩a	no ⟩d	no
T1, *delayed gift*	yes	promised, unconditional	no	yes	no ⟩c
T2, *early gift*	yes	given, unconditional	yes ⟩b1	no ⟩e	yes ⟩b2
T3, *prize*	yes	promised, conditional	yes	yes (if earned)	yes

Figure 2: Gap between Encouragement and Data Collection

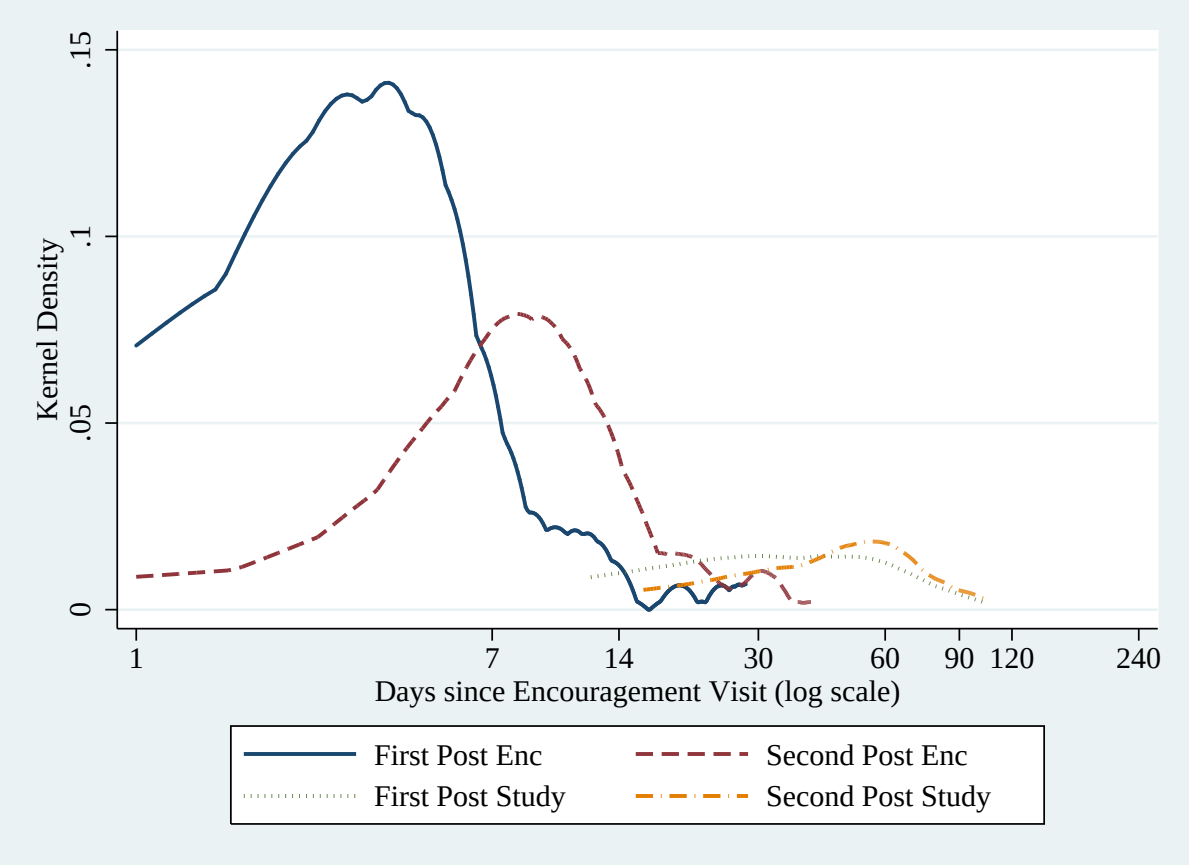

Table 2: Sample Summary Statistics

	Treatment Group					
	Control	T1	T2	T3	Avg.	p-value[a]
Average clinician age	42.54	42.54	43.43	41.00	42.41	0.85
	(11.36)	(8.57)	(9.00)	(8.99)	(9.45)	
Average Years of medical education	3.50	3.62	3.75	3.93	3.69	0.97
	(1.97)	(1.88)	(1.94)	(1.16)	(1.76)	
Average Years working as a healthworker	19.47	15.74	16.47	12.65	16.15	0.06
	(11.39)	(9.19)	(9.02)	(8.24)	(9.65)	
Average Years working with current credential	12.71	10.35	10.89	10.21	11.04	0.48
	(13.56)	(7.87)	(9.19)	(7.90)	(9.73)	
Average clinician ability	0.63	0.64	0.63	0.63	0.63	0.98
	(0.10)	(0.10)	(0.12)	(0.11)	(0.11)	
Percent female clinicians	0.35	0.25	0.17	0.38	0.29	0.26
Percent public	0.44	0.36	0.39	0.19	0.35	0.38
Percent private	0.36	0.44	0.43	0.52	0.44	0.88
Percent NGO	0.20	0.20	0.17	0.29	0.21	0.77
Percent in high volume facilities	0.52	0.48	0.43	0.48	0.48	0.92
Average days between Visit 4 and Visit 3	20.27	17.61	11.12	14.37	15.80	0.77
	(10.20)	(9.92)	(16.10)	(10.53)	(12.62)	
Average days between Visit 4 and Visit 1	22.74	27.44	25.47	24.35	24.98	0.35
	(10.69)	(26.37)	(25.21)	(24.43)	(22.40)	
Average duration of baseline	2.07	0.68	8.83†	0.58	3.19	0.33
	(4.36)	(30.55)	(32.53)	(29.93)	(27.02)	
Number of clinicians	25	25	23	21	94	
Number of patients	1176	1167	1155	940	4438	

Standard deviation in parentheses

†One value of 119 days. Once this value is excluded the average is 1.53 days

a: P-value for test of joint significant from an OLS regression of the observable variable on the treatment groups. P-values are insensitive to clustering or non-linear model specifications.

