Getting doctors to do their best: Ability, Incentives and Practice*

Kenneth L. Leonard Dr. Melkiory C. Masatu Alex Vialou

September 30, 2004

Abstract

We examine two measures of medical quality collected among clinicians in Tanzania and show evidence that properly designed organizations can *cause* any kind of clinician to provider higher quality care. We use vignettes (unblind case studies with an actor) and direct clinician observation (in which clinicians are observed in the course of their normal consultations.) Vignettes measure ability and direct observation measures practice Clinicians who work for organizations with high powered incentives provide better quality care even after controlling for ability. Since practice, not ability, is what determines the outcome of health episodes, these organizations are able to consistently deliver higher quality care.

JEL Classification: I1, O1, O2
Keywords: Vignette, Technical Quality, Health Care, Outpatient Service, Tanzania, Health Care

*This work was funded by NSF Grant 00-95235 and The World Bank, and was completed 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 design of the vignettes benefited greatly from extensive discussion with Jishnu Das.

1 Introduction

We examine two measures of medical quality collected among clinicians in Tanzania and show evidence that properly designed organizations can *cause* any kind of clinician to provider higher quality care. We use vignettes or unblind case studies with an actor and direct clinician observation (DCO) in which clinicians are observed in the course of their normal consultations. We claim that vignettes are a measure of ability and that DCO can be used to measure practice. The quality of care provided by a clinician in actual practice will be limited by ability and medical ability varies widely. Some organizations, however, are able to consistently insure that clinicians in their employ practice at levels close to their ability. Since practice, not ability, is what determines the outcome of health episodes, these organizations are able to consistently deliver higher quality care.

By examining the difference between ability and practice and not just the level of practice we are able to differentiate between the view that some organizations provide high quality care because they hire or attract high quality doctors (a signaling or selection model) and the view that some organizations provide high quality care because they cause or force doctors to be good (an incentive model). Although organizations do differ in the type of doctor they hire (or attract), they also differ systematically in their ability to encourage any type of doctor to do better.

We find that organizational form is important to outcomes in Tanzania and suggest that this result is likely to apply in many other African countries with similar health systems. Africa faces some of the most serious and urgent health crises in the world. In an environment of high morbidity and mortality, widespread poverty and weak institutions, quality care is not assured even when health care is accessible and affordable. Our data show that some clinicians routinely misdiagnose and mistreat common illness, not because of lack of training or medicines, but because they do not provide the effort necessary to find the correct diagnosis. In our trials, we found that 33% of clinicians misdiagnosed a woman with pelvic inflammatory disease (PID) and 60% mistreated this condition. PID is caused

by untreated STDs and makes a woman more susceptible to HIV/AIDS and more likely to spread the illness to partners or children if untreated. While 86% of clinicians correctly diagnosed a patient suffering from classic symptoms of tuberculosis, 67% mistreated the disease. Furthermore less than one in five clinicians informed these TB patients of the importance of taking medicine or going to referral. Proper treatment of TB requires careful management by doctors; it cannot be treated by handing a patient medicine and sending them on their way.

There is increasing recognition that Africa's health problems are global in scale and that they cannot be solved without outside help, in particular large infusions of funds to help combat particular illnesses. However, money spent on HIV/AIDS, tuberculosis and malaria (target illness of the Global Fund[1]) will have limited impact if clinicians cannot identify the patients who suffer from the illness or inform them of their condition and the available treatments. We show that despite working in an environment of severely restricted resources, changes in the organization of health systems can improve the quality of care delivered by doctors at all levels of ability.

Throughout Africa, the majority of health services are delivered by government operated public health systems. In the rural areas of most African countries the private for–profit sector plays little or no role in health care delivery. However non-governmental organizations (NGOs), in particular, value based organizations such as missions and other religious organizations, play a significant role in rural health care. Comprehensive data on their role in Africa does not exist, but there is some selected evidence of the scale of their contribution. In Tanzania, half of all hospitals and hospital beds are provided by NGOs. In Ghana NGOs provide 40 percent of clinical care and in Zimbabwe they provide 35 percent of all hospital beds. Furthermore, NGOs tend to concentrate their services in the rural areas, whereas the government does not.[2] In Zimbabwe, 96 percent of all NGO facilities are located in the rural areas (for a discussion of the role of NGOs and the private sector in Africa in general see

[1] www.theglobalfund.org.

[2] Tanzanian government services are an exception to this rule.

Gilson et al., 1997; Leonard, 2002).

NGOs are commonly seen as providing higher quality care than government services and this is frequently attributed to "[t]he ability to hire and fire employees ... and to vire at least some funds within budgets ... advantages from which government counterparts often do not benefit."(Gilson et al., 1997, pp. 295) However, NGO quality is not always better and the organization of NGO facilities is not uniform. We explicitly measure the degree to which every organization can hire and fire employees (and at what level), the level at which salary decisions are made, the independence of the chief of post in making staffing decisions and the financial independence of the facility. In theory, these features should have an impact on the the degree to which clinicians exert effort on the behalf of the patient.

This paper takes advantage of the fact that we observe clinicians practicing in government services, three church–based health systems (Lutheran, Roman Catholic and Seventh Day Adventist), an Islamic hospital, a parastatal hospital and private practitioners who are either completely independent or part of a franchise network.[3] In addition, some of these clinicians are working in urban hospitals and others are working in small rural clinics. In these many facilities quality can vary significantly. Some government clinicians are very good and some NGO clinicians are not good. Rather than assume that NGOs represent a uniform approach to health care we look at variance in organizational form and compare that to the results these organizations can achieve.

Policy Implications Testing between an incentive model and a selection or signaling model is important for policy purposes. Under a selection model (some organizations know how to pick clinicians who will perform well under all circumstances) or a signaling model (a feature of some organizations makes employment desirable for good clinicians and they therefore seek employment in these organizations) the supply of good doctors and therefore good outcomes is fixed. Unless changes in policy change the supply of good doctors, the

[3]The health services of the Church of Gospel International (COGI) are organized as a franchise in that there is no health system governing the collection of facilities, only a series of facilities that are allowed to use the name COGI.

behavior of organization only alters where good doctors work, not the number of doctors. However, under an incentive model, all doctors can be encouraged to improve their practice.

The use of vignettes and direct clinician observation as quality measurement tools We use data from evaluation of process quality at health facilities in Arusha region of Tanzania collected in two phases at 40 facilities between 2001 and 2003. For each clinician observed we, used two different evaluation methods: direct clinician observation (DCO) and vignettes. In DCO, the clinician is observed in his actual practice and his use of history taking, physical examination and health education is measured by a doctor on the research team and compared to a predetermined checklist of appropriate procedures given the presenting symptoms. When a doctor scores highly on this test it indicates that he took care to properly diagnose the condition and educate the patient about the necessary steps. With vignettes, the doctor is examined by another doctor observing him diagnose an actor who has been trained to mimic a predetermined illness condition. A similar checklist is filled and the performance of the doctor is compared to what should have been done.

DCO has been used in other quality studies (Leonard, Mliga and Haile Mariam, 2002; Mliga, 2000) and vignettes have gained increasing popularity as a tool for quality evaluation both in developing and developed countries (Das and Hammer, 2004; De Geyndt, 1995; Epstein et al., 2001; Kalf et al., 1996; Koedoot et al., 2002; McLeod, Tamblyn, Gayton et al., 1997; Murata et al., 1992, 1994; O'Flaherty et al., 2002; Peabody et al., 1994, 1998, 2000; Tiemeier et al., 2002). Performance on these tests is a function of both ability (knowing how to diagnose and what to do to reach the correct diagnosis) and will (caring enough to want to properly diagnose the patient), however, Leonard and Masatu (2003) show that vignettes and DCO do not measure the same features of quality. Because vignettes are presented by the research team and take place in a short period of time, clinicians appear to see this procedure as a test of their ability. They want to do well on the test and are willing to provide the necessary effort. As a result, vignettes produce a measure that can be seen

as a maximum level of quality or ability. When clinicians are being observed in their actual practice, there is a period of time in which they alter their behavior because they know they are being observed, but they revert to a 'normal' behavior relatively quickly. This lower level of effort is what we call practice.

