Getting clinicians to do their best: Ability, Altruism and Incentives*

Kenneth L. Leonard† Dr. Melkiory C. Masatu‡ Alex Vialou§

October 2, 2005

Abstract

The quality of health care services is an important factor in health outcomes. To what degree is the quality of care provided by a clinician in consultation limited by ability (capacity) as opposed to motivation? By measuring the ability and actual practice of a sample of clinicians in Tanzania and examining the terms of employment for these clinicians, we show that both ability and motivation are important to quality. Even after controlling for their ability, clinicians who work for organizations that use high–powered incentives are much more likely to properly diagnose and treat patients. We also show evidence that some clinicians behave in a manner that is consistent with altruism; they provide high quality independent of incentives. Changes in the incentives faced by clinicians have an important impact on quality. Significant effort has been devoted to improving the abilities of medical practitioners in developing countries; new attention should be focused on motivation.

JEL Classification: I1, O1, O2
Keywords: Incentives, Quality, Health Care, Altruism, Tanzania

*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 paper has benefitted from the comments of the audiences at the World Bank and NEUDC and the input of Sonia Lazlo and Ester Duflo. The design of the vignettes benefited greatly from extensive discussion with Jishnu Das.

†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

§2200 Symons Hall, University of Maryland, College Park, MD 20742

The quality of health care received by patients when they visit a health facility is clearly a function of the expertise of the clinician. However, if some clinicians routinely perform below their ability then expertise (or capacity, or ability), though necessary, is not sufficient for quality. In this paper, we examine a dataset from rural Tanzania in which many clinicians provide quality significantly below their capacity routinely misdiagnosing and mistreating patients (Leonard and Masatu, 2005a,b). We show that the quality of care provided is a function of capacity, but that certain types of clinicians, institutions, and institutional arrangements provide consistently higher quality than do others with exactly the same capacity. In particular, we show that quality is a function of motivation as well as capacity, and most importantly, that policy can affect motivation. The abilities of clinicians, facilities and institutions in rural Tanzania are low. However, our results suggest that there may be greater gains in overall quality from persuading clinicians to practice closer to their current capacity than from increasing their capacity with training.

While the link between health and welfare is well established, the link between health care and welfare is less clearly defined. Yet in a developing country where most ill health is due to curable illnesses such as malaria and infant diarrhea (with the notable exception of HIV/AIDS), such a link should be easy to establish. Indeed many international organizations explicitly or implicitly agree with the conclusion reached by the World Health Organization's Commission on Macroeconomics and Health (2001) that poor access to health care is one of the major impediments to balanced growth in poor countries. The argument underlying this assertion is that if poor, underserved populations had access to quality health care they would be healthier (Gertler and Gruber, 2002), which would allow them to spend more time in income-generating activities (Abel-Smith and Leiserson, 1978; Pitt and Rosenzweig, 1986; Schultz and Tansel, 1997; Townsend, 1994) This, in turn, would lead to such populations pursuing a wider range of profitable pursuits (Laxminarayan and Moeltner, 2003; Philipson, 1996), and would allow them to invest more resources in their children (Abel-Smith and Leiserson, 1978) at the same time that their children are better able to learn (Bhargava, Dean,

Jamison and Murray, 2001; Miguel and Kremer, 2004). This argument depends crucially on how "quality health care," is defined; specifically, establishing the true value of quality depends on valid measurement.

Most frequently, the quality of health care is measured by examining the capacity (or ability) of clinicians, facilities and institutions. The number of facilities or clinicians, the number of years of schooling and the experience of clinicians, the presence of diagnostic equipment or key medicines in the facility, and other static or structural features form the basis of such an evaluation. Even in those cases where quality is measured by examining outcomes, policymakers often see improvements in training or equipment as the way to remedy low–quality care. From the perspective of a patient however, the presence of diagnostic equipment, for example, is less important than the fact that the clinician is willing to use it when necessary. The fallacy that becomes apparent in the use of capacity to measure the quality of a clinician or facility, therefore, is that while quality probably cannot markedly exceed capacity, it can fall significantly short of capacity. If the actual quality of a clinician or facility is uncorrelated with capacity, the findings of studies that trace capacity to health outcomes are suspect. If, on the other hand, quality is correlated with, but falls short of capacity, is it better to increase quality using motivation or training? Even when ability is low, it is possible that closing the gap between practice and capacity has greater benefit than improving capacity. More importantly, what tools can policy makers use to close the gap between practice and ability?

Whereas clinics with absent clinicians serve little purpose (Chaudhury and Hammer, 2003), and clinicians without medicine have only limited effectiveness, our focus is on facilities with clinicians who are present and equipped with appropriate diagnostic tools and medicines. Every consultation studied in this paper represents a clinician who has the training, equipment and medicine to properly diagnose and treat the patient he is examining. We measure capacity and quality of practice by observing the activities of clinicians consulting their patients. Quality is measured by the number of appropriate inputs used by a clinician

in examining the patient. We observe and measure history taking, physical examination and health education under two sets of circumstances. With the first method (vignettes), a member of our research team acted as a case study patient for the clinician. Importantly, the clinician knew he or she was consulting a case study patient. In the second method (direct clinician observation or DCO), we observed the clinician consulting his regular patients.

Using these instruments, we develop two measures of capacity and compare the quality of practice of clinicians to their ability. Clinicians reacted to vignettes as they might to a test, striving to show their best to the evaluators, so that their scores for this instrument were higher than their scores on the DCO. Therefore, we use the behavior of a clinician on the vignette as one measure of capacity. DCO, on the other hand, allows us to measure both ability and quality in practice. The same instrument can measure both practice and capacity because of the presence of the Hawthorne effect. Even though clinicians are engaged in their normal course of consultations, they are aware of our presence, which led to a Hawthorne effect (Leonard and Masatu, 2005b). But, the impact of the Hawthorne effect fell quickly over time; most clinicians exhibited higher quality when we begin to observe them than they did 20 minutes or two hours later. This regression to type is exactly as predicted by the literature on the Hawthorne effect. Due to this change in the quality of practice over the duration of observation, we can use the first few observations of clinicians under DCO as our second measure of capacity and used the later observations as a measure of actual quality of practice.

In addition to showing that quality of practice is strongly correlated with ability, we also show that the gap between ability and quality of practice is a function of the clinicians' terms of employment. Our data includes five cadres of clinicians (from MDs to nurses) working for eight different organizations (including public, private and not–for–profit organizations) at three different types of facilities (dispensaries, health centers and hospitals). Clinicians in each of the 8 organizations face different levels of independence and accountability. Using data collected on the organizational structure of these institutions we develop a single index

of incentives and show that those clinicians receiving what we determined to be higher incentives from their organization have smaller gaps between their ability and their actual quality of practice.

Identification Reinikka and Svenson (2004) and Mliga (2000) demonstrate in Uganda and Tanzania respectively, that religious organizations provide higher quality health care than the government does, and they point to both the type of clinicians and their relationship to their employers as possible sources of this quality. Our work suggests that the institutions play an important role in motivating their employees. However, the correlation between incentives and quality shown in this paper, by itself does not prove that increasing incentives will cause clinicians to improve the quality of their care. Higher quality could be driven by clinician heterogeneity combined with selection bias, for example. Establishing the causality of incentives and performance is a difficult issue in most studies of incentives (see Chiappori and Salanie, 2002, for a summary of this literature), and with our data we cannot definitely prove that we have identified the causality. However, we do address heterogeneity and causality by controlling for both ability and non–incentivized behavior.

