Using the Hawthorne Effect to Examine the Gap Between a Doctor's Best Possible Practice and Actual Performance

by

Kenneth L. Leonard and Melkiory C. Masatu

WP 08-04

Department of Agricultural and Resource Economics

The University of Maryland, College Park

Using the Hawthorne effect to examine the gap between a doctor's best possible practice and actual performance*

Kenneth L. Leonard† Dr. Melkiory C. Masatu‡

May 12, 2008

Abstract

Many doctors in developing countries provide considerably lower levels of quality to their patients than they have been trained to provide. The gap between best practice and actual performance is difficult to measure for individual doctors who differ in levels of training and experience and who face very different types of patients. We exploit the Hawthorne effect—in which doctors change their behavior when a researcher comes to observe their practices—to measure the gap between best and actual performance. We analyze this gap for a sample of doctors, examining the impact of the organization for which doctors work on the performance of doctors, after controlling for their ability. We find that some organizations succeed in motivating doctors to work at levels of performance that are close to their best possible practice. This paper adds to recent evidence that motivation is at least as important to health care quality as training and knowledge.

JEL Classification: I1, O1, O2
Keywords: Motivation, Practice Quality, Health Care, Tanzania, Hawthorne Effect

*The data used in this paper were collected under funding from NSF Grant 00-95235 and The World Bank, and with the assistance of R. Darabe, M. Kyande, S. Masanja, H. M. Mvungi and J. Msolla. The authors are solely responsible for the data contained herein. We extend our appreciation to the Commission for Science and Technology (COSTECH) for granting permission to perform this research. The paper has benefitted from the comments of the audiences at the World Bank and the Center for Global Development.

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

‡Centre for Educational Development in Health, Arusha (CEDHA). PO Box 1162 Arusha, TZ. cmasatu@cedha.ac.tz

Training, ability and capacity are clearly necessary for the delivery of health services in low-income countries. However, there is evidence that these inputs are not sufficient; many doctors choose not to do what they have the knowledge and capacity to do.[1] In developed countries, the presumption that doctors do not always use their knowledge and skills in their patients' best interests—imperfect agency—drives much of the research on health care. In such settings, contracts and regulation are seen to improve the quality of care without increasing doctors' capacities. Clearly, most developing countries lag far behind developed countries in the capacities of their of health care sectors, but they also lag in their ability to regulate the behavior of health care providers. Thus, even in settings where capacity is clearly insufficient, imperfect agency may reduce quality below even the low level of capacity.

One important step to understanding the degree to which doctors underperform and how institutions can reduce or eliminate this behavior is to measure the gap between a doctor's best possible care and the care that he chooses to provide to his patients—coined the "know-do gap" by Maestad and Torsvik (2008), and described by Das and Hammer (2007b) and Leonard et al. (2007). In this paper, we advance an experimental methodology that allows us to measure both best possible and actual care for a doctor performing the same activities with the same types of patients, and therefore to document this gap. We examine the gap between best and actual practice for two key activities in a sample of doctors from Arusha region in Tanzania and then show how institutional features of these doctor's practices are correlated with the size of this gap.

In our study of health care quality in Arusha region, we discovered that our research team caused a distinct Hawthorne effect, in which the act of being observed alters the subject's behavior.[2] In particular, when a doctor on our research team arrived to observe a doctor

[1] For empirical evidence that doctors in developed countries underperform relative to their capacities, see Banerjee et al. (2004a,b); Chaudhury and Hammer (2004); Das and Hammer (2005, 2007a); Das and Sohnesen (2007); Filmer et al. (2000); Leonard and Masatu (2005, 2007).

[2] This effect was originally documented in Mayo (1933), and is well-described by Benson (2000). Both the methodology of the original experiment and the description of the Hawthorne effect have since been called into question (Jones, 1992; Kolata, 1998; Wickstrom and Bendix, 2000), but the original understanding of this effect has survived these debates.

in the course of his regular outpatient consultation, the observed doctor changed the way he practiced medicine and significantly improved the quality of care provided. Surprisingly, these same doctors gradually revert to their normal behavior even while the research team is present. We suggest that the arrival of another doctor puts the subject doctor under high-scrutiny and increases the implied demand for professional behavior. This effect likely depends crucially on the shared training and profession of the researcher and the subject. However, because the researchers are passive and do not provide feedback and because the subject has no direct incentive to impress the researcher, the level of scrutiny and the implied demand for professional behavior both fall over time. The fact that the subject reacts to both high and low levels of scrutiny in the presence of the research team means that we can observe high and low levels of effort with the same quality measurement instrument and with a doctor's normal patients. The high scrutiny implied when the researcher first arrives alters the way the observed doctor treats his patients but it does not alter his capacity to provide care. Thus, the superior quality of care provided in the presence of the research team reveals an achievable, higher level of care that can be compared to the actual level of care.

To demonstrate the potential for this research methodology, we examine two distinct measures of quality; diagnostic quality (effort exerted to find the correct diagnosis) and whether the doctor ordered a lab test for the patient. We measure diagnostic quality as the proportion of medically recommended questions and physical examinations actually asked or performed while examining the patient; quality is higher when doctors ask more questions and examine the patient more carefully. We show that the average doctor provides about 50% of the recommended inputs for the average patient, but increases his provision of effort by 10 percentage points when our research team first arrives. Importantly, for some organizations in our data, there is almost no gap between best and actual performance, whereas for other organizations, the gap is much higher. We argue that doctors who do not increase their performance significantly in the presence of the research team are those doctors who were already performing at high levels—levels close to their capacity—and who therefore cannot

increase the quality of care when subjected to additional scrutiny. On the other hand, doctors who do exhibit large changes in performance are those who were not performing at levels close to best practice and who therefore can easily increase their input levels in response to additional scrutiny. In other words, doctors who are normally under high levels of scrutiny have already increased their effort, whereas those who are normally under low levels of scrutiny have not.

For the use of lab tests, however, it is more difficult to differentiate high from low quality simply by observing the doctor's activities. Low levels of use might indicate a facility that is not sufficiently careful in diagnosing their patients, but high levels of use may indicate supplier-induced demand. In addition, if patients select doctors according to their condition, one doctor might only see patients requiring lab tests while another doctor only sees patients who do not need tests. Thus, use of the laboratory by itself does not reveal quality. We propose that subjecting doctors to additional scrutiny by a peer may cause them to alter their behavior in favor of professional standards. Thus, the Hawthorne effect reveals whether a doctor believes he is using the laboratory in a professional or ethical manner. We find that, whereas most doctors increase their use of laboratory tests when the research team first arrives (and subsequently decrease their use), doctors in one organization have the opposite pattern: they significantly reduce their use of tests when the research team first arrives, allowing the rate to rise over time. Importantly, this organization is suspected of engaging in supplier-induced demand to the detriment of their patients. Thus, even when we cannot objectively evaluate an organization's activities, the behavior of doctors may suggest that they do not believe they are practicing at the best possible levels of care.