Identification Strategy We use data on the level of incentives faced by each clinician to explain the different performance of clinicians in different organizations after controlling for ability. If clinicians are randomly assigned between organizations, or if the only differences between clinicians are explained by ability, cadre or level of facility (clinic, or hospital for example) then this test is sufficient; differences in practice are *caused* by differences in incentives. However, it is likely that heterogeneity in clinicians extends to dimensions beyond ability. In particular, there may exist a good type of clinician, who chooses to practice to the best of his or her ability independent of any external pressure. If these good types select into certain organizations and these organizations just happen to have high powered incentives, our results may be spurious.

In order to control for this impact of additional heterogeneity in clinician type, we examine the performance of clinicians according to three categories of inputs: history taking, physical examination and health education. Each type of input plays an important role in the outcomes of health episodes, however, they are differentially impacted by incentives. Clinicians who have low incentives to work hard prefer history taking over physical examination. Clinicians with high incentives perform more physical examination (and marginally less history taking). However, no clinician faces any external incentives to perform health education; this is an element of health care that is not supervised, measured or rewarded. The differential willingness and incentives to provide different elements of health care allows us to distinguish between good type clinicians and clinicians who perform well because they are rewarded for doing so. Those who perform well on history taking, physical examination and health education are high quality in type, whereas those who perform well on history

taking and physical examination but not on health education are induced into high quality.

We employ three tests of the impact of non-ability clinician heterogeneity. First, we test the impact of incentives on the propensity of clinicians to provide health education. If we find that incentives increase the level of health education provided (after controlling for observables), then our measure of incentives is probably picking up good type clinicians, and our estimates are biased. Second, we explicitly control for clinicians who work in value–based organizations. If value–based organizations deliberately hire higher type clinicians controlling for this impact should eliminate this bias. Thirdly, we generate a measure of the willingness to provide health education and assume this measure reveals the type of a clinician. After controlling for type we look to see if incentives have any remaining impact. We find that value based organizations do not consistently hire higher type clinicians and that it is their incentives that lead clinicians to be better not their ability to attract the right kind of clinician.

In the following section we explain the process by which data were collected. We also outline the methodology for transforming the data (based on responses to a series of items) into a single score for each clinician. Some basic comparisons between the scores are outlined and we offer strong evidence that clinicians provide less effort than they could. Section 3 develops the methodology for two types of regressions in which we test for the impact of incentives on practice and control for clinician heterogeneity. Section 4 introduces a simple policy experiment in which we ask what would happen to outcomes if we were to alter the incentives at government facilities so that they looked more like parastatal facilities. Section 5 concludes.

2 Data and Instruments

Data collection took place over a period of two years from October of 2001 to March of 2003. Forty health facilities in the rural and urban areas of Arusha region were visited at least

two times each. The full sample includes 100 practitioners; we were able to evaluate quality using both the DCO and vignette instruments in 80 clinicians.[4] Both vignettes and DCO were administered by local medical personnel (see Leonard and Masatu, 2003, for further details of the study).

2.1 Vignettes

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 clinician that he must do physical examination by posing questions. The patient then answers the question verbally. For instance, if the clinician 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. We create scores such as the number of expected items correct and differentiated these scores by history taking, physical examination or health education. Each clinician faced at least 6 vignettes identified as the malaria, PID, diarrhea, pneumonia and flu and worm infestation vignettes (an at-risk pregnancy and TB vignette were added in round 3). For the purposes of this paper we use only the malaria, diarrhea and pneumonia vignettes because they correspond well to categories in the DCO evaluation as discussed below.

Since the vignette is deliberately designed by the research team, the diagnosis given by the clinician can be compared to the correct diagnosis. There are four possible outcomes and they can be ordered; wrong, incomplete, extra and correct. Table 1 shows the results of a series of ordered probit regressions of the outcome of diagnosis on the input levels the

[4]For some clinicians we observed no patients and were therefore unable to obtain a DCO score.

variables and other characteristics of the clinician.[5] The table shows that the only variables that matter for predicting the outcome of diagnosis are the levels of inputs provided. Good doctors get the correct diagnosis because they provide the necessary level of inputs, not because they know how to diagnosis with less effort.

Physical examination and health education are important,[6] however, history taking has a significant impact on outcomes only in the case where it stands alone. In this view, history taking may only be important in as much as it feeds into physical examination; it may allow the clinician to decide what physical examination he should perform. Importantly, substitution away from physical examination into history taking will decrease the probability of a correct diagnosis.

2.2 Direct Clinician Observation (DCO)

With DCO, a member of the research team (a clinician) sits in on the regular consultations at a facility. For each consultation the observer uses a checklist of items that are expected. Defining what clinicians should have done, or what is rational, is not easy and therefore we have used physician observation checklists that identify 4 different categories of illness (fever, cough, diarrhea, and symptoms indicative of STDs). For each of these categories there is a list of expected history taking questions and well as expected physical examination procedures.

This procedure has a number of advantages, but the most important one is direct observation of the patients who are there. Even if clinicians alter their behavior when they are observed, if we examined the clinicians for repeated observations there is a strong possibility that behavior will fall towards a more accurate representation of the clinicians true behavior. The clinician becomes tired of faking it. We examine our data from the DCO for just such a trend. Figure 1 shows a non-parametric regression fit to the data for each score. The score

[5]We tested the results with the ordering wrong, extra, incomplete, correct and found little difference in the overall importance of inputs. In addition, we ran the regression as a random effect probit model on whether or not the diagnosis was correct and found the same results.

[6]It is legitimate to wonder how health education can lead to the correct diagnosis, since health education comes after the diagnosis is given. Note that the clinician does not know if his diagnosis is correct when he offers health education. If health education is dropped from this regression, we obtain the same basic results.

of each clinician is compared to the difference from his score for the first observation so as to compare clinicians. The graph shows a clear downward trend in each of the elements of diagnosis with an apparent taper after 20 observations. It seems that clinicians as a whole provide less effort the longer they are observed. Clearly each clinician is capable of providing effort similar to the effort they provide for the first few observations; clinicians are shirking.

2.3 The relationship between vignettes and consultation

Although vignettes should represent an upper bound on the performance of clinicians in actual practice, it is not clear that there should be any correspondence between what we observe in vignettes and DCO. The connection between these two instruments is illustrated by examining the set of items that are exactly the same on both instruments. For example, for the fever vignette, one of the items is "does the clinician take the patient's temperature?" The same question is present on the list of items for actual patients who present with fever. When we look at the actions of clinicians on only this item we find that, out of 329 observations, in 52% of DCO observations, the clinician checks the temperature on both the vignette and the DCO.[7] However, 35% of the time clinicians do not check the temperature in the consultation though they did in the vignette. We take this to broadly indicate the set of clinicians who knew what they were supposed to do in this circumstance, but did not. 8% of observations represent clinicians who checked the temperature in neither the vignette or DCO and 5% represent clinicians who did not check the temperature in the vignette but did check it in DCO. This last category — clinicians who did not know what they were supposed to do, but who did so anyway — is a rough measure of noise or imprecision in our instrument.

Table 2 shows the results of this comparison for all 26 possible matches in questions, in total and by history taking and physical examination. There are no matches in health education. For all possible matches 26% of the time clinicians who knew what to do did

[7]Each clinician only takes the vignette once, but may be observed in DCO multiple times. It is more precise to say "52% of observations represent a clinician who checked the temperature in the vignette and checked the temperature in DCO."

so in practice. 23% of the time clinicians who knew what to do did not do so in practice. Knowing what to do only leads to a 53% chance of doing the right thing in practice.

Figure 2 shows the relationship between vignettes and DCO as the number of consultations increase. Figure 1 showed evidence of shirking as the number of consultations increased. Here we examine the number of times that a clinician who knew what to do actually did so in practice as a function of the number of observations. Both history taking and physical examination show evidence of a drop off, although the trend is not strong for history taking. The trend for physical examination shows that about clinicians who know what to do will do so about 26% of the time when they are first being observed, but only 18% of the time as they revert to their normal practice.