One potential identification issue in many studies of incentives is that clinicians who are better than average may prefer to work for organizations that reward quality. It is also possible that clinicians who work for these organizations become better over time. Thus, it is not surprising that some organizations may be better because they attract better clinicians. By controlling for capacity, we are controlling for this type of heterogeneity. Incentives are seen as impacting the difference between capacity and practice, not the level of practice.

In addition, we should be concerned about clinicians who provide high quality (or appear to provide high quality) for reasons that are not associated with incentives or ability. For example, some clinicians in our sample provided high and constant levels of health education, which, though important for healthiness, is not monitored by organizations and is therefore not rewarded directly. These clinicians are motivated to provide health education for reasons

that are not associated with the organizational structure of their employers. They may provide health education because they truly care about the health of their patients, because their patients (perhaps more educated or discerning than some other clinicians' patients) demand it, or because they react to the presence of the research team differently (they may have a more durable Hawthorne effect, for example). For any of these reasons, these clinicians are different from other clinicians. Their willingness to provide effort for activities for which they are not compensated will spill over into compensated activities and they will appear to be motivated by incentives, when in fact, they are not. It would be wrong to draw conclusions about incentives based on this type of clinician. We refer to these clinicians as 'altruistic' because, in the context of our study, their behavior mimics altruism. However, it is important to keep in mind that this behavior can be consistent with other, non–altruistic explanations as well.

Introducing ability and 'altruism' restricts the role of incentives to those who perform at or below their ability for incentivized activities and below their ability for non–incentivized but beneficial activities. For this set of clinicians, we show that clinicians who face high–powered incentives practice closer to their ability. For the set of clinicians who perform at their ability for non-incentivized activities, incentives play no role — these clinicians always practice close to their ability.

In the following section, we discuss the data and methodology for deriving quality and ability scores. Section 2 examines the differences between ability and practice using the measures that we have developed. Section 3 offers our conclusions from this study.

1 Data and Instruments

In this section, we discuss the process by which facilities were included in our sample and introduce the two quality measurement instruments that are used in the data. We compare

the two instruments directly and we explore some of the important differences between them. In addition, we outline the method by which we develop incentive and altruism scores. Translating the data from each instrument into quality scores can be greatly improved by the use of certain statistical methods (Item Response Theory) and we discuss these methods.

1.1 The sample

K. Leonard and Dr. Masatu collected the data examined in this paper 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 for 80 clinicians.[1] Both vignettes and DCO were administered by local medical personnel (see Leonard and Masatu, 2005a, for further details of the study). The facilities in the sample were chosen to match a rural household study completed at the same time. Each facility we studied has a positive probability of being visited by a rural resident of Monduli and Arumeru districts in Arusha region. In addition, our research focuses on outpatient services for illnesses that should be treatable by any of the clinicians we observe and with medicines available at all of the facilities that we visited. It is not the case that every facility we visited sees the same types of patients with the same types of illnesses. However, we focus on a subset of illnesses that is common to all facilities and there is overlap in the types of patients seen at each facility.

The clinicians in our sample include medical officers (doctors), assistant medical officers (AMOs), clinical officers (officers), clinical assistants (assistants) and nurses of various specializations (nurses). Clinical officers and assistants are similar to nurse practitioners but they see the same patients as doctors and are rarely supervised by doctors. Nurses are not supposed to diagnosis but in the rural areas they are frequently the only health personnel present and they do diagnose patients in these circumstances. We use the term clinician to

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

refer to all cadres, and the term doctor to refer to that specific cadre only.

1.2 Vignettes

Vignettes have gained increasing popularity as a tool for quality evaluation both in developing and developed countries (Das and Hammer, forthcoming; De Geyndt, 1995; Epstein et al., 2001; Kalf et al., 1996; Koedoot et al., 2002; Leonard and Masatu, 2005a; 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). 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. Each clinician faced at least 6 vignettes (malaria, PID, diarrhea, pneumonia, flu and worm infestation) and we administered two additional vignettes (TB and an at–risk pregnancy) in the second round. 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.

1.3 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 three categories of illness (fever,

cough, and diarrhea). For each of these categories there is a list of expected history taking questions as well as expected physical examination procedures. Expected items, for both vignettes and DCO are derived from national diagnostic protocols.

This process has a number of advantages, the most important of which is the direct observation of the consultation of actual patients. However, Leonard and Masatu (2005b) show that clinicians markedly increase the quality of their practice when the research team first arrives and that this immediate impact declines rapidly over the course of observations. This diminishing Hawthorne effect allows us to observe both actual practice as well as an augmented level of care that is similar to capacity.

The median consultation lasts for 5 minutes and consists of two history taking questions, one physical examination procedure and 2 health education items (which includes telling the patient his or her diagnosis.) All physical examination is done by clinicians including taking the patient's temperature, blood pressure or even weighing a child. Some consultations lasted for less than one minute and involved only 1 history taking question; others are more involved.

1.4 The relationship between ability and practice

Vignettes present the same three case study patients to every clinician in our sample and it is therefore straightforward to compare the behavior of different clinicians with this instrument — indeed this feature is a key benefit of vignettes. In addition, both vignettes and DCO measure the effort exerted to differentiate between standard and more complicated cases. For example, a clinician who takes necessary care to differentiate between a common cold and pneumonia will not score lower if the case is a common cold then he would if it were pneumonia. This is true even though total effort exerted subsequent to identifying the illness will be greater for pneumonia than for a common cold. This allows us to use DCO to compare clinicians even if one clinician usually sees patients who have pneumonia and the

other usually sees patients who have a cold.[2] Our instruments are designed so that, even if the case mix was reversed, our assessment of quality would not change. By examining effort exerted in the initial stages of consultation and restricting our attention to patients who present with fever, cough and/or diarrhea we are able to compare vignettes and DCO for the same clinician and both vignettes and DCO across clinicians.

Even though effort on the DCO and vignette should be similar, where input provision can be directly compared between the two instruments, clinicians provide much higher levels of inputs on the vignette than they do on DCO. Twenty–six physical examination and history taking items correspond perfectly on the two instruments. For each of these inputs we can compare a clinician's input provision on the vignette to that on the DCO. Table 1 shows the percentage of DCO and vignette items correct for physical examination and history taking and the total. In addition, the last column reports the percentage of DCO items correct when the corresponding vignette item is correct. The third row of Table 1 shows that the average clinician provides the average diagnostic input 49% of the time for the vignette but only 40% of the time in practice. In addition, the last column of this row shows that providing an input on the vignette suggests that a clinician is only 53% likely to provide it in practice. Table 1 is also broken down into three types of clinicians, those who work under high, medium and low powered incentives, terms which will be defined below. Increasing incentives are associated with both the unconditional probability of correctly using a physical examination procedure and the conditional probability given that the clinician demonstrated correct use on the vignette. (No effects of incentives on history taking, which costs little effort, are apparent.)

Figure 1 is a graphical representation of the same data shown over the order of observations for the DCO. There is only one observation for each clinician on the vignette but a series of observations on DCO. The upper line on this figure shows the probability that

[2]Note that if we used outcomes as our measure of quality, all cases of common cold would exhibit a 100% cure rate, and clinicians who attracted patients with complicated illnesses could easily look worse than clinicians who attracted simple, self-limiting cases.

the average clinician asked a history taking question or used a physical examination procedure if they used that exact question or procedure on the vignette. The lower line shows the unconditional probability of asking the question or using the procedure. The average clinician never demonstrates practice that is close to his ability. Importantly, however, at the beginning of the period in which he is being examined he is much more likely to do what he knows he should do.[3]

Even for items that are directly comparable, DCO scores are lower than vignette scores, clearly demonstrating that ability is not the same as practice. Practice, as measured by DCO, is inferior to ability as measured by vignettes. Furthermore, practice as measured by DCO is lower than ability as measured by the DCO itself. Taking the first few observations of quality on the DCO as a measure of ability and comparing them to the average quality on the DCO as a measure of practice, we also observe that practice is lower than quality. Using both of these measures of ability we will show that the gap between ability and practice is a function of incentives. In the regression analysis below, we do not restrict our attention to items that are perfectly matched between the DCO and vignettes and instead develop aggregate scores of ability and practice for each clinician using all of the available information.

