Using Vignettes to Measure the Quality of Health Care
Jishnu Das (World Bank)
Ken Leonard (Maryland)

I. Introduction
No matter how one looks at it—as differences across nations or as differences within nations—poor people systematically suffer from worse health outcomes than rich people. What role does medical care play?

This paper outlines a research project that seeks to measure the quality of care, understand how quality of care varies by geographical location and sectors (private, public or Non-Governmental Organizations) and how (and whether) quality of care has an impact on health choices and outcomes. We discuss a measure of the quality of care, competence, which focuses on what doctors know, or, the maximum quality of medical advice that doctors could provide if they did all they knew to do. We show how this measure can be validated and what can be learnt by looking at correlations between competence and various attributes of the health-care provider. We propose ways in which this measure can be widely collected, at the same time arguing for (some) uniformity in cross-country studies to enable wider comparisons.

The note is structured as follows. Section II presents a prima facie case for (a) incorporating the quality of care in studies of the demand for health care and outcomes and (b) measuring the quality of care through the quality of medical advice that doctors give to patients, rather than (for instance), the infrastructure in a facility. Section III introduces vignettes as a measurement tool, describing how this data is collected and validated. Section IV presents results from recent studies; Section VI concludes with some lessons learnt, caveats and thoughts for further research.

II. Why and how should we measure quality of care?
Numerous studies have documented the role of households in producing good health outcomes—children are healthier when mothers are more educated; rich households are better able to “insure” against health shocks; rich households live in areas with better sanitation and enjoy better nutrition. Based on these studies the explanations for health outcomes among poor people have centered almost exclusively on household choices: either poor people do not use the health system as much as they should or if they do go to doctors it’s usually when it’s too late. However, recent work shows clearly, that even when the poor do visit health facilities frequently and often more frequently than the rich, their health outcomes remain dismal, the quality of the medical system must also play a large role in health outcomes.

Earlier studies sought to measure the quality of care through the presence or absence of a primary health care center, and found little or no relationship between the existence of a health care center and health outcomes. The lack of a relationship left many questions about providers unanswered: Was the lack of a relationship because the doctor was never there? Was the doctor qualified (holding a degree) and competent (knowledgeable)? The data to answer these crucial questions simply didn't exist.
The next set of studies tried to address these questions by using “structural” measures of quality; that is, quality alternatively defined by physical infrastructure, the stock of medical supplies, the total number of assigned personnel, the availability of refrigeration units, the availability of electricity or a combination of some of these (Collier and others 2003; Lavy and Germain 1994). Both studies found that health-care demand responded to structural quality—more people visited health clinics when the structural quality was higher.

A remarkable omission from these indicators is any measure of process quality, particularly the quality of medial personnel. If structural quality was well correlated with process quality, this omission could be explained because it is easier to collect data on structural quality. However, they are not well correlated and there is good reason to believe that process quality is more important than structural quality. First, structural measures such as drug availability are largely determined by the degree of subsidy and cost of transportation, making structural quality a predictable feature of owner and location, whereas process quality is more likely to vary within these parameters. To the degree that one facility is more likely to experience pharmacy stock-outs than another similar facility, it is likely to because demand is high causing misclassification. Second whereas both medicine and consultation are likely to be important to a patient’s health, households can mitigate problems with drug supply through purchases from other markets, whereas they cannot do this with medical care (see for example, Foster 1995).

We propose to measure the (maximum) quality of medical advice a patient is likely to receive when s/he consults a doctor and the correlates of this quality. This is harder to collect than structural quality since it typically involves either detailed interviews with the doctor and/or clinical observation of interactions between the doctor and a number of patients. Together with structural quality this research presents a more “complete” picture of the quality of medical advice.

III. Quality Evaluation Methods

The goal of the quality evaluation instruments used in our research has been to test the ability of medical personnel to diagnose and treat common conditions (relevance) in a setting that is similar to their normal practice (realism) while allowing comparison across facilities (comparability). Our experience suggests that the ideals of relevance, realism and comparability cannot be achieved simultaneously and that compromise is inevitable. For example, we could send the exact same person to every facility in our sample complaining that she has a headache. The consultation would be both realistic and comparable, but is limited in relevance; a patient cannot fake tuberculosis or dehydration, both common conditions on which we should evaluate practitioners. A written test could be both comparable and relevant but is not realistic; the behavior of clinicians on a written test may not be a good approximation of how they behave in practice. We have implemented two types of quality evaluation methods which make different compromises. Vignettes are standardized case studies that are not fully realistic, but are comparable across clinicians and can be used for most outpatient conditions. Direct
observation collects data on the behavior of clinicians with their own patients and is therefore realistic but, because case mix varies between clinicians, is difficult to compare across practitioners and not always relevant.