The fact that only 53% of the clinicians who know what to do actually do so suggests that there will be at best a weak correlation between vignette and DCO scores. The correlation should be stronger when we control for the response of clinicians to the passage of time. In particular, a DCO score that measures how clinicians behave on the first few observations should bear some resemblance to how they behave on vignettes.

2.4 Item response scoring

In order to compare ability and practice we create a single score for each clinician derived from both the vignette and DCO. Although the vignette is deliberately designed so that every question is important in reaching the correct diagnosis, in practice some questions will be more important than others. If we knew, *a priori* which clinicians were good and which were bad, an empirically useful item would be one that was raised by all good clinicians and not raised by any bad clinicians. A less useful item would be one that was raised by everyone. Although both items are important to diagnosis, they have a differential ability to aid the researcher in distinguishing between good and bad clinicians. In addition, there is likely to be some variance associated with each item. If we were to repeat the test again and again, it is possible that the same clinician might raise an item for some tests but not for others. In

creating an aggregate score for each clinician we seek, not only to reduce the dimensionality of the data, but also to reduce the noise in the instrument and to weigh questions according to their importance. Item response theory (Birnbaum, 1967; **?**) (IRT) is an attempt to achieve all of these goals. Das and Hammer (2004) use IRT for the analysis of vignettes. Following their methodology we use a less general two-parameter method, outlined below.

There are J items on each 'test' $j = 1 \dots J$. The response (x) for each item is either correct or wrong, indexed 1 or 0; $x_j \in \{0, 1\}$. The result of the test can be characterized as $\vec{x}$ a $J \times 1$ response vector. We define a rule s such that $s(\vec{x}) : \Re^J \to \Re$. s maps J responses to one result. θ is the single index value assigned as a result to each test. Using this result we can define, for each item j, $P_j : \Re \to [0, 1]$ such that $\Pr(x_j = 1|\theta) = P_j(\theta)$. P_j maps the latent continuous variable θ into the probability of answering the item j correctly.

We use the following rule to derive θ:

$$\frac{P_j(\theta)}{1 - P_j(\theta)} = \exp(\alpha_j \theta + \beta_j)$$
$$\log\left(\frac{P_j(\theta)}{1 - P_j(\theta)}\right) = \alpha_j \theta + \beta_j$$

This is simply a logistic regression of the discrete value x_j on the underlying latent value θ with a slope and intercept term that can vary by item. The intuition of the parameters is straightforward. When α is positive, θ increases the probability of getting an item correct. α is a measure of the importance of the underlying 'ability' for a given item and β is a measure of how likely someone is to get an item correct even if they have a low θ. When α is low or our estimate of α is insignificant, the item is not useful in distinguishing 'better' test takers from 'worse' test takers. When it is high, the question is more useful. When β is large, even 'worse' test takers are likely to answer the question correctly. When β is low, even 'better' test takers are unlikely to answer the question correctly. Choosing a simple rule such as $\theta = \sum_j x_j$ is the same as setting $\alpha_j = \alpha_k$ and $\beta_j = \beta_k \; \forall j \neq k$.

The IRT score for vignettes (θ_v) is based on 44 items over the three vignettes used. A

few clinicians were observed more than once and we use their responses on both sets of vignettes but solve for only one θ_v. The IRT score for DCO (θ_c) is based on 42 questions over three presenting conditions. Clinicians were observed many times each and we solve for only one score per clinician. In addition to the rule above we allow for a linear dropoff in practice that can vary between history taking items, physical examination items and health education items. Since some clinicians were observed for a long period of time, and other were observed for only a short period of time, controlling for observation number allows us to avoid assigning high quality to a clinician only because we did not observe long enough to see him revert to his true practice.

Table 3 shows the correlations between four possible scores, the raw vignette and DCO scores based on percentage of items correct ($\bar{x}_v$ and $\bar{x}_c$) and the IRT scores (θ_v and θ_c). The raw scores are not significantly correlated with each other, however the IRT scores are correlated at the 10% level.

2.5 Incentives

The comparison between the two evaluation methods makes it clear that shirking is a real phenomenon. The question is not whether or not it exists, but whether or not we can advance a descriptive understanding of how shirking is likely to be altered by various factors that are within the grasp of a policy maker. In addition, if we can understand the degree of shirking we can try to estimate real quality at a facility after observing the "test-altered" level of effort. For incentives we use a set of measures developed for use in this region in previous work Mliga (2000). The incentive variables suffer from multicollinearity both because there are strong correlations between organizations and because there is no variation in these variables at the facility level (and very little within an organization). Table 5 describes the incentives facing every clinician in the data by showing how incentives vary by the owner and level of facility.

The four variables that we examine may have differential impacts on how clinicians

behave. FIRE measures the ability of the chief of post or the superior to hire and fire personnel. SALARY measures the degree to which supervisors can set salaries. STAFF D measures the degree to which supervisors can choose the type and number of clinicians who work for them. FIN IND measures the degree of financial independence. We could expect at least two forces to be in play here. On the one hand, those who run and supervise health facilities need the power to reward and punish those who work for them. They might use FIRE, SALARY or STAFF D to do this. On the other hand, incentives can be driven by the exposure of a facility to the laws of demand. This type of influence is most likely to be measured by something like FIN IND. However, collinearity in our data on these measures does not allow us to compare how these factors independently impact the practice of clinicians. Nor does it make much sense to try to isolate these impacts. If we observe a facility with power but without exposure, what is the point of the power? If we observe a facility with exposure but without power, what role can exposure play?

Rather than using each measure as a independent variable, we construct, a single index representative of the level of the incentives at any facility. We employ consequently a factorial analysis to construct a scoring factor.[8] Using factorial analysis we derive a single index of incentives that we normalize to values between 0 (lowest observed incentives) and 1 (highest observed incentives).

When we consider the behavior of clinicians who work in organizations with high incentives we will not be able to differentiate between clinicians who work hard because they are supervised by someone who wants them to work hard, and clinicians who work hard because their income depends on fees collected from patients. An important feature of these incentives, for the purpose of this work is that neither type of pressure appears to leads to clinicians who do health education.

Organizational Quality and Outcome Quality The mechanism of monitoring clinicians in health care is not based on outcomes. In outpatient services the outcome is rarely

[8]See Section A.1 in the appendix for the derivation and explanation of this procedure.

learned before a patient leaves a health facility and it is prohibitively expensive to find patients after they have returned home. Instead, the regulation of clinicians is more likely to take place through the the monitoring of organizational quality Leonard (2002, 2003). Organizational quality is a measure of the quality of inputs that are provided in health care not the outcomes. In practice, every patient generates a potential record when they visit a clinic or hospital. The record will have information on vital statistics, tests ordered, results of tests, prescriptions, etc. If a patient is referred, these records are should travel with them. When a clinic is visited these records can be examined and a relatively accurate picture of quality can be created. Organizational quality is only useful to patients if it causes clinicians to provide outcome quality. Outcome quality is a measure of the set of activities that improve the health of the patient. Organizational quality is used to monitor clinicians because it does increase outcome quality. However there are aspects of outcome quality that are not part of organizational quality. Health education is one of these elements. Doctors will write down the diagnosis and the prescription. They do not write down how they explained the diagnosis to their patients. They do not record whether or not they explained the origins of the illness to patients. And they do not explain whether or not they told the patient how to avoid the illness in the future.

This is not to say that modern medicine does not believe health education is important, but rather that they have either not found a way to encourage its provision, or they have not chosen to implement such such a mechanism. More importantly, we find no evidence that patients value health education. Those clinicians who are careful to provide health education are no more likely to be highly rated by patients than those clinicians who provide no health education.

2.6 Summary statistics of scores and incentives

Table 5 shows the basic relationship between our measures of ability and practice, the incentives facing various health care providers and provider characteristics such as cadre, level

of facility, etc. Table 6 outlines the relationship of these variables to broader categories of facilities. Clinicians who work for value–based organizations do not have better ability than other clinicians, however their practice is better. Clinicians who work in organizations with high incentives have higher practice and ability than other clinicians. Furthermore, the difference between ability and practice is also higher. Organizations with medium incentives seem to do better still, but this does not control for any other factors (cadre, level etc). Hospitals and health centers have higher ability than dispensaries but worse practice. Ability follows cadre exactly as we might expect except that doctors are not significantly better than officers. Officers however, are much worse than doctors in practice. Nurses, who have the lowest ability, manage to practice at reasonable levels by practicing close to their ability.