1.5 Creating aggregate scores for the DCO and vignette

In order to compare ability and practice for the broader set of diagnostic inputs we create a single score for each clinician on each instrument. There are two concerns that we address in creating ability and practice scores. First, although we included items on the instruments because they are medically important in obtaining the correct diagnosis, in practice some questions will be more important. If we knew which clinicians were good and which were bad, we could define an empirically useful item as one that was correctly used by all good clinicians and no bad clinicians. A less useful item, for example, would be one that was raised by every clinician or that didn't discriminate by quality. Although both types of

[3]We have not included confidence intervals and these patterns are explicitly tested in the regression analysis below.

items are important to diagnosis, they have a differential ability to aid the researcher in distinguishing between good and bad clinicians. Second, the DCO does not present a simple one–off test of practice. For the vignette, each clinician was examined once for three identical conditions, producing a complete list of 31 possible inputs for every clinician. For the DCO however, there are 37 possible inputs for each clinician of which only a subset will apply to the particular illness presented. Each clinician can be observed a different number of times for any one input (from 0 to 30) and the patients seen vary in age and type of symptom even for the same input. Thus, a single DCO score needs to be able to take into account the fact that patients characteristics differ from observation to observation, that quality may fall over the number of observations and that not all clinicians were observed the same number of times for each possible input.

Scores developed according to item response theory (IRT) (Birnbaum, 1967; Bock and Leiberman, 1970) allow us to achieve all of these goals by estimating, simultaneously, a latent score for each clinician and the precision of each item. For some of the same reasons we have cited, Das and Hammer (forthcoming) also use IRT for the analysis of vignettes. Following their methodology, we use a less general two-parameter method, outlined in appendix A.1. IRT simultaneously solves for empirical weights for each item and a competence index for each clinician, and, for the version we have implemented, we can include important co-variates to control for observable patient characteristics. The result gives θ_i^V, and θ_i^C, vignette and DCO scores respectively for each clinician. θ_i^V controls for case mix by design and θ_i^C controls for case mix partially by design and partially by regression on observable patient characteristics.

1.6 Incentives

The previous discussion makes it clear that the gap between ability and practice exists. The important question is whether or not we can provide a descriptive understanding of how this gap is altered by various factors that are within the grasp of a policy maker. In this paper, we observe clinicians practicing in government services, four church–based

health systems (Lutheran, Roman Catholic, Seventh Day Adventist, and an Islamic hospital), a parastatal hospital and private practitioners who are either completely independent or part of a franchise network.[4] Nongovernmental organizations (NGOs) such as the church–run services examined in this paper, 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 [move] 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. Table 2 shows a set of measures of incentive–related organizational features developed for use in this region in previous work by Mliga (2000). 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. These incentive variables suffer from multicollinearity both because there are strong correlations between organizations and because there is no variation in these variables with-in the facility (and very little within an organization). Therefore, rather than using each measure as an independent variable, we construct a single index representative of the level of the incentives at any facility. Using factoral analysis we derive a single index of incentives that we normalize to values between 0 (lowest observed incentives) and 1 (highest observed incentives).[5] The results of the factoral analysis support a lexicographic view of the role of these four variables: a facility without the power to hire and fire its employees can never have higher incentives than a facility with this power, but the other variables do have a marginal impact on incentives.

[4]The health services of one church 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 of the church.

[5]See Section A.2 in the appendix for the derivation and explanation of this procedure.

1.7 Heterogeneity, altruism and identifying the role of incentives

In addition to incentives, there are at least three other explanations of the differences in the gap between ability and practice among clinicians. Since DCO cannot perfectly control for patient type, it may be that some clinicians face a type of patient who demands or requires higher quality care — patient heterogeneity. On the other hand, some clinicians may simply care about their patients more or may care more about our perception of their quality — clinician heterogeneity. If either patient or clinician heterogeneity is significantly correlated with but not caused by incentives, our results will be biased. In the cases of patient heterogeneity and clinicians who care about their patients, the difference in the quality we observe is real, but we would be overstating the causal role of incentives if we didn't control for them. In the case of clinicians who simply care more about our impression of them, the quality we observe is not real.

To control for these types of heterogeneity, we examine behavior that would demonstrate either patient or clinician heterogeneity but not be influenced by incentives. We use health education inputs, as measured on the DCO, as our control. Health education is not monitored by organizations, either high– or low–powered. Every clinician is trained in the use of health education and the provision of health education on vignettes shows that all clinicians are aware of its importance in the treatment of patients. However, most clinicians do not provide health education in practice, or they provide health education for only the first few observations, demonstrating the capacity but not the motivation to provide health education. There are some clinicians, however, who do provide high and consistent levels of health education.

Figure 2 shows three clinicians who represent three types of behavior that we observe in the data. Clinician 33 (leftmost) exhibits high and basically constant physical examination, but rapidly falling health education. He provides health education only for the first few observations. Clinician 32 provides low and falling levels of both physical examination and health education. Clinician 81, on the other hand, provides high and constant levels of health

education as well as physical examination. There are no clinicians who provide high health education but low physical examination, suggesting that clinicians who care to provide health education will automatically be driven to provide physical examination. There are many factors that could drive the behavior exhibited by clinician 81, including not only incentives but also patient or clinician heterogeneity. Identifying the impact of incentives will rely on comparing the incentives of clinicians who exert effort only because of organizational incentives, such as clinicians 32 and 33. Both of these clinicians are willing to display poor quality care in front of patients and the research team. Clinician 32 maintains constant physical examination (and works for an organization with high–powered incentives) while clinician 32 exhibits physical examination that is lower than his ability (and works for an organization with low–powered incentives).

In this way, we use patterns of health education to measure what we call altruism even though there are non-altruistic explanations for this behavior. We use both a discrete and a continuous measure of altruism. The discrete measure of altruism results in clinicians being assigned to either 'altruistic' or 'normal' (and in the absence of conclusive evidence, we label clinicians 'unassigned'). This labeling follows the intuition advanced in Figure 2, but uses objective methods to make the assignment. The provision of health education is predicted in a probit model with random effects at the clinician level, a clinician specific intercept and slope (with observation) and patient characteristics. All clinicians with a statistically negative slope (a fall in quality over observations) were assigned to 'normal'. Clinicians who exhibited both a positive (or flat) slope and an intercept that was above the median were assigned to 'altruistic.' Clinicians who fell into neither of these two categories were 'unassigned.' Almost 25% of the sample is defined as altruistic in this manner. Assignments to the discrete measure are made without reference to ability, cadre, organization or provision of either physical examination or health education.

In addition, we develop a continuous measure of altruism based on health education provision. We run a probit model of the provision of health inputs on cadre, patient char-

acteristics and observations and then calculate each clinician's average deviation from the predicted level of inputs. Clinicians who provide more health education inputs than would be expected (given cadre and patient characteristics) have a higher continuous altruism score. This measure does not differentiate between clinicians who provide high levels in the beginning and drop off quickly and those who start lower but provide steady levels of inputs; the average deviation from predicted trend could be the same. Nonetheless, empirically, clinicians who are labeled as altruistic by our discrete methodology have a higher continuous altruism score than clinicians who are labeled as normal.