**Vignettes**

With vignettes, as used in our studies, an enumerator trained to act as a sick person is presented to the clinician. The characteristics of the illness are completely predetermined, but, except for the first complaints presented by the patient, the practitioner must ask questions and perform physical examination to discover these characteristics. So that the case is comparable across practitioners, the patient is trained to know the answers to all possible questions whether or not they are relevant to the diagnosis. Because the patient is not actually sick, physical examination must be done in question and answer format where the doctor explains what he is looking for and the patient tells him what he would find. The quality of consultation is based on whether or not the doctor performed a series of predetermined steps in the consultation. Because doctors know they are being studied and physical examination is done through question and answer, some realism is sacrificed.

Some researchers implement a type of vignette in which either the case is not fully determined or the practitioner is asked about the use of specific procedures. In the first case the practitioner is asked to list the procedures he would use if a particular type of patient presented and in the second the doctor is asked whether or not he would use a particular procedure for a particular type of patient. Neither of these methods (which can be combined) increases the comparability or relevance of the vignettes, but they are both significantly less realistic than the vignettes that we use. The set of doctors who would know to check for signs of dehydration in a patient with diarrhea can be very different than those who recall that they should do so when they are prompted.

However, the increased realism of our vignettes does have an important cost; enumerators must completely memorize a case presentation and be sufficiently well trained to adapt to different questions posed by practitioners while still maintaining the same characteristics across all practitioners. This is much more challenging that training someone to read a list of questions off an instrument.

**IV. Validating Vignettes**

Do vignettes measure competence and is competence correlated with important underlying aspects of quality? To answer these questions, the validity of vignettes can be checked for internal consistency and the results obtained using vignettes can be compared to results obtained used more realistic instruments, in this case direct observation.

**Internal Validity with Item Response Theory (IRT)**

For any instrument that is a series of correct or incorrect responses to a list of questions, some degree of internal validation can be achieved through item response analysis or theory (IRT). Even though the vignette is not a series of questions, if it has been correctly
designed and certain conditions are met\(^1\), it can be seen as a exam in which using an item is a correct response and not using an item is an incorrect response. IRT is a method of weighting scores by difficulty of items estimates both a competence measure and standard error for all clinicians taking the test. IRT allows the researcher to analyze the relationship between the vignette and practitioner competence, identifying individual items or entire vignette cases that help to discriminate between practitioners. A reasonable summary of this information is contained in the information score, which is a measure of each vignette item’s contribution to the classification of providers.

Figure 1 shows the total contribution to classification of each vignette used in three different countries. The horizontal axis shows the competence of providers and the vertical axis the contribution to the classification. At least two important findings are clear under visual inspection. First, in India, the vignettes are not very good at distinguishing among low competence providers, though they are very good at distinguishing among high competence and between high and low competence providers. Though this result appears to be specific to India, that survey included informal sector providers who dominate the lower tail and the results shown for India would likely have been repeated in any survey which had included such providers. It is easy to identify poorly trained providers, but much more difficult to tell the difference in quality among these providers. In Tanzania and Indonesia vignettes are less useful at the tails of the distribution and better in the middle. Secondly, some vignettes are much less informative than others, but this difference may not be consistent across countries. For example, the TB vignette is the most informative in India, but the least informative in Indonesia. These information scores and the standard errors give a picture of the accuracy of the instruments.

Figure 2 looks at the contribution of two different history taking and physical examination items across the three countries. In all three countries (except at the lower tail for Indonesia) physical examination makes a larger contribution to the classification of practitioners than history taking. As in medical practice, some questions are more important than others in reaching the correct diagnosis, and good practitioners are careful to always use the important procedures. Our results suggest that practice in physical examination is one important way to distinguish low from high quality practitioners.

Comparing Vignettes and Direct Observation
The differential role of history taking and physical examination also comes up when we compare the quality of practitioners on vignettes to that given by direct observation of doctors with their regular clients. Using a checklist almost identical to some of the vignette cases, Leonard and Masatu (2005) show that, overall, the competence of clinicians on vignettes (by percentage scores) is not a good indicator of their performance in direct observations, even with item by item comparison. This suggests that competence is not a good measure of practice. The evidence, not surprisingly, suggests that competence is an upper bound on practice. In addition, the Leonard and Masatu (2005) show that performance of clinicians on physical examination items on the vignettes is a reasonable approximation of their overall performance in direct observation.

---

\(^1\) See Das and Hammer 2005, for a description of the necessary conditions for IRT analysis.
corroborating the evidence from IRT that behavior on physical examination is more accurate at classifying clinicians.