The association between the know-do gap and key institutional characteristics of a doctor's practice confirm the findings of Das and Hammer (2007b) and Leonard et al. (2007) in which ability and practice quality were measured using different instruments. Specifically, these papers show that, when tested on their knowledge of medical protocols with case study patients (vignettes), most doctors exert far more effort than they do with their normal pa-

tients and that this know-do gap decreases when doctors have extrinsic motivation to exert effort. Das and Hammer (2007b) proxy for motivation with whether a doctor practices in the private or public sector and show that, when compared to the public sector, doctors in the private sector practice at levels of diagnostic quality that are closer to their ability. Similarly, Leonard et al. (2007) proxy for motivation with the degree to which authority over fiscal and staffing decisions is decentralized to the facility and show that doctors who work under decentralized authority practice at levels closer to their ability than do doctors who work under centralized authority.

The findings in these papers rely on two untested assumptions about the relationship between quality as measured by vignettes and quality as measured by observation with regular patients. First, they assume that two doctors with similar scores on a vignette are, in fact, similar in their ability to diagnose actual patients. Second, they assume that two doctors with different practice quality scores are, in fact, different in the quality of their practice. Because the vignette measures the ability of doctors to *describe* diagnostic procedures, not their ability to *implement* diagnostic procedures, the first assumption would be violated if some doctors were good at describing procedures but unable to perform them in practice. The second assumption would be violated if the observed ability of a doctor depends on the types of patients he is diagnosing. If either of these two assumptions is violated and if the distribution of vignette-specific skills or patient characteristics is correlated with proxy measures of motivation, then concluding that motivation impacts practice quality is not justified. Consider a public and private sector doctor who have identical vignette-measured ability but demonstrate different behavior with their patients. The differences in practice quality could be driven by motivation or they could be driven by the fact that patients at public facilities suffer from illnesses that do not require extra diagnostic procedures, whereas those at private facilities suffer from illnesses that do require these procedures. The practices of these doctors will differ because they see different patients, not because they have different motivation to treat their patients; if the private sector doctor saw the public sector patients,

he would behave in exactly the same manner as the public sector doctor. Thus, it is possible that differences between ability and practice are artifacts of the two instruments used. The use of the Hawthorne effect allows us to measure ability and practice quality with only one instrument used under normal working conditions with regular patients, comparing doctors to themselves.

In the following section, we review the data on doctor quality and determinants of motivation used in the paper. Section 2 develops the link between the impact of scrutiny implied by the Hawthorne effect and a doctor's motivation to show that the Hawthorne effect can reveal the existence of gaps between best practice and actual performance. Section 3 examines the association between the proxy measure of motivation and the know-do gap exposed by the Hawthorne effect. In addition, we discuss the significance of these changes in behavior and their implications for patient outcomes. Section 4 concludes.

1 Data and Instruments

The primary data used in this paper were collected over a period of two years from October of 2001 through March of 2003. Thirty-nine health facilities in the rural and urban areas of Arusha region were visited at least two times each. Doctors who were present at these facilities during any of the visits were evaluated for competence and performance using case-study patients and direct observation respectively. Direct observation allows us to measure both quality and whether the doctors ordered lab tests for their patients. In 2005, we collected additional data at 11 facilities in Arusha municipality, developing a quality measurement instrument that allows us to measure quality even when we do not directly observe the doctor.

1.1 Measures of Quality

The research team used the direct clinician observation (DCO) instrument to measure the actual performance of doctors with their regular patients. DCO measures compliance with Tanzanian protocol and is designed to be sensitive to the limited resources available in the facilities we survey. Every doctor visited was trained in protocol and had the resources at his or her disposal to follow it. Protocol requires history taking (such as asking the patient the duration of the illness or whether diarrhea is accompanied by vomiting) and physical examination (such as taking the patient's temperature or auscultating the chest). With the DCO instrument, a doctor on the research team sits in on the examined doctor's consultations. For each consultation, the observer fills a protocol checklist designed to match patients presenting with fever, cough or diarrhea. For other conditions, there is a more general history taking protocol and one physical examination protocol item. 80 doctors were observed directly and evaluated over 1100 consultations.

In addition, each of these doctors was evaluated using vignettes, which are case-study patients presented by an actor. Vignettes have gained increasing popularity as a tool for quality evaluation both in developing and developed countries (Das and Hammer, 2005, 2007b; McLeod et al., 1997; Murata et al., 1992, 1994; Peabody et al., 1994, 1998, 2000; Tiemeier et al., 2002). There are many possible ways of implementing a vignette; we use the unblind case study with an actor. There are two researchers present: a 'patient' and an examiner. The examiner, after introductions, never speaks, he only observes. The 'patient' presents herself as a patient would, entering the room from outside and leaving after the consultation. She describes her symptoms and answers questions as a patient would. It is explained to the doctor that he must do physical examination by posing questions. The patient then answers the question verbally. For instance, if the doctor says "I would take the patient's temperature", the 'patient' would say "the temperature is 38.5." The examiner then fills a checklist of the expected inputs including expected history taking questions, physical examination items and health education points. Each doctor was tested in their ability for

six typical cases: malaria, pelvic inflammatory disease, diarrhea, pneumonia, flu and worm infestation. 103 doctors were evaluated using the case-study patients.

Additional data were collected in urban Arusha in 2005, using the retrospective consultation review (RCR) instrument. This instrument uses the same checklist as the DCO instrument and it is filled by interviewing patients who have just left the consultation. The RCR questionnaires were administered to 320 patients at 11 facilities. 211 of these patients visited one of the 12 doctors directly observed by the team, and the remainder visited clinicians at the same facilities but who were never observed. On average, we have data on 6 consultations before the team arrived and 11 after we arrived. For consultations that were also observed by the research team, Leonard and Masatu (2006) show that the results from the RCR and DCO instruments are well correlated.