Table 3: Experimental Results

	whether clinician provided specific item as reported by patient in exit interview (0/1)			
	(1)	(2)		(3)
	all items	all items	primed items	all items
peer scrutiny	0.035***	0.035***		0.044***
	(0.012)	(0.012)		(0.014)
post–peer scrutiny	0.014	0.015		0.017
	(0.013)	(0.013)		(0.015)
Visits 4 and 5: Post–Encouragement				
Overall	0.026	0.020	0.016	0.033*
(control is omitted cat.)	(0.016)	(0.016)	(0.010)	(0.017)
T1	0.026	0.032*	-0.018	0.018
(promised book as gift)	(0.018)	(0.018)	(0.013)	(0.018)
T2	0.044**	0.040**	0.013	0.053***
(given book as gift)	(0.019)	(0.019)	(0.014)	(0.020)
T3	0.028	0.031	-0.012	0.038
(gift to be given as prize)	(0.022)	(0.022)	(0.015)	(0.024)
Visits 7 and 8: Post–Study				
Overall	0.096***	0.083***	0.040***	0.099***
	(0.021)	(0.021)	(0.015)	(0.022)
T1	0.051**	0.045**	0.017	0.046**
(follow up, given book)	(0.022)	(0.022)	(0.019)	(0.021)
T2	-0.005	-0.009	0.013	-0.001
(follow up)	(0.021)	(0.021)	(0.019)	(0.021)
T3	0.003	0.003	-0.001	0.007
(follow up, book awarded)	(0.022)	(0.023)	(0.022)	(0.022)
patient order	-0.002	-0.002		-0.001
	(0.002)	(0.002)		(0.002)
patient order after enc.	-0.003	-0.003		-0.0073
	(0.007)	(0.007)		(0.008)
Observations	96251	96251		84918

Marginal Effects reported. Standard Errors in Parentheses: significance at 1% (***), 5% (**), and 10% (*)
Controls for Gender, G, Age (A) Model is non-linear logistic model controlling for individual clinician latent practice quality, discrimination and difficulty factors for each specific item: standard errors are derived from 500 bootstrapped samples.
(1): Full sample, all items
(2): Full sample, first column shows all items, second column shows the differential effect for primed items
(3): Restricted Sample, only health workers who finished all stages of the research, for all items.

