

# Which Doctor? Combining Vignettes and Item-Response to Measure Doctor Quality

Jishnu Das (DECRG, The World Bank)  
Jeffrey Hammer (DECRG, The World Bank)\*

## Abstract

We develop a method in which vignettes—a battery of questions for hypothetical cases—are evaluated with item response theory to create a metric for doctor quality. The method allows a simultaneous estimation of quality and validation of the test instrument that can be used for further refinements. The method is applied to a sample of medical practitioners in Delhi, India. The method gives plausible results, rationalizes different perceptions of quality in the public and private sectors and pinpoints several serious problems with health care delivery in urban India. The findings confirm for instance, that the quality of private providers located in poorer areas of the city is significantly lower than those in richer neighborhoods. Surprisingly, similar results hold for providers in the public sector with important implications for inequities in the availability of health care.

---

\*Corresponding author: Jishnu Das ([jdas1@worldbank.org](mailto:jdas1@worldbank.org)). The modules used in this study were designed in consultation with Dr. Tejvir Singh Khurana and many discussions with Ken Leonard and Asim Khwaja. The pilot and survey was implemented by Jishnu Das and Jeffrey Hammer with N. Deepak, Pritha Dasgupta, Sourabh Priyadarshi, Poonam Kumari and Sarasij Majumdar, all members of The Institute of Socio-Economic Research on Development and Democracy Delhi (ISERDD). Further support from Purshottam, Rajan Kumar, Ranjit Gautam and Simi Bajaj, often under trying circumstances, is gratefully acknowledged. We also thank Dr. Arvind Taneja, Veena Das, R. K. Das for comments and suggestions; to the panel of physicians led by Dr. Jonathan Ellen and Dr. Zahida Khwaja for their cooperation in evaluating treatments; to Shruti Haldea for excellent research assistance; to Carolina Sánchez for her continuing support and to seminar participants at the World Bank. Finally, the project would not have been possible without the cooperation and enthusiasm of the participating providers as well as administrators of the various public sector facilities surveyed. The research was funded by a research grant from the World Bank. The findings, interpretations and conclusions expressed in this paper are those of the authors and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments they represent. Working papers describe research in progress by the authors and are published to elicit comments and to further debate.

## 1. Introduction

The "quality" of medical care is an important determinant, of both the demand for health services and health outcomes. How to measure quality, though, is more problematic. Quality has been 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). A remarkable but understandable omission from these indicators is the quality of medical personnel themselves, particularly since they account for the largest component of cost and arguably make the greatest contribution to health outcomes in these facilities.

The omission is remarkable since, as an explanation of demand, it is the nature of the advice given that is actually specific to the facility. The presence or absence of drugs, for example, indicates the degree of subsidy that a service represents (if drugs are distributed free of charge in public facilities). This measure is not informative about the inherent quality of advice or likely outcome of visiting the facility as opposed to self-medicating. Further, public facilities with high demand due to high quality are more likely to face "stock outs" if drugs are free, resulting in potential misclassifications. From the point of view of household optimization, the response to improvements in facility characteristics that can be purchased in a market (medicines are a prime example) will be less than improvements in non-tradable characteristics such as provider quality (on this see Foster 1995; for an application to schooling, Das and others 2003).

These problems are especially exacerbated in environments such as urban India, the focus of this study, where (a) the private sector is the primary provider of care and there is little variation in out-patient infrastructure, (b) health insurance is virtually non-existent so that most spending is out-of-pocket and (c) the *de jure* regulation and certification of providers does not translate into *de facto* enforcement so that households are faced with a bewildering variety of providers characterized by very different competence levels (Mahal and others 2001, Kakar 1988, Rohde and Viswanathan 1995 and Jessani 1997).

Moreover, recurrent expenditure on medical supplies and infrastructure is small compared to that on personnel. India spent over 60 percent of its recurrent health budget (which accounts for 97 percent of all expenditure) on salaries in 1990 (Reddy and Selvaraju 1994). Given the predominance of salaries in the health budget, a critical policy question is whether expenditures are efficiently allocated. Are public facilities providing a service that cannot be obtained in the (usually very large) private market?