1.2 Doctors and Organizations

The doctors in our sample include nurses of various specializations, clinical assistants, clinical officers, assistant medical officers (AMOs), and medical officers (MOs). Clinical assistants have an elementary school education and three years of medical training. Clinical officers traditionally have O level education and two years of medical training. AMOs are clinical officers with two additional years of training. MOs have both an A level education and five years of university–level medical training. Nurses are not supposed to diagnose but in the rural areas they are frequently the only health personnel present and they do diagnose patients in these circumstances. With the exception of nurses, all clinicians examined in this study diagnose patients, prescribe medicines, and are addressed using the title "doctor." Following the convention in Tanzania, we refer to these clinicians as doctors, even though most of these para-professionals are not full medical officers.

Most doctors in the sample, as in Tanzania, work in the public service in government–run health facilities. In addition, there are seven other types of organizations delivering care in the area, one parastatal hospital (owned by the government but operated as an

independent entity) five private facilities (considered one type of organization) and five faith-based nongovernmental (NGO) organizations operated by the Lutheran, Roman Catholic, Seventh Day Adventist and Church of Gospel International (COGI) churches and the Ithna Asheri Mosque.[3]

1.3 Measures of Institutional Characteristics

We examine the role of institutions using three categorizations of organizations. First, we examine each of these eight organizations as distinct categories. Second, we analyze the performance of the public sector by comparing it to all other organizations combined (non-public). And third, we take advantage of a study of all these organizations conducted by Mliga (2000) and place all facilities on a scale measuring the decentralization of decision–making authority. This third methodology takes into account the fact that the labels applied to some of the organizations are misleading. For example, in a pattern that is not uncommon in Tanzania (Kanji et al., 1992), the COGI facilities are actually private facilities that have franchised the church's name, allowing them to provide services under a preferable tax status. Kanji et al. (1992) suggests that if a facility is franchised to an NGO that does not operate any independent health facilities then that facility cannot be subject to any medical supervision from the franchising organization. For our purposes, therefore, such a facility is private. The index of decentralization reflects these facts, and in addition, the facts that some NGO facilities are highly centralized and more similar to the public sector, while others are decentralized and more similar to the private sector.

The variables used to create the index of decentralization include: a dummy variable indicating whether the chief of post can hire and fire personnel; the level at which salaries are set (national / regional / local); the degree to which the chief of post can use local funds to pay salaries and buy medicines (low / medium / high); and the level at which choices about staffing are made (national / regional / local). These measures are highly correlated

[3]Ithna Asheri is a Shia branch of Islam, the largest school of Shia thought.

and jointly determined, so we examine the impact of an overall decentralization score, not the marginal impact of each characteristic. We create a single index of decentralization by using the first factor from a factor analysis of these variables entered as dummy variables representing each category within each of the four variables (11 categorical variables).[4] Table 1 summarizes the determinants of the index, showing a regression of the index on the ability to hire and fire, decentralization of salary decisions, the degree of local control over financial decisions and the decentralization of staffing decisions.[5] The greatest weight is put on the ability of the chief of post to hire and fire and the three other characteristics have smaller but significant weights. The index of decentralization varies across organizations and across facilities within organizations, but does not vary within a facility.

Effective organizations are likely to be those that manage to provide high–powered incentives for quality to their employees. In this setting, the technology for providing incentives combines medical supervision with either punishment or reward. Although health care suffers from asymmetric information in the doctor–patient interaction, doctors can evaluate the effort and activities of other doctors. Thus, supervising doctors visit facilities and, by observing the activities in that facility, they can assess the quality of care that is provided. In theory, a stakeholder supervises every facility we study (in a single-doctor private practice the stakeholder and the doctor are the same person). In practice, supervision in some organizations is perfunctory. One doctor, who was frequently supervised, stated that in a typical visit the supervisor asked that all logbooks requiring a signature be brought to him as he sat in his still-running (air-conditioned) vehicle. Our index implicitly states that such supervision visits are less likely when the supervisor has the power to act on what he would find if he left the car. Thus, the differences between organizations are not whether they are supervised, but whether the supervisor has the authority to act on what he or she

[4]The factor analysis examines 12 distinct types of facilities, across the organizations studied. There are six significant factors, but the first Eigen value is twice the size of the second and most of the variation is explained by the first factor.

[5]For this regression, the categories local, regional, and national are represented as 3/2/1, respectively, and low, medium, and high as 1/2/3, respectively.

discovers. Our measure of decentralization, therefore, captures the potential effectiveness of supervision while the actual level of supervision may vary. Importantly, we control for doctors who work in single–doctor private practices because the degree of decentralization may have a non–linear impact on doctors for whom there is no outside stakeholder; decentralization with outside supervision is fundamentally different from decentralization without outside supervision.

1.4 Summary Statistics

Table 2 shows the number of consultations observed, the percentage of items correctly used and the average decentralization score for each of the organizations examined in the data. The only two organizations identified by name are the public sector and the collection of purely private facilities. We separate summary statistics by the two different data sets examined. Note that the decentralization score shown is the average over all facilities owned by a particular organization, and for some organizations, there is variance across types of facilities. Note that doctors in the best organization perform 75% of the items required by protocol, suggest that protocol is not an absolute measure of quality.

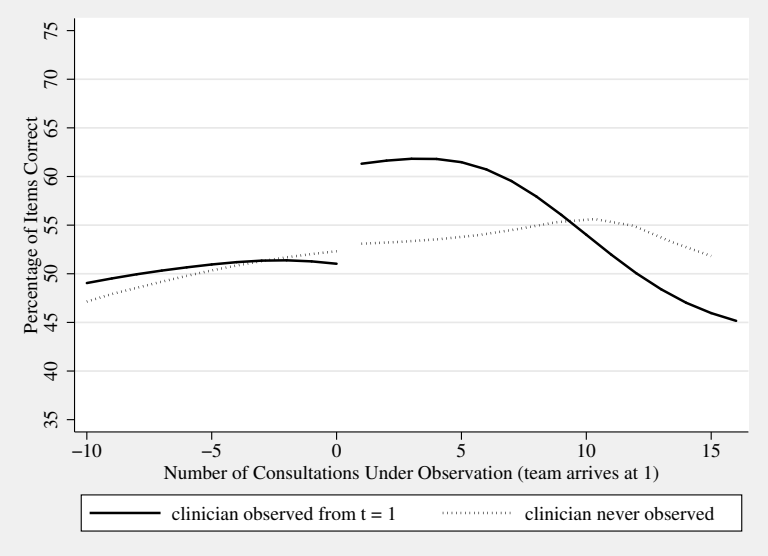
This table illustrates the limited ability of the data to examine the behavior of individual organizations—for some of the organizations studied, we have few observations of patients. This is always because there were few patients on the day we visited, a not uncommon event in rural facilities. Clearly, there is large variation in the performance of some organizations. For example, the one private facility in the rural area of our study has almost no patients and provides poor diagnostic quality, but the four urban private sector facilities are much better and see more patients. The unbalanced nature of the data on organizations (there are too few facilities and doctors in some of the organizations) as well the unbalanced nature of the Hawthorne effect for some doctors (there are too few patients to allow us to observe changes in quality) limit the practical uses of this data to comparing the public sector to the non-public sector and to demonstrating the potential usefulness of the Hawthorne effect

methodology.