1.8 Summary statistics of scores and incentives

Table 3 shows the basic relationship between our measures of ability and practice, the incentives facing various health care providers and provider characteristics such as cadre and level of facility as well as summary statistics for both discrete and continuous altruism scores. Most of the raw scores show what we would expect. The exception is the apparent anomaly around incentives which is not shown in the regression analysis.

The cadre of clinician is negatively associated with incentives, with most doctors working for organizations with low–powered incentives and most clinical officers working for organizations with high–powered incentives. On average, clinicians who work for organizations with high–powered incentives have the highest ability as well as the lowest practice. This is driven by one private facility (# 8) in which the contrast is particularly strong. Our regressions analyses, which permit a fuller and more complicated analysis, will show the opposite impact of incentives, as we hypothesized.

The percentage of clinicians that fall into the discrete category of 'altruistic' and normal is reported. The missing category is clinicians who cannot be confidently assigned to a category. We expect a greater proportion of altruistic clinicians in NGOs, but it is not particularly surprising that 'altruistic' clinicians can be found in all three types of organizations. Doctors are more likely to be 'altruistic' and there are no 'altruistic' nurses in our

sample. Importantly, 'altruism' is well distributed among the three categories of incentives (low, medium and high). 'Altruistic' clinicians score higher on the continuous altruism score and exhibit both greater ability and practice than normal clinicians. As expected, the raw difference between ability and practice is less than for normal clinicians. By the continuous measure of altruism, 'altruism' is well distributed throughout the various organizations, with government facilities and one NGO and private organization having the least altruistic clinicians on average. We do not show continuous altruism by cadre, since this variable was used in the derivation of the score and averages for each cadre are therefore zero.

Clinicians with the highest level of ability tend to be doctors, to look altruistic, work for organizations with high–powered incentives and work in hospitals. None of these findings is surprising. Largely speaking, these are the same practitioners who do well in practice. The regression analyses in the following section allows more rigorous testing of the patterns shown in this table.

Patient Heterogeneity The design of the instruments means that we are comparing clinicians' abilities and practices for roughly the same types of illnesses. However other characteristics of patients may vary in other ways as well. The sample of facilities was designed to include every facility commonly visited by poor residents of the rural and peri–urban areas of Monduli and Arumeru districts. Every facility in the sample is accessible to the poor. However, two facilities in our sample (labeled in Table 3 as the only facilities of organizations 6 and 7) are also very frequently visited by the urban middle and upper class. Although both facilities are also visited by the poor, it may be that quality is high simply because of the nature of the average or marginal patient. In fact, the chief of post at facility number 7 has a regional reputation and his name was known in every rural village we visited. In addition, he earned his reputation serving most of his career as a doctor in the regional referral hospital (run by the government). The quality of care that patients receive when they visit this doctor is not driven by the organization that he works for and the fact that

he now works under high–powered incentives is misleading for the purposes of this analysis. As suggested by the work of Das and Hammer (2005), it could be very important to control for this type of heterogeneity.

We claim that behavior in health education can control for patient heterogeneity and the data appear to support this claim. As shown in Table 3, 50% of the clinicians in these facilities are coded as being 'altruistic', a much higher fraction than in any other organization. However, we have controlled for this kind of behavior by controlling for 'altruism'. The doctor with the regional reputation is the clinician listed as number 81 and shown in Figure 2. Whether or not the doctors at these two facilities are good because they see different patients, or they see different patients because they are good, measuring health education allows us to set aside this behavior and concentrate on the 'normal' clinicians in our sample.

2 Analysis

The aim of our analysis is to examine the role of ability, incentives and altruism in the practice of clinicians and to test the significance of incentives in a framework that can reasonably be seen as identified. Practice quality should be improved by ability and motivation. Motivation can come from within as well as without and we examine the behavior of clinicians that we have classified as 'altruistic' as well as those who are not classified as 'altruistic'. As with normal clinicians, we expect 'altruistic' clinicians to provide higher quality care for their level of ability. However, although normal clinicians will be motivated by incentives, 'altruistic' clinicians are either unresponsive or less responsive to incentives. Thus, two otherwise identical 'altruistic' clinicians should provide the same quality care even if they work for dissimilar organizations. In addition, by controlling for continuous altruism, we allow for further complexity in the behavior of clinicians. Providing high levels of health education suggests some non–organizational source of motivation and by controlling for continuous altruism (provision of health education) we seek to partially control for these other sources

of motivation.

In the following empirical analysis, we test each of three different models on three different data frames, for a total of 9 regressions. Model I is a simple regression of practice on ability and incentives. Model II compares the importance of ability and incentives between 'altruistic' and normal practitioners and Model III examines the relationship between practice, ability and incentives, controlling for continuous altruism. The first data frame is a clinician by clinician comparison of overall DCO and vignette scores. The second frame examines the provision of each DCO item using the overall vignette score as one of the independent variables. The third frame examines the provision of DCO items without reference to vignette scores, using the initial Hawthorne effect to differentiate between ability and practice. All three data frames provide qualitatively similar results across all three models.

The base specification (Model I) for each of the three data frames is outlined in the following equations:

$$\theta_i^C = \beta + \alpha\theta_i^V + \gamma I_i + \zeta Z_i + \epsilon_1 \quad (1)$$

$$\Pr(x_{ijt} = 1) = f(\beta_j + \alpha_j\theta_i^V + \gamma I_i + \kappa_j t + \hat{\kappa} I_i \cdot t + \zeta Z_i) + \epsilon_2 \quad (2)$$

$$\text{where } j \in [1, 31]$$

$$\Pr(x_{ijt} = 1) = f(\beta_j + \alpha_j\tilde{\theta_i^C} + \kappa_{j'} t + \hat{\kappa}_{j'} I_i \cdot t + \zeta Z_i) + \epsilon_3 \quad (3)$$

$$\text{where } j \in [1, 31] \quad \text{and } j' \in [\text{ht}, \text{pe}]$$

i is the clinician index, j the DCO item index, and t indexes the order of observation on the DCO. θ_i^V and θ_i^C are the vignette and DCO IRT scores, I_i is the index of incentives and Z_i is a vector of clinician and patient characteristics. $\tilde{\theta}_i^C, \alpha, \beta, \gamma, \kappa$ and ζ are estimated parameters that measure endogenously defined ability, the contribution of ability, an intercept, the contribution of incentives and the dropoff with the number of observations, respectively.

The first data frame (Equation 1, Table 4) is a direct comparison between the overall practice (θ^C) and ability (θ^V) scores with only one observation per clinician (80 observations).

This corresponds to the assumption that vignettes measure ability and DCO measures practice. Equation 1 is solved as an OLS regression, the results of which are shown in the first column of Table 4 as Model I.

The second data frame (Equation 2, Table 5) examines the probability of a clinician providing an item on the DCO as a function of characteristics of the item as well as the ability of the clinician as defined by the overall vignette score (θ^V). There are a total of 13,395 observations corresponding to multiple observations over 31 different items for 80 clinicians. Each item has a specific intercept, and interaction between ability and observations ($\beta_j, \alpha_j, \kappa_k$). In addition, we specify $\hat{\kappa}$ which is the average change in dropoff as as incentives increase. As with the first level, this assumes that vignettes represent ability and that DCO represents practice quality, but it also allows for an impact over the number of observations. Equation 2 is solved as a probit model with clustered errors (at the facility type), and the results are shown in the first column of Table 5.