3 Analysis

Figure 1 and Figure 2 show clear evidence of shirking in clinician behavior. We observe ability in at least two forms: scores on the vignette and scores for the first few observations of actual consultations. Furthermore, Table 5 shows a strong relationship between incentives and the change from vignette scores to DCO scores. In this section we develop a more explicit test of this relationship.

3.1 Incentives to provide effort

We begin with a general concept of practice (P) and ability (A) and incentives (I). Practice at any moment in time is a function of ability, incentives, time and an idiosyncratic element.

$$P_{it} = \beta + \alpha \cdot A_i + \delta \cdot I_i + \gamma \cdot A_i \cdot I_i + \kappa \cdot I_i \cdot t + \epsilon_{it} \tag{1}$$

Incentives could change the ability of any practitioner (δ), change the rate at which ability is transformed into practice (γ) or change the degree to which issues of time play a role (κ).

As we have seen evidence of shirking both in the gap between ability and practice and

in the dropoff of practice over time, we will test for the impact of incentives in each of these areas. We will examine two different specifications for this test:

$$Pr_{ijt}(x_j = 1) = \beta_j + \alpha_j \cdot \theta_{v,i} \quad + \delta \cdot I_i + \gamma \cdot I_i \cdot \theta_{v,i} \quad + \kappa_j \cdot t + \epsilon_{ijt} \tag{2}$$

$$Pr_{ijt}(x_j = 1) = \beta_j + \alpha_j \cdot \theta_{c,i} \qquad\qquad\qquad\qquad + \kappa_j \cdot t + \hat{\kappa}_j \cdot I_i \cdot t + \epsilon_{ijt} \tag{3}$$

Incentives impact the gap between ability and practice Equation 2 models the probability of getting an item correct on DCO as a function of the item itself (α_j and β_j), the IRT score on the vignette for clinician i, ($\theta_{v,i}$), an intercept term for incentives ($\delta \cdot I_i$), the IRT score interacted with incentives ($\gamma \cdot \theta_{v,i} \cdot I_i$) and an item specific response to the order of observation ($\kappa \cdot t$). In this specification we expect the vignette score to be an important determinant of practice and test for the role of incentives through both δ and γ.

Incentives impact the rate at which clinicians alter their behavior in DCO Equation 3 models the probability of getting an item correct on DCO as a function of an endogenously determined clinician score (θ_v) item specific slope and intercepts (α_j and β_j) a item type specific dropoff ($\kappa_j \cdot t$) and an item type specific dropoff interacted with incentives ($\hat{\kappa}_j \cdot I_i \cdot t$). We do not calculate a separate dropoff for each item (as above) but rather a dropoff by history taking, physical examination and health education categories. We expect a significant dropoff for each of the three types of items ($\kappa_j < 0$) and a positive interaction term of each dropoff with incentives ($\hat{\kappa}_j > 0$). This would suggest that on average clinicians provide less and less effort as time passes, but that clinicians who work with high incentives either do not provide less effort ($\hat{\kappa}_j + \kappa_j > 0 \rightarrow \hat{\kappa}_j > -\kappa_j$), or that they dropoff less than do other clinicians ($0 < \hat{\kappa}_j < -\kappa_j$).

The sign of the dropoff for health education and this dropoff interacted with incentives allows us to test whether incentives increase health education scores. If $\hat{\kappa}_{j \in he} > 0$ this suggests that incentives are correlated with clinicians who do well on health education. If incentives are driving behavior (rather than being correlated with high type clinicians) then

we should find $\hat{\kappa}_{j\in he} = 0$.

3.2 Selection Bias

In addition to testing the two models of practice listed above, we investigate the possibility of selection bias caused by unobserved heterogeneity. It is likely that there is an inherent trait of some clinicians which causes them to be of good type. These clinicians care for their patients under any circumstance. They may or may not have greater ability, but their nature will lead them to practice at levels that are close to their ability. If such clinicians are randomly distributed in the health care system it will lead at worst to attenuation bias. However, if good type clinicians are more frequently found working for NGOs or value-based organizations (and these organizations use high powered incentives), our coefficients for incentives will be upwardly biased and any finding of positive effects from incentives may be spurious. NGOs may be able to find such clinicians in the hiring process, or such clinicians may find it preferable to work for NGOs and value–based organizations and self-select into employment.

We use two measures of clinician type to try to control for the impact of heterogeneity. First, we create a measure of how much a clinician cares for the patient by looking at whether or not clinicians provide above average levels of health education. To do this we first run the regression:

$$Pr_{ijt}(x_j = 1) = \beta_j + \alpha_j \cdot \theta_{v,i} + \kappa_j \cdot t + \epsilon_{ijt} \tag{4}$$

Then we create a score for each clinician based on

$$\text{CARES}_i = \frac{\sum_{it,j\in he} x_j - Pr_{ijt}(x_j = 1)}{N_{it,j\in he}} \tag{5}$$

where $N_{itj\in he}$ is the number of items observed in health education for each clinician. Clinicians with a high score for this variable are those who provide above average levels of health education. CARES is a measure of the frequency with which clinicians provide health

education above the predicted amount of health education.

Second, we use a dummy variable for whether or not a clinician works for a value–based organization. Incentives vary among value–based organizations and therefore we can test for the impact of both a value–based organization and the level of incentives.

Each of these two measures are tested as additional measures of incentives in the model of Equation 2. The values of VALUE–BASED and CARES are recorded in Table 7. Clinicians who care are more likely to be found in value–based organizations and less likely to work for organizations with low incentives. However there is clear variation in the distribution of these types of clinicians. The Lutheran and Roman Catholic services are both large religious presences in this area and yet they attract (or find) such clinicians at different rates. The Seventh Day Adventist church (SDA) is world renowned for its health care philosophy which places heavy emphasis on preventive habits and health education. Yet this organization is significantly less attractive to high type clinicians than the Church of Gospel International, a denomination with no known health care presence.

3.3 Results

Table 8 shows the results of three probit regressions based on the specification of the test in Equation 2. Regression A represents the simple test of incentives on the relationship between ability and practice. Regression B includes the variable CARES and regression C includes the dummy variable VALUE-BASED. In all three regressions the impact of ability (α) is positive and significant, and the change in practice over time (κ) is negative and significant. For regression A, incentives increase practice (δ) and the rate at which ability is translated into practice (γ). For regression B and C both the intercept and the slope effect are positive. However in regression B but only the intercept is significantly positive whereas for regression C the opposite is true.

CARES has a very significant impact on practice. Doctors who care do better than doctors who do not. However, even after taking this effect into account, incentives still improve the

quality of practice. Value–based organizations do not appear to be differentially good at attracting good doctors. They are better than non-value based organizations, but this is because they use higher incentives.

Table 9 shows the results of the probit regression based on Equation 3. The rates of dropoff for the average clinician are all negative and significant; practice declines with time. High powered incentives restore this loss, but only for physical examination. In fact the results are such that clinicians who work for the organizations with the highest level of incentives ($I_i = 1$) increase the quality of their practice over time (as measured by physical examination). However, these same clinicians do worse than the average clinician on history taking and much worse than the average clinician on health education. Incentives emphasize the use of physical examination over the use of history taking and health education.

These results are consistent with clinicians who find it easy to add history taking items when they are first being observed. After time the history taking drops off because they are reverting to their normal practice. However those clinicians who are constrained by incentives continue to provide the all–important physical examination inputs.

As a whole, these results suggest that clinicians change their practice in the face of high–powered incentives. There is heterogeneity in clinician type that is not measured by ability and clinicians who are of the good type are much better than clinicians who are not. However, all clinicians can be made better by high–powered incentives.

4 Policy Implications

The previous section exposed the role of incentives in making any type of clinicians a better practitioner. Thus it is particularly interesting to compare how two clinicians from the same cadre, working in a same level facility but where the incentive structure varies will deliver health care. Our analysis has measured practice in terms of scores, however to examine the policy implications we translate these scores into outcomes.