2 The Hawthorne Effect as Additional Scrutiny

The Hawthorne effect refers to a situation in which an individual's behavior changes when they realize they are being observed. It is characterized by a positive but temporary change in some measurable behavior in a situation in which there was no deliberate attempt to affect behavior (Benson, 2000). The doctors observed in Tanzania were told explicitly that the research would not impact them in any way, however, they may have reacted to the mere presence of other doctors as if there were a "perceived demand for performance" (Campbell et al., 1995). Thus, in this setting, it is useful to describe the Hawthorne effect as a temporary response to increased scrutiny from a professional peer.

Figure 1: The Hawthorne Effect on Quality



The figure shows smoothed average percentage of items required by protocol as measured from patient exit interviews performed immediately after the consultation. The dashed line shows percentage provided for patients seen immediately before and after the research team arrives at a facility who visited a doctor who was never directly evaluated by the research team. The solid line shows the percentage provided for doctors who were observed by the research team starting at t = 1.

Leonard and Masatu (2006) document the full pattern of the Hawthorne effect with a small sample of doctors practicing in Arusha region in Tanzania. Because we used a patient exit survey, we could collect data for three types of patients: patients who had consultations before the team arrived at a facility, patients consulted after the team arrived whose consultations were observed by the research team, and patients consulted after the research team arrived whose consultations were not observed by the research team. Patients in this third group were seen by doctors who were not evaluated by the research team, but who practice at facilities where other doctors were evaluated. Figure 1 shows the pattern of quality as estimated from patient responses for observed and unobserved doctors. For doctors who were observed, there is a significant jump in quality when the team arrived. However, for doctors who were never observed, there is no significant change in quality. Figure 1 also shows that the Hawthorne effect is temporary; quality rapidly returns to levels similar to those found in the absence of a research team.[6]

The changes in quality observed with the Hawthorne effect can be seen as reflecting differences between best and actual practice. Before the research team arrives, observed quality is equal to normal practice quality. When a researcher arrives, every doctor practices to the best of his ability and after time has passed, every doctor returns to his normal level of ability. Thus, the gap between best and actual practice seen with the Hawthorne effect is a function of the degree to which doctors do not normally practice close to their ability. Doctors who always provide maximum effort show no gap and doctors who regularly shirk show a large gap. Formally, the Hawthorne effect represents changes in the level of scrutiny faced by doctors. All doctors choose to provide quality (q) that is equal to a fraction (λ) of their best possible quality (θ) where $\lambda \in (0, 1)$ and $q \in (0, \theta)$. This fraction is a function of the professional scrutiny at the time that quality is chosen: $\lambda = \lambda(s)$.

The baseline level of scrutiny for each doctor is unknown but the Hawthorne effect in-

[6]Leonard and Masatu (2006) use regression analysis to verify the significance of the change in quality before and after observation, the gradual fall in quality as time passes for observed doctors, the unchanging quality before the team arrives and the unchanging quality for doctors who are never observed.

creases scrutiny. Thus, if $\partial\lambda^2/\partial^2 s < 0$ then the change in the share of best possible quality is greater with additional scrutiny when the baseline level of scrutiny is lower and lower when the baseline level of scrutiny is higher. Practically, this means that doctors who face high levels of motivation on a regular basis have little room to react to additional levels of scrutiny, whereas those who are not otherwise motivated can easily change their behavior. Note, the baseline level of scrutiny may differ for different activities. In particular, it is possible that some organizations provide high levels of scrutiny for diagnostic quality, but low levels of scrutiny for appropriate laboratory test use. In such a case, additional scrutiny would have little impact on diagnostic quality, but may have a larger impact on laboratory use.

In the following section, we examine the changes in the provision of diagnostic quality and the use lab tests. Quality is measured by the probability that doctor j would implement diagnostic input k, from among all diagnostic inputs that are required by protocol for the given patient i: prob$(x_{ijk} = 1)$. The use of lab tests is the probability that doctor j would order any lab test for patient i: prob$(l_{ij} = 1)$. Each probability is a function of the item, patient characteristics, the level of additional scrutiny and the baseline ability and motivation of the doctor. Given the nature of our data, we investigate the role of each of these factors in two different empirical specifications: the change in quality when the research team arrives and the change in quality as the research team continues to observe consultations. We examine diagnostic quality in each specification, but the use of lab tests in the second specification only.[7]

2.1 Changes in scrutiny when the research team arrives

For the small study of 11 facilities, we have data from exit interviews that allows us to compare the quality of care provided before and after the research team arrives. To study the impact of additional scrutiny, we focus on the immediate impact of the additional scrutiny

[7]We do not have data on the use of lab tests for patients whose consultations were never observed.

when the research team arrives and test whether the reaction to this scrutiny varies with our measures of institutional characteristics. Although the data set is small, we can take advantage of the facts that nine of the doctors were practicing in facilities with at least two doctors, that the selection of the doctor to observe was random, and that we have data on the quality of care provided by these unobserved doctors as well as observed doctors. Restricting attention to facilities with paired doctors and to consultations that occurred soon before or after the arrival of the team (four consultations before and after) allows us to pursue what is essentially a triple difference strategy. We compare the difference between the change in quality when the research team arrives for doctors who were observed and doctors who were never observed and then examine how this net change in quality with increased scrutiny varies with the institutional variables we are studying.

We implement this triple difference strategy in a random effect probit regression of the probability that each doctor implemented a given required input, as a function of a constant (α), facility effects (ϵ_f), whether there was a researcher present at the facility (P), whether the consultation was observed by a researcher (O) and whether the consultation was observed by a researcher interacted with the institutional variables. Thus, for example, using the decentralization index for each facility (D_j) we estimate:

$$\text{prob}\left(x_{ijk}=1\right)=f\left(\alpha+\beta_1 P+\beta_2 D_f+\beta_3 O+\delta D_f\cdot O\right)+\epsilon_f+\epsilon_{ijk} \tag{1}$$

The coefficient on decentralization when the doctor is observed (δ) is our estimate of the differential impact of scrutiny on doctors who work in decentralized facilities. If doctors who work in decentralized facilities face high levels of scrutiny in the absence of the research team, then their reaction to additional scrutiny should be smaller than doctors who work in centralized facilities and the coefficient should be negative.