The third level of the data (Equation 3, Table 6) is similar to the second except that ability is endogenously determined from the DCO instrument without reference to the scores on the vignettes. There are a total of 19,602 observations, corresponding to multiple observations over 31 items for 84 clinicians (we can include the 4 clinicians for whom we did not observe vignette scores). The specification is otherwise similar to Equation 2 except that we do not specify item–specific dropoff rates, focusing instead on physical examination and history taking (using the index j'). Equation 3 is solved as a non-linear logit model, and the results are shown in the first column of Table 6.

Models II and III use the same three levels of data as model I but add terms for discrete altruism (δ^A, δ^N) and continuous altruism (H) respectively. Both incentives (I) and continuous altruism (H) are scaled to vary between 0 (lowest) and 1 (highest).

2.1 Results

Three models are specified on three data frames as represented in Tables 4, 5 and 6. The results show that ability matters in all specifications and that incentives are important in determining the relationship between ability and practice. In addition, there is some weak evidence that practitioners characterized as 'altruistic' respond to incentives differently than do those characterized as normal.

There are three specifications represented in Table 4. In each specification, practice is significantly correlated with ability. In the second specification, the coefficients for altruistic and normal clinicians are not significantly different from zero, but a test of the hypothesis that they are both equal to zero is rejected. The coefficients on incentives are all significant and positive except the coefficient on incentives for altruistic practitioners (γ^A), which, though large, is not significant. The coefficients suggest that 'altruistic' practitioners and those with high continuous altruism scores are better in practice, but these coefficients are not significant.

Table 5 examines the probability that a given item is provided, controlling for patient, clinician and item characteristics. The impact of incentives is tested both through a direct effect and through a change in the degree to which clinicians change their behavior over time. Incentives should make clinicians better and they should make them maintain constant quality over the time they are being observed. Each item has a specific dropoff rate estimated and we report the average dropoff rate, κ. The impact of incentives on physical examination, for example, varies from $\kappa \cdot t$ (lowest incentives) to $\gamma + (\kappa + \hat{\kappa}) \cdot t$ (highest incentives) in model I, from $\kappa \cdot t$ to $\gamma^N + (\kappa + \hat{\kappa}^N) \cdot t$ for altruistic practitioners in model II, and from $\kappa \cdot t$ to $\gamma + (\kappa + \hat{\kappa}) \cdot t$ in model III for practitioners with the highest level of continuous altruism. γ is significant and positive in both model I and II and γ^N is significant in model II, whereas γ^A, though large, is not significant; incentives matter for normal but not for altruistic clinicians. The dropoff for the average item (κ) is significantly negative, reflecting the initial Hawthorne effect, but there is not a significant change in the slope with incentives. This is probably due

to the fact that the slope is specified as a change from ability as defined by the vignettes. The coefficients for ζ^A and ζ^N suggest that 'altruistic' clinicians are better than normal clinicians, but the coefficients are not significant. However, clinicians with higher continuous altruism scores (more health education) are better than the average clinician (ζ^H in model III).

In the third data frame, shown in Table 6 there is no coefficient for ability, but a test that all the endogenously defined clinician effects ($\tilde{\theta_I^C}$) are equal rejects the null hypothesis, suggesting important differences between clinicians that are not explained by organization, patient, or item effects. As a result of the specification we cannot test whether or not 'altruistic' clinicians are different from normal clinicians overall; such tests have to be contained within an examination of the change in quality over the course of observations. The impact of incentives is seen in the slope of quality over observations. This varies from κ_{pe} (lowest incentives) to $\kappa_{pe} + \hat{\kappa}_{pe}$ (highest incentives) on physical examination in model I, from κ_{pe} (lowest incentives) to $\kappa_{pe} + \hat{\kappa}_{pe}^N$ (highest incentives) for normal clinicians on physical examination in model II, and from κ_{pe} (lowest incentives, lowest altruism) to $\kappa_{pe} + \hat{\kappa}_{pe} + \tilde{\kappa}_{pe}$ (highest incentives, highest altruism) for altruistic clinicians on physical examination in model III. In the first column we see that even clinicians with the highest level of incentives exhibit essentially the same decline in the provision of history taking inputs as clinicians with lowest incentives. However, clinicians with differing incentives are different on physical examination inputs and the results show that clinicians with high–powered incentives actually increase their provision over time (-0.042+0.068). This impact is even stronger when we differentiate by 'altruistic' and normal clinicians, with 'altruistic' clinicians exhibiting an increase over time (-0.042+0.086) and normal clinicians exhibiting a very large increase over time (-0.042 + 0.555). As with the undifferentiated result, there is no significant pattern of differences in history taking. The third column (model III) allows for an impact of both incentives and continuous altruism. Altruism significantly alters the slope such that a clinician with the highest level of altruism maintains almost constant quality in both physical examination and

history taking even if he works for an organization with low–powered incentives. Even after controlling for this impact, we find that a clinician under high incentives will maintain a constant level of effort in physical examination (-0.068+0.061).

The magnitude of the impact of incentives does not change very much across models within data frames, suggesting that controlling for altruism, either in discrete or continuous form, is less important that we had thought. The fact that altruism is well–distributed across organizations is evidence that its impact is orthogonal to the impact of incentives. The overall implication remains that organizations that use high–powered incentives get more out of their clinicians. The impact is strongest for physical examination and weak for history taking. This is not surprising since the cost (in effort or discomfort) of each physical examination procedure is much higher than for history taking. If there is shirking we expect to see its impact on costly activities such as physical examination.

2.2 Policy Implications

In order to properly weigh the advantages of gains in ability over changes in incentives we would have to have some idea of the relative costs of these policy alternatives, and we do not. However, we can get some sense of the relative importance of the two approaches to improving quality from our regressions. The results shown in Table 4 show the impact of changes in ability (θ_i^V) without controlling for the cadre of the clinician. The scores for ability have been normalized and the scores for incentives vary from 0 to 1. Using the coefficients from column 3 (0.261, for incentives and 0.126 for ability) the regression coefficients suggest that going from the lowest to the highest incentives is equivalent to a two standard deviation increase in ability, after controlling for continuous altruism. As a rough comparison, the difference between the least qualified clinicians (nurses) and the most qualified clinicians (doctors) is two and a half standard deviations in quality (see Table 3). In order to increase the overall quality of consultation (as measured on the DCO score, θ_i^C) transforming a government

facility into a private facility is roughly equivalent to 8 years of formal education.[6]

Our instincts suggest this is too high a return to incentives. Using the coefficients shown in Table 5 we get more reasonable figures. The regressions shown in Table 5 control for cadre and are based on the probability of getting a particular item correct, not an overall practice quality score. Column three of this regression shows that a complete change in incentives (0 to 1) is equivalent to about 0.41 standard deviations of ability if you ignore the interaction with observations or 0.62 standard deviations of ability if you include the interaction with observations ((0.118+20*0.003)/0.288). According to Table 3 the difference between a clinical assistant and a clinical officer is 0.8 standard deviations in ability. The government of Tanzania is currently forcing all clinical assistants to upgrade to clinical officers with two years of additional schooling. None of the upgraded assistants are in this data set, so it is not clear that two years of schooling will achieve a 0.8 standard deviation increase in ability, but if it did, these regression results suggest the same result could have been achieved by increasing incentives.

Neither of these two tables should be read as calibrating the equivalence of schooling and incentives. The results discussed above should be seen as evidence that incentives are economically as well as statistically important in the drive to increase quality.