Practice is better described with the DCO data, however, information on the diagnosis can be obtained from the vignette instrument. It would be useful to be able to infer the probability of getting the correct diagnosis using the IRT scores from the DCO observations. As a first step we need to compare the distribution of each score. If we can translate a given DCO score into a vignette score we can infer the probability of getting the correct diagnosis from the DCO score.

Table 5 displays the θ_v and θ_c scores for each cadre within each facility. Both sets of scores are constrained to range between -1 and 1.[9] However, a score of 1 on the vignette is not necessarily equivalent to a score of 1 on the DCO. In order to compare these two ranges of scores we return to the results underlying Table 2. Examining only those items on the two tests that are exactly the same, we found that, on average, clinicians answered 49% of vignette items correctly and 40% of DCO items correctly. Thus, for all observations and for the items for which we can directly compare the vignette and the DCO, the DCO response rate is 81% of the vignette response rate. If we restrict ourselves to observations that are greater than 20 (the approximate point at which the dropoff appears to level in Figure 1) this ratio becomes 60%. Thus we hypothesize that the highest possible score on the DCO is approximately equivalent to a score of 0.6 on the vignette. Other possible scores on the DCO are translated into vignette scores based on this initial translation. This is a very approximate method for translating scores so we also follow out policy experiment assuming that a score of 1 on the DCO is equivalent to a score of 1 on the vignette. This equivalent score is labeled $\theta_c^\star$ Figure 3 shows a kernel density distribution for θ_v and $\theta_c^\star$.

From among the clinicians represented in Table 5 we identify two clinicians very similar in their ability measure (θ_v). One doctor works at a governmental hospital and the other one at a parastatal hospital. The parastatal hospital is governmental–owned but does not receive funds from the government. To stay in business, the parastatal can seek profits and its manager has the ability to hire or fire staff. The parastatal hospital is a high incentive facility,

[9]Mean and variance of the IRT score are unidentifiable, so the scale is arbitrarily set.

while the governmental–run hospital has low incentives. Table 10 shows that each of these two clinicians has roughly the same ability, but their practice scores diverge significantly. Figure 3 shows the ability of both the government and parastatal doctors at the same level and then shows how the ability of each practitioner drops off. Importantly the practice of the government doctor is significantly worse than the practice of the doctor in the parastatal hospital. At the parastatal hospital the doctor's practice score is high, and this should lead to a higher probability of getting the correct diagnosis. If we translate DCO and vignette scores 1 for 1, this clinician is likely to diagnose diarrhea in an infant correctly 80% of the time. Using our hypothesized shift (DCO scores are 60vignette scores), we expect he will diagnose correctly 69% of the time. In comparison the clinician at the government facility will diagnose the patient correctly only 65 or 51 percent of the time.

Policy experiment Parastatals and government owned health facilities are both owned by the government, but fall under very different organizational schemes. It is conceivable that government services could be reorganized so that they looked more like parastatal services. This might involve very large changes in the fees charged at both institutions and these changes could lead to large changes in welfare. However, in the context of this experiment, we only ask if a change in incentives would lead to an improvement in outcomes.

If we increase the incentives faced by the government clinician so that they are the same as those faced by the parastatal clinician, we find that his ability to diagnose increases by between 11 and 17 percent for a complicated malaria case, between 19 and 23 percent for a pneumonia case and between 19 and 23 percent for a diarrhea case. These changes will lead to very large changes in outcomes.

5 Conclusion

Clinicians do not universally practice to the best of their abilities, however, organizations can use incentives to increase the quality of care delivered by clinicians. Comparing the scores

of clinicians on vignettes and direct clinician observation (DCO) we find that clinicians who work in organizations that have high powered incentives achieve higher quality practice after controlling for ability, cadre and the level of facility. In addition, we find that clinicians who work for organizations with high powered incentives maintain the same level of quality in physical examination as time passes, whereas other clinicians experience a significant dropoff in their level of practice. Incentives have the same impact across these two data sets.

We do not reject the possibility that there are at least two types of clinicians: there are clinicians who provide high quality care only because they face strong incentives to do so and there are clinicians who would provide high quality care under any circumstance. Furthermore, some organizations appear to be able to attract these clinicians. However, even when we control for the possibility of good-type clinicians we find that increased incentives increases the likelihood that a clinician will provide high quality care.

These results provide important impetus to policy reform in health care services in Africa; they suggest that the quality of care can be improved among with the existing stock of medical personnel. Our simple policy experiment suggested that the rate of correct diagnosis could be significantly improved if clinicians were induced to work harder.

References

Online documents are available at `www.arec.umd.edu/~kleonard/research.html`

Bennett, S., B. McPake, and A. Mills, eds, *Private Health Providers in Developing Countries: Serving the Public Interest?*, London and New Jersey: Zed Books, 1997.

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.

Das, Jishnu and Jeffrey Hammer, "Which Doctor?: Combining Vignettes and Item-Response to Measure Doctor Quality," mimeo, The World Bank 2004.

De Geyndt, "Managing the Quality of Health Care in Developing Countries," World Bank Technical Paper 258, The World Bank, Washington, D.C. 1995.

Epstein, SA et al., "Are psychiatrists' characteristics related to how they care for depression in the medically ill? Results from a national case-vignette survey," *Psychosomatics*, 2001, *42* (6), 482–489.

Gilson, L. et al., "Should African Governments contract out clinical health services to church providers?" In Bennett, McPake and Mills, eds (1997) chapter 17.

Kalf, Annette JH et al., "Variation in diagnoses: Influence of specialists' training on selecting and ranking relevant information in geriatric case vignettes," *Social Science and Medicine*, 1996, *42* (5), 705–712.

Koedoot, CG et al., "Palliative chemotherapy or watchful waiting? A vignettes study among oncologists," *Journal of Clinical Oncology*, 2002, *20* (17), 3658–3664.

Leonard, Kenneth L., "When States and Markets Fail: Asymmetric Information and the Role of NGOs in African Health Care," *International Review of Law and Economics*, 2002, *22* (1), 61–80.

_ , "African Traditional Healers and Outcome–Contingent Contracts in Health Care," *Journal of Development Economics*, 2003, *71* (1), 1–22.

_ **and Melkiory C. Masatu**, "Comparing vignettes and direct clinician observation in a developing country context," mimeo, under revise and resubmit at *Social Science & Medicine*, Columbia University 2003.

_ **, Gilbert Mliga, and Damen Haile Mariam**, "Bypassing Health Facilities in Tanzania: Revealed Preferences for Observable and Unobservable Quality," *Journal of African Economies*, 2002, *11* (4), 441–471.

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: Macmillan, 2000, chapter 8. Available at http://repositories.cdlib.org/uciaspubs/editedvolumes/5/.

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

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

O'Flaherty, M. et al., "Low agreement for assessing the risk of postoperative deep venous thrombosis when deciding prophylaxis strategies: a study using clinical vignettes," *BMC Health Services Research*, 2002, *2* (16).

Peabody, John W. et al., "Quality of care in public and private primary health care facilities: structural comparisons in Jamaica.," *Bull Pan Am Health Organ*, 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 Amer-*

ican Medical Association, 2000, *283*, 1715–1722.