This strategy explicitly controls for selection bias caused by the fact that doctors choose where to work because it compares doctors to themselves (before and after being observed)

and to other doctors who work in the same facility, not to doctors in other facilities.

2.2 Changes in scrutiny as the research team continues to observe

The impact of scrutiny can also be seen in the change in behavior after the researcher has been present for a longer period. The larger data set has information on quality collected by observers and therefore does not contain any data on the quality of care provided before the research team arrives. However, we can measure the impact of scrutiny by modeling scrutiny as decreasing with the length of time that the researcher has already been present. Thus, scrutiny $s_i = -1 * \{\#$ of consultations since the team arrived$\}$.[8]

We model the probability that a doctor will provide an input that is required by protocol (prob$(x_{ijk} = 1)$), as a function of an item specific effect (α_k), illness characteristics ($\vec{Z}_i\gamma_i$), the level of additional scrutiny at the time patient i is seen (s_i), doctor-level effects ϵ_j (which reflect both ability and baseline scrutiny) and an additional error term. We test the hypothesis that the reaction to additional scrutiny is a function of the degree of decentralization for each doctor, D_j as well as for the fact that some doctors practice in single-doctor practices S_j. Thus, we estimate:

$$\text{prob}\left(x_{ijk} = 1\right) = f\left(\alpha_k + \vec{Z}_i\gamma_i + \beta_1 s_i + \beta_2 D_j s_i + \beta_3 S_j s_i\right) + \epsilon_j + e_{ijk} \tag{2}$$

where β_1 is the average impact of additional scrutiny, β_2 is the impact of scrutiny when the doctor works in a facility that is more decentralized, and β_3 is the impact of scrutiny for doctors who work in single-doctor practices. The hypothesis that additional scrutiny has a smaller impact for doctors who normally face high levels of motivation, therefore, translates into the hypothesis that $\beta_2 < 0$.

In addition, we examine the reaction to scrutiny by public/non-public and in each of the

[8] Alternative specifications, including the negative of the log of the number of previous consultations under scrutiny, and the inverse of the number of previous consultations under scrutiny, produce essentially identical results.

organizations in our data for which we have adequate observations. As with the previous strategy, this analysis compares doctors only to themselves, not to other doctors in other types of facilities. However, we do not compare doctors to other doctors in the same facility and cannot measure quality before the research team arrives. Thus, we control for the patient and illness characteristics that might otherwise impact the variation in quality.

To analyze the probability of using a lab test, we follow the same basic strategy, but use only organizational categories, not the degree of decentralization and whether a facility is non-public.

3 Analysis

In this section, we ask whether the size of the Hawthorne effect can be explained with our proxy measures of motivation. We examine the impact in two different data sets, looking first at the reaction to the arrival of the research team, and second at the reaction to the continuing presence of the research team.

3.1 Motivation and the reaction to the arrival of the research team

Here we compare the behavior of doctors before and after the research team arrives, comparing the differences between observed and unobserved doctors in decentralized and centralized facilities. We follow the specification of effort shown in Equation 1 using a random effects probit regression The analysis is restricted to the four consultations before and after the research team arrives and to facilities in which observed doctors can be paired with unobserved doctors.

Table 3 shows three specifications of institutional characteristics. Column 1 shows each of the organizations as a dummy variable interacted with whether or not the consultation was observed (owner 2 and 8 are not represented in this data). Column 2 shows whether the consultation was observed and a dummy variable representing whether the organization

is non-public interacted with whether the consultation is observed. Column 3 uses the decentralization score instead of the non-public dummy. In this data, there is almost no difference between the category non-public and the decentralization score because all non-public facilities are monitored by authorities in the municipality and therefore authority is local.

The presence of a research team at the facility does not change the probability that a doctor will provide a given input, but the fact that a consultation is observed by a member of the research team has a strong and significant impact on that probability. Column 1 and 2 show that most of this change in quality when the doctor is observed is driven by public sector doctors. The only significant change in quality shown in column 1 is for public sector doctors—although some of the coefficients for other organizations are positive they are not significant. This basic result is confirmed in columns 2 and 3. Note that the total change in quality when the research team arrives for non-public and decentralized facilities is the sum of the coefficient for high scrutiny and the coefficient for the institutional variable interacted with high scrutiny. Doctors in non-public facilities and doctors in decentralized facilities have a much smaller reaction to increased scrutiny.

3.2 Motivation and the decline in scrutiny over time

Here we examine the pattern of quality after the research team arrives, using the number of previous consultations observed as a proxy for declining scrutiny. When the number of previous consultations is low, scrutiny is high, so a positive coefficient for the level of scrutiny indicates that quality or lab tests are declining as the team remains. We examine the differential response to scrutiny by the same three institutional measures as above, except that we add a categorical variable indicating single-doctor practices to the decentralization score. For organizations 2 and 8, the average number of consultations observed is less than 7, and therefore we drop these organizational category variables in column 1, though we retain the data. We follow the specification of effort shown in Equation 2. In addition, we examine

the impact of scrutiny on the use of lab tests using only the organization categories. We examine patterns for both diagnostic quality and lab test use with a random effect probit regression.

Column 1 shows that the average doctors in three organizations increase the quality of care provided when they are under high levels of scrutiny, but doctors in other organizations do not have a statistically significant reaction to quality. Column 2 shows that the average doctor increases the quality of care provided when under high scrutiny, but that doctors in non-public facilities are significantly different from the average doctor, and their net change in quality is essentially flat. Column 3 shows the same basic result as column 2, that decentralization is associated with doctors who have a much smaller change in quality when they are under high levels of scrutiny. The coefficient for single-doctor practice is not significant, indicating that, by this measure, doctors who are their own stakeholders are not different from other doctors who face similar levels of decentralization.

Column 4 examines the changes in the use of lab tests when the doctor is under high levels of scrutiny. In three of the organizations examined, the average doctor increases his use of lab tests when he is under high levels of scrutiny. However, in one of the organizations, the average doctor actually decreases the use of lab tests when he is under high levels of scrutiny. Note that these same doctors did not change their diagnostic quality.