Table 4: Summary of differences in percent change in effort between treatments.

				Visit 4, 5				Visit 7, 8			
		scrutiny	post scr.	Ce	T1e	T2e	T3e	Cf	T1f	T2f	T3f
scrutiny			-2.06*	-0.95	1.70	3.49**	1.86	6.06***	11.12***	5.53***	6.34***
post–scr.				1.11	3.76**	5.55***	3.92**	8.12***	13.18***	7.59***	8.40***
	Ce				2.64*	4.44***	2.80!	7.01***	12.06***	6.48***	7.28***
Visit	T1e					1.79	0.16	4.37**	9.42***	3.83**	4.64**
4, 5	T2e						-1.63	2.57*	7.63***	2.04*	2.85*
	T3e							4.21**	9.26***	3.67**	4.48***
	Cf								5.05***	-0.53	0.27
Visit	T1f									-5.59***	-4.78**
7, 8	T2f										0.81
	T3f										

		Ce+f	T1e+f	T2e+f	T3e+f
scrutiny		2.56*	6.41***	4.51***	4.10***
post–scr.		4.62***	8.47***	6.57***	6.16***
	Ce+f		7.70***	3.90!	3.08
Visit 4,	T1e+f			-3.79!	-4.62*
5, 7, 8	T2e+f				-0.83

Each number represents the difference between the column coefficient and the row coefficient from column 2 of Table 3 multiplied by 100 to represent percentage points. Visit 4, 5, 7, 8 is the average of the coefficients for Visit 4, 5 and Visit 7, 8: the average impact of the two treatment periods. Significance values come from non-paremetric bootstraps, the percentage of times out of 500 that the bootstrapped difference is of the same sign as the reported difference. Significance at 1% (***), 5% (**), 10% (*), and 15%(!)

Table 5: Robustness Tests with All Items

	whether clinician provided specific item as reported by patient in exit interview (0/1)			Protocol Adherence
	(1)	(2)	(3)	(4)
peer scrutiny	0.020***	0.017**	0.025**	0.026***
	(0.004)	(0.007)	(0.010)	(0.009)
post–peer scrutiny	0.004	0.005	0.006	0.016*
	(0.004)	(0.008)	(0.010)	(0.010)
Visits 4 and 5: Post–Encouragement				
Overall	0.010*	0.002	0.012	0.011
(control is omitted cat.)	(0.006)	(0.011)	(0.014)	(0.013)
T1	0.017**	0.018	0.019	0.024
(promised book as gift)	(0.006)	(0.011)	(0.016)	(0.015)
T2	0.031***	0.036***	0.037**	0.057***
(given book as gift)	(0.006)	(0.009)	(0.017)	(0.015)
T3	0.017**	0.016	0.019	0.028*
(gift to be given as prize)	(0.007)	(0.012)	(0.018)	(0.016)
Visits 7 and 8: Post–Study				
Overall	0.055***	0.040***	0.068***	0.064***
(no follow up)	(0.005)	(0.010)	(0.014)	(0.013)
T1	0.028***	0.032***	0.022	0.034**
(follow up, given book)	(0.007)	(0.010)	(0.016)	(0.016)
T2	0.000	0.016	-0.003	0.022
(follow up)	(0.008)	(0.013)	(0.018)	(0.017)
T3	-0.003	0.016	-0.012	0.008
(follow up, book awarded)	(0.008)	(0.013)	(0.018)	(0.017)
patient order	-0.001***	-0.001	-0.001	-0.002
	(0.000)	(0.001)	(0.001)	(0.001)
patient order after enc.	-0.001	0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	96251	96251	96251	4381