3 Conclusion

Although it is clear that clinicians do not universally practice to the best of their abilities, 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

[6]Clinical assistants and most nurses have elementary schooling plus three years of medical training and doctors have an A-level education plus 5 years of university training, for a minimum difference of 8 years of schooling.

same level of quality in physical examination as time passes under observation, whereas other clinicians experience a significant drop–off in their level of practice. Incentives have the same impact whether we compare overall ability and quality for each clinician or whether we examine the provision of inputs against ability derived from vignettes or ability endogenously determined from the DCO.

We find strong evidence that there are multiple types of clinicians and that these different types of clinicians may react very differently to incentives. We identify a type of clinician who behaves in a manner that is consistent with altruism. 'Altruistic' clinicians are distributed among many different organizations, including the government services. There is some weak evidence that these clinicians are not impacted by incentives.

These data do not represent a perfect experiment; no clinician faces any variance in the level of incentives and clinicians were not randomly assigned incentive levels. Clinicians selected into or were selected by the organizations for which they work. The terms and conditions of employment differ in many respects not reflected in this data. Baring an experimental framework in which clinicians can be reassigned to organizations, more information on the role of incentives will come from a better understanding of the process by which clinicians choose or are chosen by organizations. In addition, the sample of clinicians is small, with only 80 clinicians in some regressions. However, since the sample includes a variety of organizations that are relevant to policy makers and clinicians in these organizations were randomly chosen for evaluation, the findings should suffer only from lack of precision or attenuation bias. The role of incentives is clear despite the lack of precision.

Quality health care is clearly lacking in Africa, and this not a new issue. It is not the case that clinicians are already well trained and equipped and clearly some training is necessary for any clinician to practice medicine. However, the issue of motivation has not received attention. Not only is motivation a way to improve care, but increasing capacity for unmotivated clinicians may have very little benefit. In Tanzania, the historically positive relationship between the government and the NGO sector allows many opportunities for the

government to learn from the management of NGOs (Gilson et al., 1997; Leonard, 2002; Mliga, 2000). Even when governments do not directly use the capacity of NGOs to advance the goal of quality they can learn from their management and attempt to design a health service sector that encourages clinicians to exert effort. The continued focus on capacity is both misleading and unproductive.

References

Abel-Smith, B. and A. Leiserson, *Poverty, Development, and Health Policy*, Geneva: World Health Organization, 1978.

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.

Bhargava, A., T. Dean, L. J. Jamison, and C. J. L. Murray, "Modeling the Effects of Health on Economic Growth," *Journal of Health Economics*, 2001, *20*, 423–440.

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

Bock, Richard D. and M. Leiberman, "Fitting a response Model for Dichotomously Scored Items," *Psychometrika*, 1970, *33*, 179–197.

Chaudhury, Nazmul and Jeffrey S. Hammer, "Ghost doctors : absenteeism in Bangladeshi health facilities," Policy Research working paper 3065, World Bank 2003.

Chiappori, P.A. and B. Salanie, "Testing Contract Theory: A Survey of some recent work," working paper 11, INSEE 2002.

Das, Jishnu and Jeffrey Hammer, "Money for Nothing, The Dire Striats of Medical Practice in Delhi, India," mimeo, The World Bank 2005.

_ **and** _ , "Which Doctor?: Combining Vignettes and Item-Response to Measure Doctor Quality," *Journal of Development Economics*, forthcoming.

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.

Gertler, P. and J. Gruber, "Insuring consumption against illness," *American Economic Review*, 2002, *92* (1), 51–76.

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.

Laxminarayan, R. and K. Moeltner, "Malaria, Adaptation and Crop Choice," Presentation at NEUDC, RFF 2003.

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.

_ **and Melkiory C. Masatu**, "Comparing vignettes and direct clinician observation in a developing country context," *Social Science and Medicine*, 2005, *61* (9), 1944–1951.

_ **and** _ , "Measuring the quality of health care and the Hawthorne effect," mimeo, University of Maryland 2005.

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.

Miguel, Edward and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 2004, *72* (1), 159–217.

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), 1–3.

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

Philipson, T., "Private Vaccination and Public Health: An Empirical Examination for U S Measles," *Journal of Human Resources*, 1996, *31* (3), 611–30.

Pitt, Mark and Mark Rosenzweig, "Agricultural Prices, Food Consumption, and the

Health and Productivity of Farmers," in "Agricultural Household Models: Extensions, Applications and Policy," Baltimore: Johns Hopkins University Press, 1986, pp. 153–65.

Reinikka, Ritva and Jakob Svenson, "Working for God?," Discussion Paper Series 4214, Centre for Economic Policy Research 2004.

Schultz, T. Paul and Aysit Tansel, "Wage and Labor Supply Effects of Illness in Cote D'Ivoire and Ghana: Instrumental Variable Estimates for Days Disabled," *Journal of Development Economics*, 1997, *57* (2), 251–286.

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.

Townsend, Robert M., "Risk and Insurance in Village India," *Econometrica*, 1994, *62* (3), 539–591.

World Health Organization and Jeffrey Sachs, *Macroeconomics and health: investing in health for economic development*, Geneva: World Health Organization, 2001.

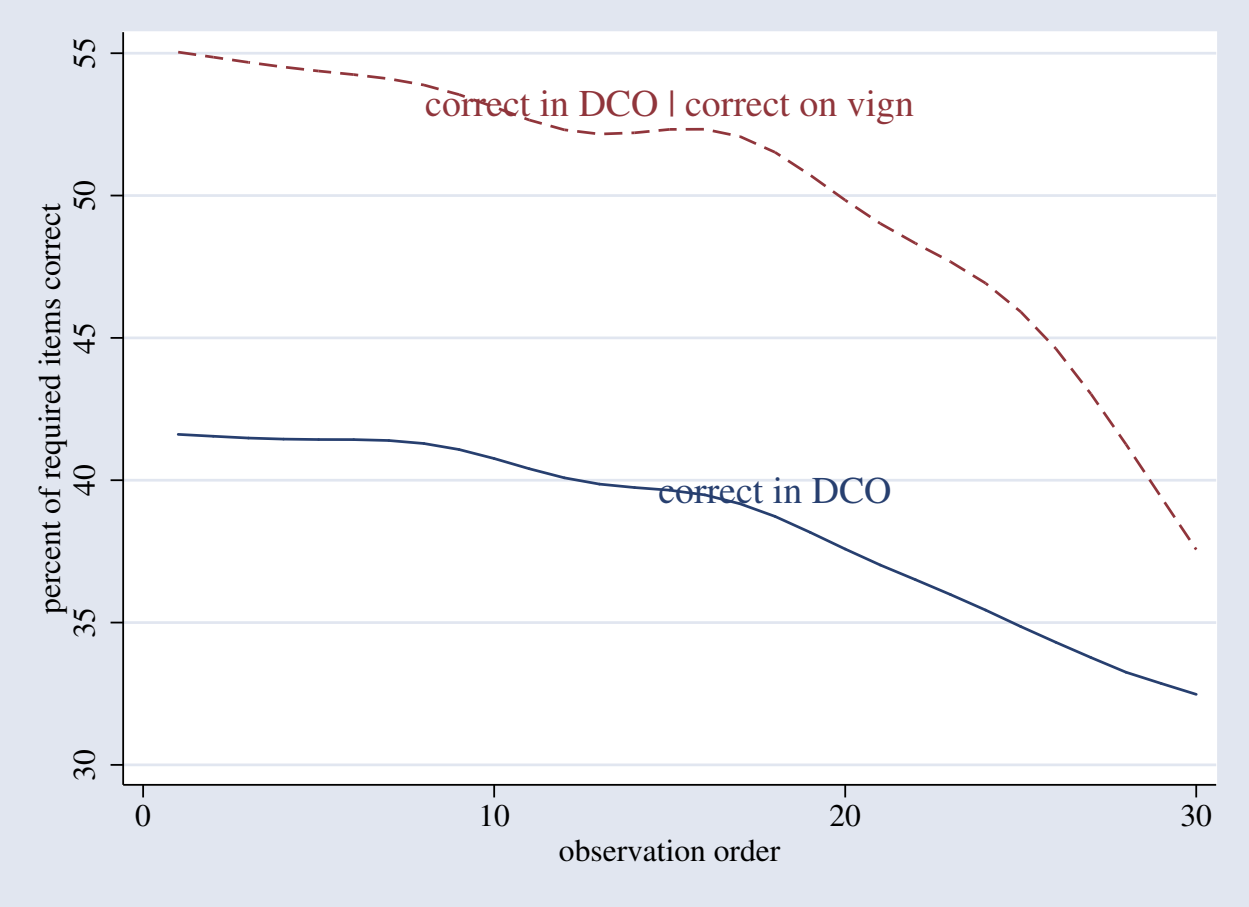
Table 1: Comparison of % correct responses to items that appear on both the DCO and Vignettes