V. Some Results

Now that we have a measure of quality based on competence, what do we do with it? As a first pass, we can try to benchmark the quality of care, and provide some information about whether care is high quality or not. Next, we can try to look at difference in the quality of care, perhaps across geographical or income groups.

Results on the baseline quality of care

Despite the evidence that performance on vignettes is likely to be an upper bound, the overall quality of care is low, although there is considerable variation across countries and even within countries over time. In India doctors completed only 26 percent of the tasks that were medically required for a patient presenting with tuberculosis and only 18 percent for a child with diarrhea. Similarly, doctors in Tanzania completed less than a quarter (24 percent) of the essential checklist when faced with a patient with malaria and 38 percent for a child with diarrhea. That tuberculosis kills more patients every year than every other infectious disease combined in India and that Tanzania is a country with 63-96 thousand malaria deaths a year is particularly worrying in this context. Indonesian providers perform better than those in India for both patients presenting with tuberculosis and diarrhea, but presents a disturbing time-trend. As Barber, Gertler and Harimurti (2005) document, a virtual hiring freeze in the public sector led to a decline in the quality of care between 1993 and 1997, with a 7 percent drop in the percentage of checklist items completed in the vignettes.

Results on correlates of quality of care across countries

Though the quality of care is generally low, it is not evenly distributed. In Delhi, high quality doctors appear to be self-sorting into locations where they serve richer clients, even within the public health system. In Indonesia, quality is much higher in the relatively wealthier Java Bali regions than other regions. In Tanzania, the quality of care available in the rural areas is much lower than that available in the urban areas, even though rural facilities in Tanzania score relatively well on the structural quality measures. In addition, in Tanzania, the NGO sector delivers marginally better health care, but also manages to provide relatively constant quality across the rural and urban divide.

In addition, there is little evidence to suggest that purely private health care is better than public health care. Although the private care available in urban areas and the wealthy areas of towns is generally superior to that available in the rural or poor neighborhoods, this is no different than the pattern in public health facilities. Private providers in the rural and poor areas are not superior to public providers in those areas.

VI. Discussion
Do vignettes measure aspects of medical quality that matter?
In Tanzania, doctors who were examined with the vignette were asked to provide a diagnosis at the end of each case. Because all practitioners examined were trained in the same medical system, their diagnoses can be compared to the ‘true’ diagnosis for that case. Practitioners who scored better on the vignette were significantly more likely to get the right diagnosis. In addition, Barber and Gertler (2005) find that health outcomes are worse in facilities with lower vignette scores.

Clearly, vignettes do not measure everything that is important in health care and a measure of practice quality that could be compared across clinicians would give a much better picture of health care access than vignettes do. Nonetheless, vignettes make an important contribution to knowledge because they allow some understanding of the distribution of competence, which is correlated with the distribution of practice quality. In addition, competence is less likely to be affected by incentives and therefore competence gives a picture of quality that is more static and potentially harder to change. Changing the motivation of a competent doctor may be useful, but it cannot make much of a difference for an incompetent provider. In addition, the limited view of competence can offer some evidence that even a good practice quality score could not. For example, it may not be surprising that clinicians who serve poor people are more likely to be lazy and to shirk their activities than they would if they worked in wealthy areas, but why are clinicians who serve the poor less competent than clinicians who serve the rich? Somehow doctors are selecting into these areas based on their competence. This should be easy to understand for private providers, but how does this happen for public sector providers? Measuring practice quality by itself would not allow us to compare clinicians in this manner.

Why does quality vary systematically across locations, even for competence?
None of our studies to date have explicitly studied the way that clinicians choose or are posted to health facilities. Although it would seem that practitioners sort by competence according to the expected benefit of a health practice, in Delhi there is very large variation in the competence of private practitioners in poor areas, suggesting that many private practitioners (who could easily move) remain in poor areas. We don’t yet have a good way of addressing these questions.

Issues for the implementation of vignettes
As we have alluded to above, implementing vignettes requires extensive training and some degree of adaptation to the local setting. The questions that a case patient should be prepared to answer will generally differ from one country to another, particularly for questions that are not medically relevant, but for which answers must be standardized. Thus, it is unlikely to be the case that an entire vignette manual could be designed that would work anywhere in the world. However, there are potentially important benefits to using the same cases across countries. TB is the same in India or Indonesia and clinicians are only graded on whether or not they ask medically relevant questions. The fact that a doctor in Indonesia may ask questions about religious practices and a doctor in India may ask questions about the food eaten for dinner the previous night should not prevent us from comparing competence across doctors albeit with less confidence than we can
compare doctors within countries. Since there is greater variance in training, placement and incentives across than within countries, these studies can only improve our understanding of how competence is delivered.

References
Figure 1 Information by Vignette and Country
Figure 2 Information differentiated by history taking and physical examination items