The omission is also understandable. Apart from logistic problems and the high human capital requirements of measuring provider quality, there is little consensus on how it should be measured. This is particularly true of low-income countries characterized by multiple medical systems and degrees. While some progress has been made (Peabody 2000 and Leonard 2003) there are a number of problems that remain unresolved. One fundamental issue is the construction of a metric that can be used to gauge the validity of the measurement tool—how does measured quality compare to the "true" quality of the provider? The problem is particularly severe when providers respond optimally to incentives so that practice is a poor reflection of knowledge. In this case, using observed practice to validate measures of knowledge (as in Peabody, 2000) is invalid.

This paper addresses the problem of measuring the "knowledge" or "competence" dimension of provider quality through the use of clinical vignettes in combination with Item Response Theory (IRT) methods (Hambleton, Swaminathan and Rogers 1991). The combination of vignettes along with item response provides two benefits. First, clinical vignettes, by standardizing the cases used to judge quality, allow us to abstract from the provider's case-load mix that may reflect unobserved selection criteria. Second, item response theory uses the test data to generate a measure of the "usefulness" of the test (in a sense made more precise below) for measuring competence, eliminating the need for a separate, independent metric of comparison. Statistically this technique is related to principal components or factor analysis in that it extracts a measure of a latent variable—a student's general grasp of knowledge in school examinations, knowledge of medical procedures in our analysis—from a set of response vectors.

These techniques are used with data collected from 205 public and private providers in 7 localities of Delhi by the authors over a two-year period. The choice of these particular providers and neighborhoods is discussed below; in essence the sample is tied to a pre-existing household survey and represents the available universe of providers for households in the parallel study. The providers cover a very broad range of skills and qualifications and the particularities of an urban environment allow us to investigate a number of questions regarding the relationship between competence, market organization and provider choice in a broader research project.

This paper focuses on the measurement problem that is a necessary first stage for examining the question of provider choice and organization of the market for health care. The results are encouraging. The study finds that despite tremendous variation, the use of vignettes allows for a reasonably precise measurement of competence. The method performs well in distinguishing between providers with high and low competence although it is better at distinguishing among relatively competent providers than distinguishing within the lower end of the competence distribution.

Based on this competence index the paper presents some results on the treatment patterns of providers and the structure of the health care market in Delhi. Some surprising findings emerge. With regard to the treatment patterns, the study shows that a provider classified as highly competent by the index satisfies only a very weak condition—the ability to distinguish life-threatening situations and act accordingly, either through treatment at the clinic itself or through referrals. A significant proportion are unable to diagnose such conditions and the competence index effectively captures the provider's ability to recognize such conditions. In this sample, 28 percent were unable to diagnose a textbook case of uncomplicated pulmonary tuberculosis and 44 percent were unable to diagnose and refer a standard case of pre-eclampsia (a complication of pregnancy that both requires immediate care and is responsible for a substantial fraction of maternal deaths).

What correlates with competence? Households in poor neighborhoods are worse off in a number of ways. On the one hand, more competent providers are predominantly in richer areas and on the other, even within the public sector less competent providers are in poorer localities. Apart from the neighborhood, medical training accounts for the most significant proportion of the variation in competence with the expected sign. Finally, on the job experience (or tenure in the particular neighborhood) has little impact. If anything, it

leads to a slight decline, perhaps because any learning-by-doing is more than balanced by improvements in the training received by recent graduates (vintage effects). The distribution of competence across poor and rich neighborhoods is the same for young and old providers, consistent with the hypothesis that observed differences arise due to sorting rather than differential depreciation in competence.

The remainder of our paper is structured as follows. Section 2 highlights some problems in the literature concerning the measurement of competence. Section 3 outlines the statistical theory on which the analysis is based as well as the econometric issues that arise in the use of the competence variable. Section 4 then presents the vignettes methodology and the sampling strategy. Section 5 presents results and Section 6 concludes with suggestions for improvements in the instrument for future studies of this kind.

## 2. Measures of Quality

The measurement of quality of care has never been an easy task. Such measures are proposed for very different purposes and have, *inter alia*, measured knowledge of medical practitioners, their behavior in clinical settings and the outcomes of medical treatments. Confusing the three, which are all determined by and should be measured by different means, has been a serious problem. While the ultimate outcome of interest is the actual health improvement of a patient, it is important to distinguish inherent characteristics of the provider (knowledge or competence) from the behavior of the provider in different settings and incentive environments or the outcomes they actually achieve. Further, complications arise for a variety of reasons such as the difficulty of correcting for differences in case-loads, strategic behavior on the part of the provider and selection bias on the part of the patient.