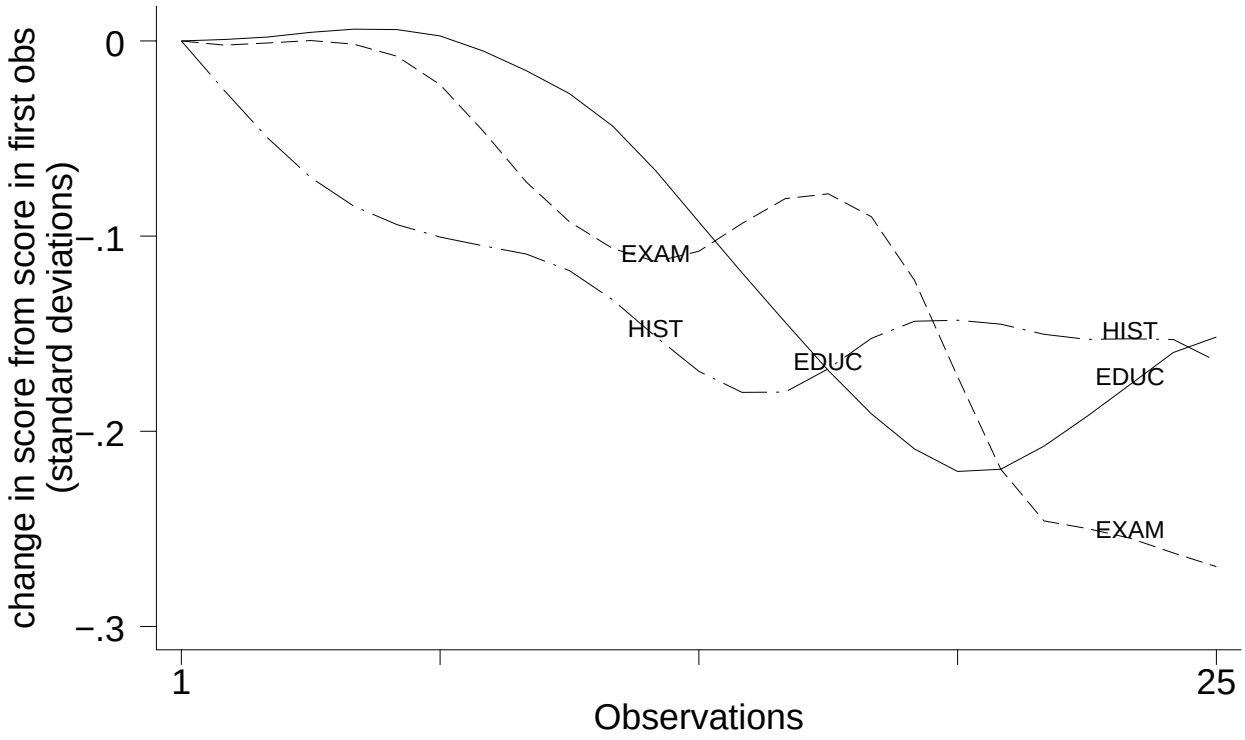
Tiemeier, H et al., "Guideline adherence rates and interprofessional variation in a vignette study of depression," *Quality & Safety in Health Care*, 2002, *11* (3), 214–218.

Table 1: Ordered Probit Models of Diagnosis Outcome

Regression	A		B		C		D		E	
variable	coef	stderr	coef	stderr	coef	stderr	coef	stderr	coef	stderr
EDUC	0.214	(0.076)**	0.210	(0.076)**	0.232	(0.078)**				
EXAM	0.304	(0.075)**	0.317	(0.078)**	0.346	(0.080)**	0.384	(0.078)**		
HIST	0.065	(0.070)	0.074	(0.072)	0.091	(0.074)	0.115	(0.073)	0.153	(0.073)**
Control for vignette number		NO		NO		YES		YES		YES
Control for cadre category		NO		YES		YES		YES		YES
obs		282		282		282		282		282
likelihood		-319.89		-319.49		-295.90		-300.44		-312.75

Ordered Probit of diagnosis outcome (wrong, incomplete, extra, correct) . Cut-off levels not reported. Model A, B and C are for vignettes 1, 3 and 4. F-test that all skill variables are jointly equal to zero fails to reject. Likelihood ratio test suggest that the cadre variable are not significantly different from zero as a group. Likewise, tenure and Experience are never significant. On the contrary, the vignette variable are significantly different from zero.

Figure 1: Change in average consultation score over number of patients seen for (DCO)



The smoothed lines shown represent the change in the number of history taking, physical examination and health education items answered correctly compared to the first observation. The average score is declining with the number of observations. The average number of items correct is normalized within each type of symptom so that scores can be aggregated. Points are derived from a kernel density estimation with a quartic kernal, bin width of 6 observations evaluated at each number of observations between 1 and 25. Observations over 25 are set equal to 25.

Table 2: Comparison of responses to items that appear on both the DCO and Vignettes

vign/DCO pair	obs	yes/yes	no/no	yes/no	no/yes
	#	%	%	%	%
total	6844	26	37	23	14
history taking	3887	28	33	24	15
physical examination	2957	24	41	22	13

There are 26 items that appear on both the vignette and DCO instruments. 16 are history taking items and 10 are physical examination items. The number of observations is much larger than the number of vignettes administered because there are multiple questions and multiple observations of the DCO instrument. We can group these responses into four mutually exclusive categories: correct on vignette and DCO, incorrect on both, correct on vignette and incorrect on DCO and incorrect on vignette and correct on DCO. The percentages of all responses that fall into each of these four categories are reported in this table.

Figure 2: Change in the ratio of items correct on DCO to items correct on vignette over the number of patients seen

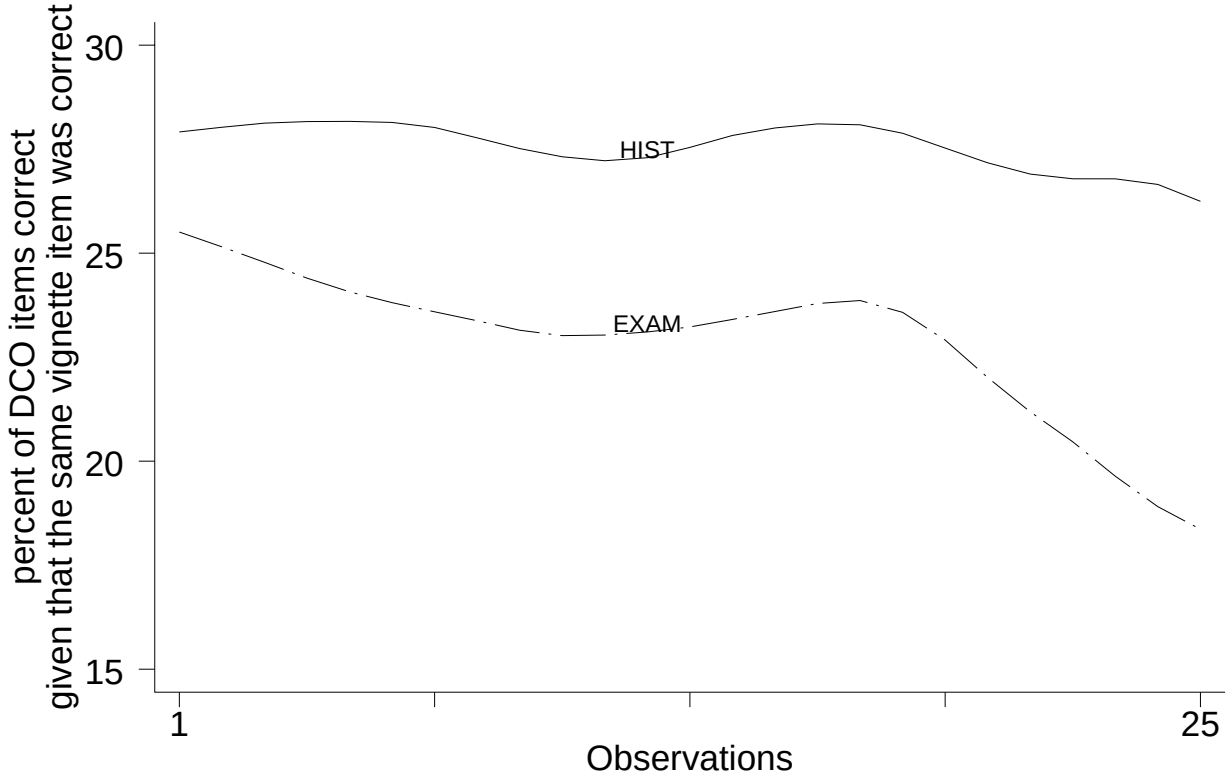


Table 3: Correlations between raw scores ($\bar{x}$) and IRT scores (θ)

score	$\bar{x}_v$	$\bar{x}_c$	θ_v
$\bar{x}_c$	0.16		
θ_v	0.74**	0.23**	
θ_c	0.34**	0.76**	0.21*

80 observations for each pair
** indicates significance at the 5 percent level
* indicates significance at the 10 percent level

Table 4: Organizational Variables reflecting incentives by owner and level of facility

		observations		Staff control[a]				
Owner	Level	facilities	clinicians	FIRE	SALARY	FIN IND	STAFF D	LABEL
Govt	Disp	15	20	no	national	low	national	govt
Govt	HCenter	4	8	no	national	low	national	govt
Govt	Hospital	3	9	no	national	low	national	govt
Luth	Disp	4	6	yes	regional	medium	regional	ngo
Luth	Hospital	1	5	yes	regional	high	local	ngo
RC	Disp	2	2	yes	local	high	regional	ngo
RC	Hospital	1	3	yes	local	high	loc/reg	ngo
SDA	Disp	3	4	yes	national	high	regional	ngo
COGI	Disp	3	3	yes	local	high	local	priv
Parast.	Hospital	1	2	yes	local	high	local	priv
Islamic	Hospital	1	2	yes	local	high	local	priv
Private	Disp	1	1	yes	local	high	local	priv

Govt: government; Luth: Lutheran; RC: Roman Catholic; SDA: Seventh Day Adventist; COGI: Church of Gospel International; Parastatal: Parastatal run by the Arusha International Conference Centre; Islamic: Ithna Asheri Mosque.

a: Variables derived from (Mliga, 2000, pp. 213).