3.3 Organizations, Institutions and the Know-Do Gap

We have presented data on the size of the Hawthorne effect by organization categories, whether an organization is public sector or not, and the degree of decentralization in decision-making authority. In some cases (as in the analysis of lab tests), examining the data by organization categories produces some interesting results, however, in general this data is not well suited to analyzing organizations. There are only two organizations with a significant number of doctors and facilities (the public sector and owner 4) and in some of the facilities there were so few patients observed by our team that we could not observe changes in

behavior due to the duration of scrutiny. On the other hand, the difference between the public and the non-public sector is significant whether measured by a categorical variable indicating the non-public sector, or by the measure of decentralization. Despite the large variance in non-public sector performance, the data clearly show that the average non-public sector doctor has a smaller know-do gap than the average public sector doctor.

We have introduced the measure of decentralization as a potential way to differentiate among organizations, and clearly, the know-do gap is decreasing in the degree of decentralization. However, in this case, this variable does not do a better job of explaining the data than the simple dummy variable indicating the non-public sector. Thus, in this analysis, we are unable to make the case that the decentralization of decision-making authority adequately explains the differences between organization studied. Part of our failure to helpfully describe the differences among organizations is because the Hawthorne effect methodology requires a reasonable sample of patients (about 10 per doctor), and many of the organizations studied simply have too few patients.

3.4 Organizations and Lab Tests

We show that most doctors increase their use of lab tests when faced with additional scrutiny, either because they are increasing their diagnostic quality and realize that a lab test is needed, or because they know that a lab test is indicated by the patients' condition, but would normally have ignored this. However, one organization in our data displays the opposite pattern, reducing the use of lab tests when they fall under additional scrutiny. As an isolated finding, this result does not indicate a problem. However, doctors on the research team repeatedly expressed concerns about the use of lab tests by this organizations because they were not indicated by any of the patient's symptoms. As a result of these concerns, at the conclusion of the study we had a conversation with the chief of post in one of the facilities owned by the organization in question. He stated that the organization had explicitly asked doctors to use more lab tests because many of their patients had infrequent contact with

the health care system and it was therefore useful to perform lab tests (like a urine analysis or blood test) to look for undetected conditions. We pointed out that this was fine if the patients agreed, but that since they were paying for tests, it seemed dishonest to let them believe the tests were indicated by their symptoms. He agreed and hoped his doctors were properly informing patients. He also stated that it was time to revisit the policy because of concerns about its use. This was the only organization for which there was any discussion about the use of lab tests and these issues arose after the data were collected and before these results were analyzed. We are not suggesting that there was dishonest intent, but our methodology correctly identified an aberrant behavior worthy of further investigation, demonstrating the usefulness of the Hawthorne effect methodology. Clearly, the doctors we studied were uncomfortable with the policy, particularly when they were asked to see it in the light of their professional and ethical standards.

3.5 The significance of the Hawthorne effect induced changes in quality

The patterns of changes in diagnostic quality observed with the Hawthorne effect are significant, but do they matter? It is not particularly surprising that doctors change the way they practice medicine because they are nervous about the arrival of another doctor, and doctors who are rarely supervised may be more nervous than those who are frequently so. We claim that these change in behavior are significant for two reasons. First, because they affect the one outcome for which we have data and second, because the magnitude of the changes in behavior can be tied to important changes in outcomes as seen in the case study (vignette) patients.

We asked patients if they were satisfied or very satisfied with the quality of care they received and Leonard (2008) examines the changes in patient satisfaction as doctors increase their use of diagnostic inputs. They show that patients are more likely to be very satisfied with the quality of care when the doctor increases his quality because the research team has

arrived.

In addition, data from the vignettes show that the probability of correct diagnosis is increasing in the use of diagnostic inputs. Table 5 examines the probability that a doctor would give the correct diagnosis over six vignettes as a function of diagnostic inputs and doctor characteristics. Since the vignettes were designed by the team to mimic specific conditions we know which diagnoses were correct. Diagnostic inputs were measured using an instrument similar to the DCO instrument, recording whether doctors used history taking and physical examination inputs required by protocol. Table 5 shows that doctors give the correct diagnosis because they provide diagnostic effort, not because of their training, experience or tenure. In addition, it shows that physical examination is much more important than history taking. Increasing the use of physical examination by 1 percentage point leads to a 1.9% increase in the probability of providing the correct diagnosis. On the other hand, an increase in the use of history taking leads to a small and statistically insignificant increase in the probability of correct diagnosis.

Table 6 examines the patterns of diagnostic input provision with the Hawthorne effect, differentiating between physical examination and history taking. Overall, the average doctor provides more of both type of input when he is first observed. However, doctors in decentralized facilities are different from the average doctor for physical examination, but not different for history taking. The average doctor in a decentralized facility exhibits very little change in physical examination while he is being observed by the research team, but does exhibit a decline in history taking over that same period. Thus, doctors in centralized facilities are increasing their use of physical examination when the research team first arrives. Table 6 suggests that the average doctor in a centralized facility changes his use of physical examination inputs by about 20 percentage points between when the team first arrives and 10 consultations later. This change in diagnostic quality is approximately equivalent to 1 standard deviation in the distribution diagnostic quality over the whole sample. If the link between physical examination and diagnosis is the same with an actual patient as it is with

a case study patient, then we predict that the difference between the probability of being properly diagnosed when the team first arrives and the probability after 10 consultations is approximately 38 percentage points (0.02 * 10 * 1.9 = 0.38).

In reality the difference will be much smaller than this estimate, because most patients who visit any doctor are in fact suffering from the presumptive diagnosis; the diagnosis given when doctors do not exert effort. For example, most patients who visit with symptoms of malaria are in fact suffering from malaria, and if the doctor exerts no effort, but gives the presumptive diagnosis, he will be correct. The vignettes were specifically designed to test a doctor's ability to differentiate common from less common illnesses. Thus, the return to diagnostic effort is likely to be much smaller for the common illness. Nonetheless, the changes in effort observed are significant, both statistically and for health impacts.

It is also interesting to note that doctors in decentralized facilities are not imperturbable. They do change their behavior when the research team arrives and ask more history taking questions. The arrival of an outside research team does have an impact on the way even very good doctors provide care, although it does not change actual quality.

4 Conclusion

By using the experimental intervention implied by Hawthorne effect—as a reaction to both the arrival of the research team and the continued presence of the research team—we show that the average doctor is capable of providing much higher levels of diagnostic quality and that these changes in quality could improve important outcomes for many patients. Importantly, the changes in quality caused by the arrival of the research team vary widely across doctors and we show that these changes reflect the baseline motivation of doctors. In other words, those doctors who normally practice at levels close to their abilities do not change their behavior as much as doctors who normally practice at levels much lower than their abilities. Thus, the Hawthorne effect demonstrates the size of the gap between a

doctor's best possible practice and their actual performance.