provided on	obs #	DCO %	Vignette %	DCO cond. on Vign † %
		Total		
history taking	3887	43	52	54
physical examination	2957	37	46	52
total	6844	40	49	53
		High Powered Incentives		
history taking	637	42	56	46
physical examination	466	51	51	71
total	1103	45	53	57
		Medium Powered Incentives		
history taking	255	47	42	57
physical examination	194	48	43	61
total	449	48	42	60
		Low Powered Incentives		
history taking	2995	42	51	55
physical examination	2297	33	44	48
total	5292	49	48	52

Shown is the percentage of appropriate items correctly implemented for vignettes and DCO. There are 26 items that appear on both the vignette and DCO instruments (16 history taking and 10 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 report the the percent correct on the DCO for the subset of cases in which the same item was correct on the vignette.

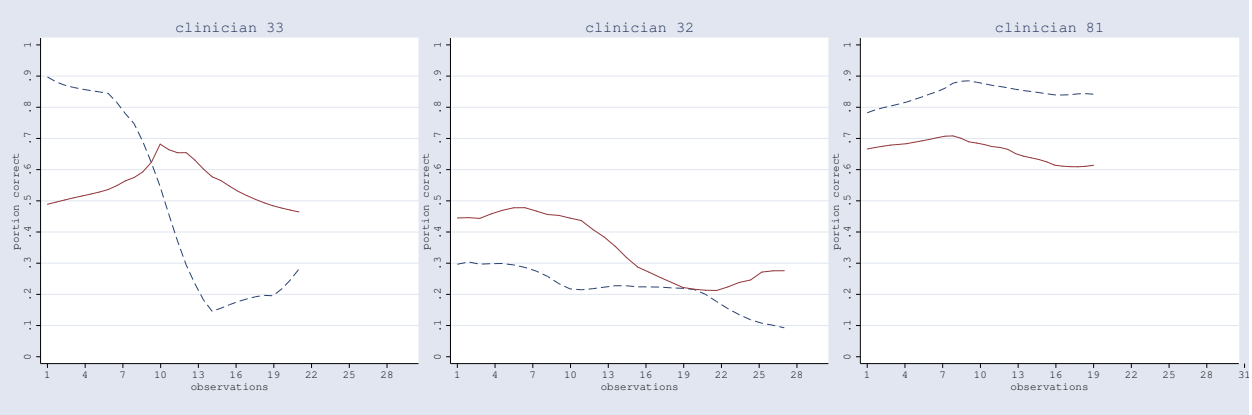
†The conditional probability of getting an item correct on DCO given that it was correct on the vignette.

Figure 1: Pattern of matches between DCO and vignettes compared to the order of observations on the DCO



The smoothed lines shown represent the change in the number of physical examination and health education items answered correctly as a percentage of possible rational items. The lines are derived from scores controlling for patient characteristics, using a local average regression with an Epanechnikov kernel and a bin width of 3 observations.

Figure 2: Three Examples of clinician behavior on physical examination and health education



The dashed line is health education and the solid line is physical examination.
The smoothed lines shown represent the change in the number of physical examination and health education items answered correctly as a percentage of possible items. The lines are derived from scores controlling for patient characteristics, using a local average regression with an Epanechnikov kernel and a bin width of 6 observations.

Table 2: 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
1	Disp	15	35	no	national	low	national	govt
1	HCenter	4	13	no	national	low	national	govt
1	Hospital	3	13	no	national	low	national	govt
2	Disp	4	6	yes	regional	medium	regional	ngo
2	Hospital	1	5	yes	regional	high	local	ngo
3	Disp	2	3	yes	local	high	regional	ngo
3	Hospital	1	4	yes	local	high	loc/reg	ngo
4	Disp	3	3	yes	national	high	regional	ngo
5	Disp	3	5	yes	local	high	local	ngo
6	Hospital	1	2	yes	local	high	local	priv
7	Hospital	1	4	yes	local	high	local	ngo
8	Disp	1	2	yes	local	high	local	priv

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.

FIN IND: The ability of a facility to use local funds to pay salaries and buy essential medical supplies to run the facility.

STAFF D: Location at which medical staffing decisions occur (for example composition of staff).

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

Table 3: Average IRT scores for vignettes (θ^V) and DCO (θ^C) by owner and type

owner	level	cadre	obs	I incentives[a]	δ^A altruistic[b]	δ^N, normal[b]	H, altruism[c]	θ^V †	θ^C ‡
Owner of the facility									
1 (govt)	—	—	61	0	21%	72%	-0.043	-0.076	-0.066
2 (ngo)	—	—	11	0.830	9%	36%	0.256	-0.323	0.1
3 (ngo)	—	—	7	0.945	0%	57%	-0.048	-0.253	-0.166
4 (ngo)	—	—	3	0.67	33%	0%	0.118	0.873	2.046
5 (ngo)	—	—	5	1	20%	60%	0.211	0.608	0.277
6 (priv)	—	—	2	1	50%	50%	0.025	0.565	0.067
7 (ngo)	—	—	4	1	50%	50%	0.13	1.595	0.519
8 (priv)	—	—	2	1	0%	0%	-0.293	0.226	-4.188
Type of facility									
—	Disp	—	54	0.297	19%	52%	0.017	-0.245	-0.136
—	H Center	—	13	0	15%	85%	-0.037	0.385	-0.046
—	Hospital	—	28	0.515	25%	68%	0.028	0.424	0.186
Cadre of clinician									
—	—	doctor	4	1.00	75%	25%		1.582	0.378
—	—	AMO	8	0.375	25%	63%		.219	0.186
—	—	officer	48	0.336	19%	60%		.392	0.047
—	—	assist	20	0.142	25%	65%		-.44	-0.222
—	—	nurse	15	0.297	0%	67%		-.96	-0.237
Level of Incentives									
—	—	—	20	High	20%	50%	0.035	0.462	-0.297
—	—	—	14	Medium	14%	29%	0.226	-0.067	0.517
—	—	—	61	Low	21%	72%	-0.043	-0.076	-0.066
Discrete Altruism									
—	—	—	19	0.294	yes	no	0.196	0.943	0.204
—	—	—	58	0.226	no	yes	-0.082	-0.187	-0.051

a: Incentive scores were transformed to the range between 0 and 1. b: Discrete altruism (normal, altruistic and unassigned.) c: continuous altruism (standardized to mean of zero and standard deviation of 1.) †: θ^V is normalized with high values indicating higher capacity. ‡: θ^C is normalized with high values indicating higher practice.