Marginal Effects reported. Standard Errors in Parentheses: significance at 1% (***), 5% (**), and 10% (*)
Controls for Gender, G, Age (A) (2): Logit regression with dummies for item effect, controlling for gender (G) and age (A) of patient;
(3) Logit regression with random effects at the unique patient level, with dummies for item effect
(4) Fixed effect linear regression with fixed effects for each unique item, errors are clustered at the unique patient level;
(5) Fixed effect linear regression of patient average (protocol adherence), with fixed effects for each clinician. Errors clustered at the clinician level.

Table 6: Robustness Checks: primed and Unprimed Items

	whether clinician provided a specific item as reported by patient in exit interview (0/1)		
	(1)	(2)	(3)
peer scrutiny	0.022***	0.015*	0.027**
	(0.004)	(0.008)	(0.011)
post–peer scrutiny	0.008	0.002	0.01
	(0.005)	(0.010)	(0.013)
Visits 4 and 5: Post–Encouragement			
Overall	0.002	-0.001	0.004
(control is omitted cat.)	(0.006)	(0.012)	(0.015)
T1	0.02***	0.021*	0.021
(promised book as gift)	(0.007)	(0.011)	(0.016)
T2	0.028***	0.035***	0.035**
(given book as gift)	(0.006)	(0.010)	(0.017)
T3	0.018**	0.017	0.019
(gift to be given as prize)	(0.007)	(0.012)	(0.017)
Visits 7 and 8: Post–Study			
Overall	0.044***	0.033***	0.051***
(no follow up)	(0.006)	(0.011)	(0.015)
T1	0.025***	0.029**	0.017
(follow up, given book)	(0.007)	(0.011)	(0.017)
T2	-0.001	0.015	-0.003
(follow up only)	(0.009)	(0.013)	(0.018)
T3	-0.003	0.017	-0.012
(follow up, book awarded)	(0.009)	(0.013)	(0.018)
Visits 4 and 5: Post–Encouragement, primed Items			
Overall	0.019**	0.017***	0.027**
(control is omitted cat.)	(0.007)	(0.006)	(0.012)
T1	-0.014	-0.013	-0.008
(promised book as gift)	(0.012)	(0.010)	(0.015)
T2	0.011	0.005	0.013
(given book as gift)	(0.010)	(0.009)	(0.015)
T3	-0.007	-0.006	-0.002
(gift to be given as prize)	(0.012)	(0.010)	(0.016)
Visits 7 and 8: Post–Study, primed Items			
Overall	0.037***	0.033***	0.065***
(no follow up)	(0.008)	(0.006)	(0.012)
T1	0.016	0.014	0.019
(follow up, given book)	(0.013)	(0.010)	(0.014)
T2	0.004	0.001	-0.003
(follow up only)	(0.014)	(0.012)	(0.016)
T3	-0.001	-0.005	0.001
(follow up, book awarded)	(0.015)	(0.013)	(0.017)
patient order	-0.001	-0.002	-0.001
	(0.001)	(0.001)	(0.002)
patient order, after enc.	-0.001	0.001	-0.001
	(0.001)	(0.001)	(0.001)
Observations	96251	96251	96251

Notes: see Table 5

A Appendix

A.1 Item difficulty and discrimination scores

The IRT analysis of dichotomous data measures the ability of each clinician and the difficulty and discrimination levels of each item. Difficulty measures the difficulty of the item for all participants and discrimination measures the degree to which better clinicians are more likely to get an item 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 7: Items in the RCR list, changes with scrutiny and encouragement and difficulty and discrimination scores

Item type	Item	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)***
	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 8: RCR questions (II)

Item type	Item	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 9: RCR questions (III)

Item type	Item	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 as 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 as 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 10: RCR questions (IV)

Item type	Item	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						
	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 11: RCR questions (IV)

Item type	Item	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)***