We examine the determinants of this gap by three measures of institutional characteristics: organization categories, whether the organization is non-public sector, and the degree of decentralization of decision-making authority. We find that the average doctor who works for a non-public sector organization or who works in a facility with decentralized decision-making authority exhibits a much smaller change in quality when the research team arrives. In fact, doctors in some of the best organizations appear to change their behavior only in unimportant ways and maintain high quality whether observed by the team or not. These doctors react to the presence of the research team by increasing their use of history taking, but maintain their use of (the much more important) physical examination. These results, in turn, suggest that the differences in performance between the public and the private sector are not driven by the abilities of doctors in those sectors, but by their motivation to provide quality. This in turn has important policy implications for health care in Tanzania, implications that appear to apply to most developing country health systems.

Combined with the analysis of motivation and health quality in Leonard et al. (2007) and Das and Hammer (2007b), this paper suggests that improvements in health care quality may come as much from focusing on motivation as they do from focusing on training. Given the dismal state of health in developing countries, and the enormous potential for improvements in health status with access to appropriate and high quality medicine, a better understanding of the mechanisms necessary to address motivation in health workers is an essential part of the way forward.

The Hawthorne effect has traditionally been raised as a potential problem in research settings; can the researcher know that what he is seeing is what would happen if he were not there? However, some new research in experimental economics suggests that the Hawthorne effect demonstrates the importance of interpersonal utility; the fact that subjects may derive utility from being perceived to be more charitable, public good–minded, honest or professional than they really are (see Gneezy and List, 2006; Leonard and Masatu, forthcoming;

Levitt and List, 2007, for example). Thus, the presence of another doctor causes a doctor to behave in a more professional manner. In medicine, the impact of increased professionalism is easy to observe; doctors provide more effort and therefore better diagnostic quality. However, perhaps because the observer provides no feedback, the gain in interpersonal utility is short lived and the subject rapidly returns to his or her original level of effort (this return to original effort is also noted in Gneezy and List, 2006). This paper makes the argument that the Hawthorne effect may not be something to be studiously avoided in research settings, but something to be studied more closely, and in particular something to be exploited so as to improve our understanding of the determinants of performance. In settings where the performance of individuals is a combination of ability and effort, the Hawthorne effect may be particularly useful because it does not impact ability, but appears to significantly impact effort. When the changes in effort are driven by professional or ethical concerns, the Hawthorne effect allows us to observe both the best possible care and actual performance, and to examine the determinants of this gap.

References

Banerjee, A., A. Deaton, and E. Duflo, "Health care delivery in rural Rajasthan," *Economic and Political Weekly*, 2004, pp. 944–950.

_ , _ , and _ , "Wealth, health and health services in rural Rajasthan," *American Economic Review*, 2004, *94* (2), 326–330.

Benson, P. G., "The Hawthorne Effect," in W. E. Craighead and C. B. Nemeroff, eds., *The Corsini Encyclopedia of Psychology and Behavioral Science*, 3 ed., Vol. 2, NY: Wiley, 2000.

Campbell, JP, VA Maxey, and WA Watson, "Hawthorne Effect: Implications for Prehospital Research," *Annals of Emergency Medicine*, 1995, *26* (5), 590–594.

Chaudhury, Nazmul and Jeffrey S. Hammer, "Ghost doctors : absenteeism in Bangladeshi health facilities," *World Bank Economic Review*, 2004, *18* (3), 423–441.

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

_ **and** _ , "Location, location, location: Residence, Wealth and the Quality of Medical Care in Delhi, India," *Health Affairs*, 2007, *26* (3).

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

_ **and Thomas Pave Sohnesen**, "Variations In Doctor Effort: Evidence From Paraguay," *Health Affairs*, 2007, *26* (3).

Filmer, D., J. Hammer, and L. Pritchett, "Weak links in the chain: a diagnosis of health policy in poor countries," *World Bank Research Observer*, 2000, *25* (2), 199–224.

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

Jones, Stephen R. G, "Was There a Hawthorne Effect?," *The American Journal of Sociology*, 1992, *98* (3).

Kanji, N., P.M. Kilima, and P.M. Munishi, "Quality of Primary Curative Care in Dar-Es-Salaam," 1992. Unpublished paper.

Kolata, Gina, "Scientific Myths that are too good to die," *The New York Times*, 1998.

Leonard, Kenneth L., "Is patient satisfaction sensitive to changes in the quality of care? An exploitation of the Hawthorne Effect," *Journal of Health Economics*, 2008, pp. 444–459.

_ **and Melkiory C. Masatu**, "The use of direct clinician observation and vignettes for health services quality evaluation in developing countries," *Social Science and Medicine*, 2005, *61* (9), 1944–1951.

_ **and** _ , "Outpatient process quality evaluation and the Hawthorne Effect," *Social Science and Medicine*, 2006, *63* (9), 2330–2340.

_ **and** _ , "Variation in the quality of care accessible to rural communities in Tanzania," *Health Affairs*, 2007, *26* (3), w380–w392.

_ **and** _ , "Moving from the Lab to the Field: Exploring Scrutiny and Duration Effects in Lab Experiments," *Economic Letters*, forthcoming.

_ **,** _ **, 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.

Levitt, Steve and John List, "What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?," *Journal of Economic Perspectives*, 2007, *21* (2), 153–174.

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

Mayo, Elton, *The Human Problems of an Industrial Civilization*, New York: MacMillan, 1933.

McLeod, P. J., R. M. Tamblyn, D. Gayton et al., "Use of Standardized Patients to Assess Between-Physician Variations in Resource Utilization," *Journal of the American Medical Association*, 1997, *278*, 1164–8.

Mliga, Gilbert R., "Decentralization and the Quality of Health Care," in David K.

Leonard, ed., *Africa's Changing Markets for Human and Animal Health Services*, London, also available at http://repositories.cdlib.org/uciaspubs/editedvolumes/5/: Macmillan, 2000, chapter 8.

Murata et al., *Prenatal Care: A Literature review and quality assessment criteria*, Rand Corp, 1992.

_ **and** _ , "Quality Measures for Prenatal Care," *Archives of Family Medicine*, 1994, *3* (1), 41–9.

Peabody, John W. et al., "Quality of care in public and private primary health care facilities: structural comparisons in Jamaica.," *Bulletin of the Pan American Health Organization*, 1994, *28*, 122–141.