Table 4: First Data Frame: Regression of θ^C on θ^V and incentives for each clinician

Model	I	II	III
α (θ^V, ability)	**0.149 (0.065)**		**0.126 (0.070)**
α^A (θ^V, ability for altruistic)		0.119 (0.149)	
α^N (θ^V, ability for normal)		0.113 (0.082)	
Incentives			
γ (I_i)	**0.311 (0.138)**		**0.261 (0.159)**
γ^A ($\delta_i^A \cdot I_i$, altruistic)		0.428 (0.397)	
γ^N ($\delta_i^N \cdot I_i$, normal)		**0.403 (0.138)**	
ζ^A (δ_i^A, altruistic type)		0.294 (0.509)	
ζ^N (δ_i^N, normal type)		0.183 (0.503)	
ζ^H (H_i, continuous altruism)			0.414 (0.418)
constant	**1.881 (0.082)**	**1.666 (0.493)**	**1.895 (0.085)**
observations	80	80	80
R^2	0.11	0.13	0.11

OLS regression of θ_i^C as the dependent variable. Robust standard errors reported in parentheses. Variables significant at the 10% level are shown in bold font.

Table 5: Second Data Frame: Regression of DCO input items on vignette defined ability and incentives

Model	I	II	III
α (θ^V, ability)	**0.385 (0.112)**	**0.321 (0.117)**	**0.288 (0.1090)**
Dropoff in scores over observations			
κ (t, dropoff)	**-0.035 (0.015)**	**-0.033 (0.015)**	**-0.035 (0.015)**
κ ($t \cdot H_i$, altruism)			0.000 (0.011)
Incentives			
γ (I_k, all clinicians)	**0.194 (0.061)**		**0.118 (0.049)**
Incentives interacted with type			
γ^A ($I_k \cdot \delta_i^A$, altruistic)		0.195 (0.151)	
γ^N ($I_k \cdot \delta_i^N$, normal)		**0.220 (0.087)**	
Incentives interacted with observations			
$\hat{\kappa}$ ($t \cdot I_i$, all clinicians)	0.004 (0.003)		0.003 (0.004)
Incentives interacted with observations and type			
$\hat{\kappa}^A$ ($t \cdot I_i \cdot \delta_i^A$, altruistic)		0.001(0.016)	
$\hat{\kappa}^N$ ($t \cdot I_i \cdot \delta_i^N$, normal)		0.002 (0.005)	
Type			
ζ^A (δ_i^A, altruistic)		0.033 (0.122)	
ζ^N (δ_i^N, normal)		-0.158 (0.118)	
ζ^H (H_i, continuous altruism)			**0.590 (0.162)**
Patient Characteristics	Included	Included	Included
Clinician Cadre	Included	Included	Included
α_j	Included	Included	Included
β_j	Included	Included	Included
κ_j	Included	Included	Included
constant	**-0.468 (0.205)**	-0.351 (0.227)	**-0.515 (0.207)**
obs	13395	13395	13395
pseudo R^2	0.16	0.16	0.17

All regressions are a probit regression with x_{ijt} (outcome for clinician i on item j for observation t, where outcome is 0 or 1) as the dependent variable. Robust standard errors (corrected for correlation at the facility type (owner and level), reported in parentheses. Variables significant at the 10% level are shown in bold font.

Table 6: Third Data Frame: Regression of DCO input items on DCO defined ability and incentives

Model	I	II	III
$\widetilde{\theta^C}$ (clinician effect)	endogenous	endogenous	endogenous
Regular dropoff rate			
κ_{ht} (t_{ht}, history taking)	**-0.022 (0.005)**	**-0.022 (0.005)**	**-0.034 (0.008)**
κ_{pe} (t_{pe}, physical exam)	**-0.042 (0.006)**	**-0.040 (0.006)**	**-0.068 (0.009)**
Dropoff interacted with incentives			
$\hat{\kappa}_{ht}$ ($t_{ht} \cdot I_i$)	-0.007 (0.007)		-0.010 (0.007)
$\hat{\kappa}_{pe}$ ($t_{pe} \cdot I_i$)	**0.068 (0.008)**		**0.061 (0.008)**
Dropoff interacted with incentives and type			
$\hat{\kappa}^A_{ht}$ ($t_{ht} \cdot I_i \cdot \delta^A_i$, altruistic)		0.012 (0.016)	
$\hat{\kappa}^A_{pe}$ ($t_{pe} \cdot I_i \cdot \delta^A_i$, altruistic)		**0.086 (0.016)**	
$\hat{\kappa}^N_{ht}$ ($t_{ht} \cdot I_i \cdot \delta^N_i$, normal)		0.219 (0.135)	
$\hat{\kappa}^N_{pe}$ ($t_{pe} \cdot I_i \cdot \delta^N_i$, normal)		**0.555 (0.173)**	
Dropoff interacted with incentives and altruism			
$\tilde{\kappa}_{ht}$ ($t_{he} \cdot I_i \cdot H_i$)			**0.032 (0.016)**
$\tilde{\kappa}_{pe}$ ($t_{pe} \cdot I_i \cdot H_i$)			**0.065 (0.019)**
α_j	Included	Included	Included
β_j	Included	Included	Included
Patient characteristics	Included	Included	Included
obs	19602	19602	19602
log likelihood	11644	11639	11639

All regressions are a probit regression with x_{ijt} (outcome for clinician i on item j for observation t, where outcome is 0 or 1) as the dependent variable. Standard errors reported in parentheses. Variables significant at the 10% level are shown in bold font.

A Appendix

A.1 Item Response Scoring

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 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 latent competence for a given item. 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 (use a particular input). When β is low, even 'better' test takers are unlikely to answer the question correctly. Choosing a simple rule such as $\theta = \sum_j x_j = \bar{x}$ 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 31 possible inputs 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 37 possible inputs 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 control for patient characteristics and allow for a linear fall in practice quality over observations that can vary between history taking and physical examination items.

Whereas vignettes and DCO measure the provision of inputs, IRT produces an underlying latent variable. The raw average of inputs provided on the vignette ($\bar{x}_v$) is not correlated with the raw average of inputs provided on DCO ($\bar{x}_c$) (p-value of 0.12), but the vignette IRT score (θ^V) is correlated with the DCO IRT score (θ^C) with a correlation coefficient of 0.27 (p-value of 0.01). This suggests that, for our sample, ability and practice are not correlated but the latent competencies underlying ability and practice are correlated.

A.2 Factorial Analysis of Incentives

To create an incentive score we used dummy variables representing all possible levels at which control is exercised over salary decisions, financial decisions and staffing decisions

and performed a factor analysis. There are 6 significant factors and the first factor has an eigenvalue that is twice the size of the second and is straightforward to interpret as the intensity of incentives. We create our incentive score from this first factor. Table 7 demonstrates the relationship between each element of incentives and our overall incentive score. Each element is positive and significant. The greatest weight is put on the ability to hire and fire and the three other factors have approximately similar weights.

Table 7: Regression of overall incentive scores on individual incentive contributions

Variable	coef	std. err
FIRE	1.245	(0.030)
SALARY	.175	(0.007)
FIN IND	.119	(0.019)
STAFF D	.121	(0.010)
constant	-2.017	(0.017)