FIRE: Can the head of this facility hire and fire personnel?

SALARY: Level at which salary decisions are made. National = 1; regional = 2; local = 4.

FIN IND: The ability of a facility to pay salaries and buy essential medical supplies to run the facility: national = 1; regional = 2; local = 4.

STAFF D: Location at which medical staffing decisions occur (for example composition of staff): national = 1; regional = 2; local & regional = 3; local = 4.

LABEL: The overall label that would be applied to this facility; government, NGO or private.

Table 5: Average IRT scores for vignettes (θ_v) and DCO (θ_c) by owner

owner	clinicians	value-based?	incentives[a]	level	cadre	θ_v[b]	θ_c[c]	difference[d]
COGI	1	NO	1.00	Disp	assist	0.20	0.38	0.24
	1	NO	1.00	Disp	officer	0.33	0.09	0.10
	2	NO	1.00	Disp	doctor	0.03	0.56	1.64
	4	NO	1.00	average	average	0.16	0.40	0.90
Govt	6	NO	0.00	Disp	nurse	-0.37	-0.54	0.23
	11	NO	0.00	Disp	assist	-0.26	-0.33	0.32
	11	NO	0.00	Disp	officer	0.16	-0.20	-0.13
	1	NO	0.00	Disp	doctor	0.37	0.11	0.05
	1	NO	0.00	HCenter	nurse	0.06	-0.34	-0.22
	3	NO	0.00	HCenter	assist	0.04	0.02	0.54
	7	NO	0.00	HCenter	officer	0.37	-0.45	-1.31
	2	NO	0.00	HCenter	doctor	-0.23	-0.10	0.85
	1	NO	0.00	Hospital	assist	-0.30	-0.46	0.27
	13	NO	0.00	Hospital	officer	0.22	-0.33	-0.47
	1	NO	0.00	Hospital	doctor	0.42	-0.37	-1.01
	57	NO	0.00	average	average	0.03	-0.31	-0.15
Islamic	1	YES	1.00	Hospital	officer	0.35	-0.31	-0.74
	2	YES	1.00	Hospital	doctor	0.67	0.07	-0.62
	3	YES	1.00	average	average	0.56	-0.06	-0.66
Luth	5	YES	0.58	Disp	nurse	-0.29	0.02	1.38
	3	YES	0.58	Disp	officer	0.00	-0.07	0.30
	6	YES	0.92	Hospital	officer	0.15	-0.07	0.12
	14	YES	0.71	average	average	-0.03	-0.03	0.70
Parastatal	2	NO	1.00	Hospital	doctor	0.19	0.17	0.55
	2	NO	1.00	average	average	0.19	0.17	0.55
Priv	1	NO	1.00	Disp	officer	0.05	-0.47	-0.45
	1	NO	1.00	average	average	0.05	-0.47	-0.45
RC	2	YES	0.82	Disp	nurse	-0.55	-0.04	1.60
	2	YES	0.82	Disp	assist	0.02	-0.41	-0.29
	3	YES	1.00	Hospital	officer	-0.03	-0.37	-0.11
	7	YES	0.91	average	average	-0.17	-0.33	0.12
SDA	4	YES	0.73	Disp	officer	0.33	-0.05	-0.46
	4	YES	0.73	average	average	0.22	-0.05	-0.46

Govt: government; Luth: Lutheran; RC: Roman Catholic; SDA: Seventh Day Adventist; COGI: Church of Gospel International; Parastatal: Parastatal run by the Arusha International Conference Centre; Islamic: Ithna Asheri Mosque.

a: Incentive scores were normalized to be between 0 and 1.

b: θ_v is restricted to be between -1 and 1 with high values indicating higher capacity

c: θ_c is restricted to be between -1 and 1 with high values indicating higher practice

d: The difference is the raw difference normalized to a mean of zero and standard deviation of 1. Positive values indicate clinicians who are above average in the distance between θ_v and θ_c; clinicians who are above average in translating ability into practice.

Table 6: Average IRT scores for vignettes (θ_v) and DCO (θ_c) by incentives, facility type and clinician cadre

owner	clinicians	value-based?	incentives[a]	level	cadre	θ_v[b]	θ_c[c]	difference[d]
average	28	YES				0.03	-0.12	0.17
average	64	NO				0.05	-0.25	-0.06
average	57		LOW			0.03	-0.30	-0.16
average	14		MEDIUM			-0.05	-0.10	0.51
average	21		HIGH			0.15	-0.01	0.14
average	42			Disp		-0.06	-0.19	0.27
average	13			HCenter		0.18	-0.27	-0.40
average	37			Hospital		0.16	-0.23	-0.27
average	14				nurse	-0.32	-0.30	0.69
average	18				assist	-0.13	-0.24	0.28
average	44				officer	0.20	-0.26	-0.39
average	16				doctor	0.21	0.11	0.39

a: Reported here are incentives categorized into low medium and high where low is less than 0.33, and high is greater than 0.58.
b: θ_v is restricted to be between -1 and 1 with high values indicating higher capacity
c: θ_c is restricted to be between -1 and 1 with high values indicating higher practice
d: The difference is the raw difference normalized to a mean of zero and standard deviation of 1. Positive values indicate clinicians who are above average in the distance between θ_v and θ_c; clinicians who are above average in translating ability into practice.

Table 7: Average scores for CARES BY OWNER, VALUE-BASED, INCENTIVES, LEVEL AND CADRE

owner	clinicians	value-based?	incentives[a]	level	cadre	cares[a]
COGI	4	NO	HIGH	average	average	1.24
Govt	57	NO	LOW	average	average	-0.19
Islamic	3	YES	HIGH	average	average	-0.23
Luth	14	YES		average	average	0.84
Parastatal	2	NO	HIGH	average	average	0.27
Priv	1	NO	HIGH	average	average	-1.13
RC	7	YES	HIGH	average	average	-0.33
SDA	4	YES	HIGH	average	average	0.11
average		YES				0.25
average		NO				-0.10
average			LOW			-0.19
average			MEDIUM			0.64
average			HIGH			0.15
average				Disp		0.22
average				HCenter		-0.21
average				Hospital		-0.30
average					nurse	0.44
average					assist	0.03
average					officer	-0.21
average					doctor	0.22

Govt: government; Luth: Lutheran; RC: Roman Catholic; SDA: Seventh Day Adventist; COGI: Church of Gospel International; Parastatal: Parastatal run by the Arusha International Conference Centre; Islamic: Ithna Asheri Mosque.

a: The variable cares is the (normalized) clinician average residual for health education items of a regression that predicts response on all DCO items by vignette scores (θ_v). It indicates clinicians who perform above expected number of health education questions.