_ **and** _ , "The Effects of Structure and Process of Medical Care on Birth Outcomes in Jamaica," *Health Policy*, 1998, *43* (1), 1–13.

_ **and** _ , "Comparison of Vignettes, Standardized Patients, and Chart Abstraction: A Prospective Validation Study of 3 Methods for Measuring Quality," *Journal of the American Medical Association*, 2000, *283*, 1715–1722.

Tiemeier, H et al., "Guideline adherence rates and interprofessional variation in a vignette study of depression," *Quality & Safety in Health Care*, 2002, *11* (3), 214–218.

Wickstrom, G and T. Bendix, "The "Hawthorne effect?" what did the original Hawthorne studies actually show?," *Scandinavian Journal of Work Environment & Health*, 2000, *26* (4), 363–367.

Table 1: Regression of the decentralization score on institutional characteristics

Variable	coef	std. err
The ability to hire and fire personnel (yes/no)	1.245	(0.030)***
The level at which salary decisions are made	.175	(0.007)***
Local control over financial decisions	.119	(0.019)***
The level at which staffing decisions are made	.121	(0.010)***
constant	-2.017	(0.017)
observations (facilities)	39	

*** indicates significance at the 1% level.

Table 2: Summary Statistics

	Direct Observation						Patient Exit Interview				
	% correct	decntl	facilities	doctors	cons	items	% correct	decntl	facilities	cons	items
Public sector	41%	0.04	20	52	830	13513	50%	0.04	3	104	1996
owner 2	56%	1.00	3	4	68	1137					
owner 3	58%	1.00	1	3	43	583	68%	1.00	1	42	662
owner 4	47%	0.81	9	4	63	1087	52%	1.00	1	26	416
owner 5	53%	1.00	2	1	37	245	75%	1.00	1	13	191
Private sector	18%	1.00	1	1	2	28	55%	1.00	4	73	1310
owner 7	39%	0.96	2	6	47	853	61%	0.96	1	24	400
owner 8	59%	0.68	2	3	10	172					

Table 3: The reaction to changes in scrutiny when the research team arrives

	Dep Var: whether doctor provides a specific input required by protocol as reported by the patient in an exit interview		
	(1)	(2)	(3)
Team is present at facility (0/1)	-0.065	-0.05	-0.05
	[0.077]	[0.077]	[0.077]
High scrutiny:		0.303	0.314
(whether the consultation is observed)		[0.114]***	[0.117]***
Institutional variables interacted with high scrutiny			
public sector	0.337		
	[0.113]***		
owner 3	-0.013		
	[0.196]		
owner 4	0.197		
	[0.169]		
owner 5	-0.398		
	[0.266]		
Private sector	0.231		
	[0.197]		
owner 7	0.18		
	[0.227]		
non-public		-0.234	
		[0.136]*	
decentralization score			-0.249
			[0.143]*
Institutional variables			
non-public		0.379	
		[0.085]***	
decentralization score			0.4
			[0.091]***
Constant	0.094	0.02	0.003
	[0.048]*	[0.074]	[0.077]

1,849 possible items observed over doctors at 9 facilities in which one doctor was observed by the research team and at least one doctor was not observed by the research team. Random effect probit regression on whether or not a given input was provided as measured by patient exit interviews. Standard errors in brackets, *, **, *** indicates significance at the 10%, 5% and 1% level. The data is restricted to facilities where observed doctors can be paired with unobserved doctors, and include data for four observations before the team arrives and four observations after the team arrives for each doctor.

Table 4: The reaction to changes in scrutiny as the research team remains

	Dep Var: whether the doctor			
	provides a specific input required by protocol			orders a lab test
	(1)	(2)	(3)	(4)
Level of additional scrutiny‡		0.015	0.015	
	[0.002]***	[0.002]***		
Institutional variables interacted with the level of additional scrutiny				
public sector	0.015			0.068
	[0.002]***			[0.028]**
owner 2	0.009			0.108
	[0.004]**			[0.020]***
owner 3	0.001			0.069
	[0.009]			[0.073]
owner 4	-0.007			-0.085
	[0.008]			[0.040]**
owner 5	0.05			0.267
	[0.027]*			[0.162]
owner 7	0.001			0.09
	[0.008]			[0.044]**
non-public		-0.012		
		[0.004]***		
decentralization score			-0.016	
			[0.006]**	
single-doc practice			0.009	
			[0.007]	
patient characteristics		Included		Included
DCO item fixed effects		Included		
illness characteristics				Included
Constant	-0.525	-0.566	-0.532	-1.019
	[0.158]***	[0.158]***	[0.160]***	[0.358]***
Observations	12,143	12,143	12,143	745

80 doctors observed over 12,143 diagnostic input observations (columns 1 through 3) and 745 consultations (column 4). Standard errors in brackets. *, **, *** indicate significance at the 10%, 5% and 1% level.
‡The level of scrutiny is the number of consultations since the research team arrived, times -1. All regressions are random effect probit regressions with a doctor random effect.

Table 5: Determinants of the correct diagnosis for a case study patient

	Dep Var: whether the doctor gives the correct diagnosis (0/1)
History taking inputs	0.037
(percent of inputs required by protocol)	[0.383]
Physical Examination Inputs	1.876
(percent of inputs required by protocol)	[0.404]***
Medical Officer	0.374
	[0.444]
Nurse	0.108
	[0.317]
Years of Training	-0.026
	[0.048]
Tenure (log of years)	-0.026
	[0.048]
Experience (log of years)	0.031
	[0.072]
Observations	598
number of unique clinicians	103

Standard errors in brackets. *** indicate significance at the 1% level. Random effect probit regression on whether or not the doctor was able to correctly diagnose a case study patient (vignette).

Table 6: The Hawthorne effect by history taking and physical examination inputs

	Dep Var: whether doctor provides a specific input required by protocol
Level of additional scrutiny‡	
history taking	0.012
	[0.003]***
physical examination	0.02
	[0.004]***
Decentralization score interacted with the level of additional scrutiny	
history taking	0.004
	[0.007]
physical examination	-0.04
	[0.008]***
Single-doc practice interacted with the level of additional scrutiny	
history taking	0.009
	[0.009]
physical examination	0.01
	[0.009]
patient characteristics	Included
DCO item fixed effects	Included
Constant	-0.296
	[0.167]*

80 doctors, 12,143 diagnostic input observations. Standard errors in brackets. *, **, *** indicate significance at the 10%, 5% and 1% level. ‡The level of scrutiny is the number of consultations since the research team arrived, times -1. The regression is a random effect probit regressions with a doctor random effect.