Table 8: Regression of DCO input items on θ_v and incentives

Regression	A		B		C	
variable	coef	stderr	coef	stderr	coef	stderr
α (θ_v)	0.320	(0.043)**	0.322	(0.044)**	0.505	(0.056)**
κ (t, observation)	-0.038	(0.007)**	-0.038	(0.012)**	-0.039	(0.007)**
δ (I_i, incentives)	0.081	(0.012)**	0.041	(0.012)**	0.031	(0.022)
γ ($I_i \cdot \theta_v$)	0.022	(0.011)**	0.0141	(0.012)	0.141	(0.025)**
VALUE–BASED					0.038	(0.046)
VALUE–BASED$\cdot I_i$					-0.273	(0.051)**
CARES			1.228	(0.055)**		
CARES$\cdot I_i$			-0.155	(0.064)**		
Controls for Cadre	YES		YES		YES	
$\alpha_{j\in[2,42]}$	Included		included		included	
$\beta_{j\in[2,42]}$	Included		included		included	
$\kappa_{j\in[2,42]}$	Included		included		included	
obs	18870		18870		18870	
likelihood	-11074.49		-10816.30		-11057.96	

Notes here

Table 9: IRT Regression of DCO input items on dropoff in performance over observations

variable		coef	std err
	θ_c	endogenous	
Regular Dropoff Rate			
history taking	$\kappa_{j\in ht}$	-0.021	(0.005)**
physical examination	$\kappa_{j\in pe}$	-0.035	(0.005)**
health education	$\kappa_{j\in he}$	-0.026	(0.006)**
Dropoff Rate interacted by Incentives			
history taking	$\hat{\kappa}_{j\in ht} \cdot I_i$	-0.017	(0.007)**
physical examination	$\hat{\kappa}_{j\in pe} \cdot I_i$	0.042	(0.008)**
health education	$\hat{\kappa}_{j\in he} \cdot I_i$	-0.036	(0.011)**
	$\alpha_{j\in[2,42]}$	Included	
	$\beta_{j\in[2,42]}$	Included	

Notes here

Figure 3: Approximate relationship between ability and practice and the relationship between ability and practice for a parastatal and government doctor

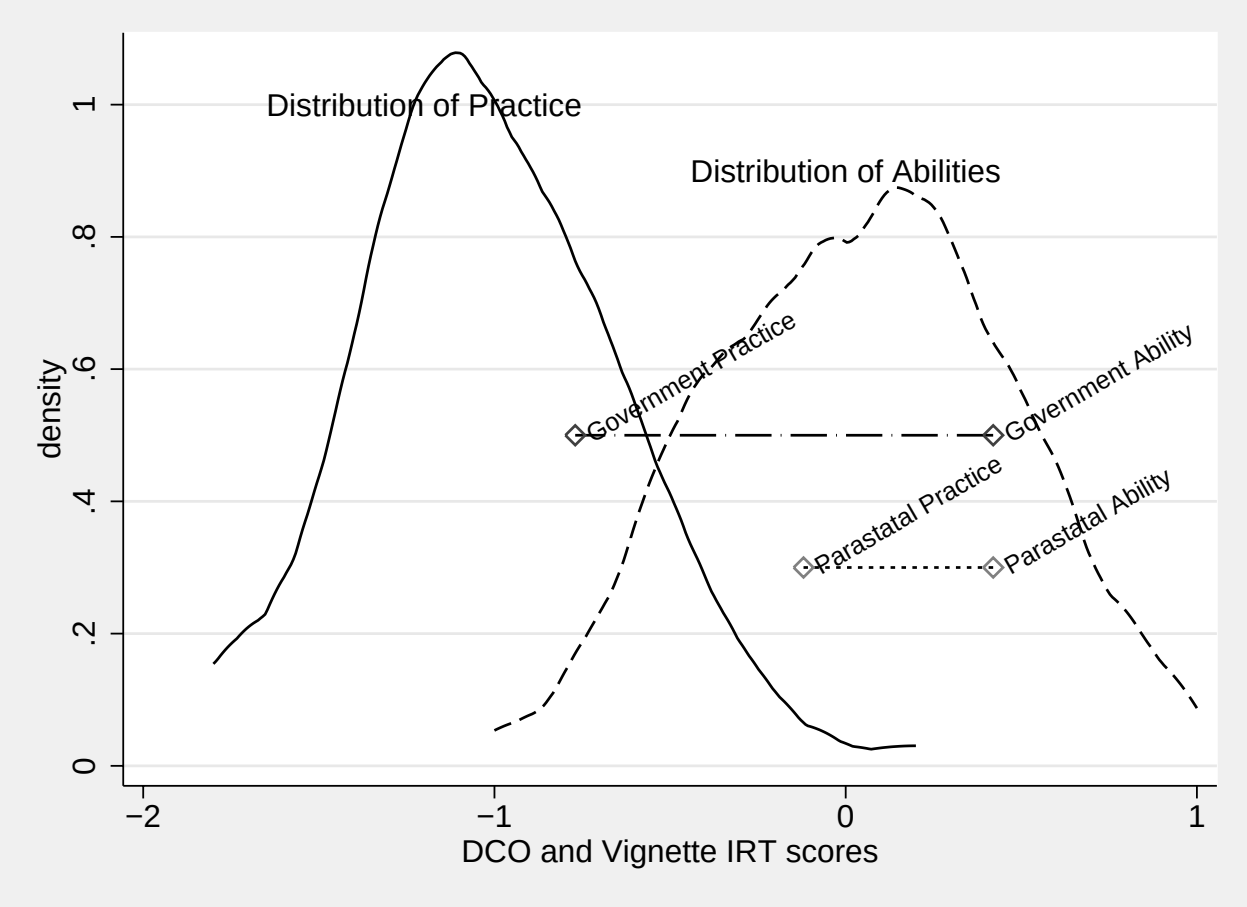


Table 10: Results of Policy Experiment for Parastatal and Government Doctor

Doctor	Parastatal			Government			Possible Government			
	ability	practice		ability	practice		simulated practice		Net Gain	
		Upper	Lower		Upper	Lower	Upper	Lower	Upper	Lower
θ_v	0.42	0.30	-0.10	0.42	-0.37	-0.77	0.31	-0.09		
	Probability of Correct Diagnosis									
Fever	26%	23%	14%	31%	11%	6%	28%	17%	17%	11%
Pneumonia	49%	45%	31%	55%	27%	17%	50%	36%	23%	19%
Diarrhea	83%	80%	69%	86%	65%	51%	84%	74%	19%	23%

Results using DCO scores have been derived using the probabilistic distribution of getting a correct diagnosis. Upper bound results are computed without any adjustment. Lower bound results have been adjusted subtracting 0.6 points in the DCO scores.

A Appendix

A.1 Factorial Analysis of Incentives

Our analysis of incentives is based on the construction of a general index; an index that can characterize each facility with respect to its incentive variables. Such a latent variable, reflecting the underlying scores of our observed incentive variables can be obtained through factor analysis.

Table 11 shows the results of factor analysis on our incentive variables.

Table 11: Factors derived from Factor Analysis

Factor	Eigenvalue	Difference	Proportion	Cumulative
1	3.307	2.690	0.866	0.866
2	0.617	0.565	0.162	1.027
3	0.052	0.088	0.014	1.041
4	-0.036	0.085	-0.009	1.032
5	-0.121	.	-0.032	1.000

Principal factors; 3 factors retained. The largest eigenvalue represents the amount of variance explained by the first axis.

Table 12: Factor Loadings

	Factor			
Variable	1	2	3	Uniqueness
FIRE	0.909	0.220	-0.095	0.116
SALARY	0.811	-0.363	0.078	0.204
FIN IND	0.958	0.002	-0.117	0.069
STAFF D	0.947	0.038	0.137	0.082
EXPAT	0.085	0.659	0.067	0.553

Table 13: weight on incentives scores to obtain the incentive factor

Scoring Coefficients

Variable	weight
FIRE	0.20785
SALARY	0.10236
FIN IND	0.36839
STAFF D	0.35924
EXPAT	-0.01673

Scoring coefficients using Stata 8.0. have been computed through the default regression scoring method. Bartlett methodologies lead to approximately the same results.

The first factor (Table 12) is clearly a measure of the intensity of the incentives. In fact all the variables appearing in the first column of the factor loading table are positive. After derivation of the loadings and without rotating the factors, scores are obtained with

the weight of each incentive parameters (Table 13). The two main variables appear to be 'financial independence and 'staff dependency'. The factor is normalized with mean 0 and standard deviation 1.