Correcting for case-loads is a particularly difficult problem for all methods of measurement based on outcomes. Better doctors see more difficult cases either by referral or by reputation. One way of addressing this problem has been to avoid provider characteristics as indicators of quality entirely, and to focus instead on physical quality indicators such as medication and infrastructure. A second approach is to use either standardized-patients or vignettes, both of which present all providers with the same case-load mix thus abstracting from the joint distribution of provider competence and patient characteristics.

Under the standardized patient (SP) framework, every provider is visited by an identical (and anonymous) "patient" and evaluated on the basis of performance vis-a-vis this patient. In contrast, vignettes present providers with a fixed number of cases, acted out (with the knowledge of the physician) by the surveyor. For each case, the provider behaves "as-if" the surveyor is the presenting patient and conducts the case accordingly. As with the SP, providers are then assessed on the basis of their performance in each case. Two different approaches have been followed in comparing SPs and vignettes as measures of competence

One approach has been to treat the SP framework as the gold standard against which to measure performance in the vignettes. This for instance is the approach taken by Peabody and others (2000) in their comparison of vignettes, standardized patients and chart abstraction. Based on the comparison, they conclude that vignettes are an inexpensive and accurate measurement tool for measuring provider quality.<sup>1</sup>

---

<sup>1</sup>Though, given that the methods were compared only on the basis of their average scores and not on the correlation between scores on different methods, it is unclear what conclusions can be drawn from the study.

A contrasting view is that what providers do may be very different from what they *know* they should do. For instance, Rethans and others (1991) in a survey of 36 physicians in Netherlands show that SPs and vignettes generate very different responses. The authors find that treatment in practice is worse than treatment in vignettes and that the correlation between competence in vignettes and in the clinical settings was an insignificant  $-0.04$ . More recently, Leonard (2003) shows that the results from vignettes are not comparable to those from direct clinical observation (of actual patients, not SPs). More startlingly, for some questions a correct answer in the vignettes yields no further information regarding provider behavior in the clinical setting. Leonard (2003) concludes that vignettes are a useful quality evaluation tool but are not a complete replacement for direct observation.

An incentive based view of the comparison between SPs, vignettes and direct clinical observation is that the performance of a provider is the outcome of an optimizing process. In particular, performance in the clinical setting depends on provider competence, the effort expended and the provider's expectations regarding the patient's future behavior, each of which would reflect conditions of the market. For example, consider the problem of follow-up—will the patient return when asked? If the provider thinks it unlikely that the patient will return (based on her best estimate) she will adjust her behavior accordingly by perhaps not suggesting diagnostic tests or by providing medicines that can act under a number of different contingencies.<sup>2</sup>

The outcome of this joint maximization can be different for vignettes, SPs, and clinical observation, albeit in unknown directions. For instance, vignettes could differ from SPs in that, for the latter, the characteristics of the patient are exactly the characteristics of the person that the provider sees—the SP cannot pretend to be a different person. Consequently, conditioning on observables, the provider will choose as her best guess of the patient's future behavior the conditional mean of her own patient population and this could affect her performance in the case.<sup>3</sup> To the extent that a vignette controls not only the case-load mix but also the patient-mix, it thus affords greater standardization of confounding factors and more reliable estimates of provider competence, but not of case-load controlled clinical behavior.

The findings of both Rethans and others (1991) and Leonard (2003) suggest that an important issue then is the construction of a metric with which to gauge the validity of the vignettes instrument. In particular, if optimizing behavior on the part of the provider invalidates the use of the SP as a gold standard for the vignettes, how to measure the informational content of measures based on vignettes becomes a critical question. The following discussion shows how the use of item response addresses this issue.

### 3. Item Response Overview<sup>4</sup>

Item Response Theory (IRT) was first (and continues to be) used in the field of psychometrics (Rasch 1960) to understand the relationship between some condition of the patient (for instance depression) and

---

<sup>2</sup>In US hospitals for example, standard of care for patients with sexually transmitted diseases who may be unlikely to return is very different for those with whom the provider has a regular relationship. For the former, standard of care requires that a broad spectrum antibiotic be used that can work on a number of different strains, but has worse side-effects.

<sup>3</sup>For instance, a SP presenting to a provider in a locality where all patients comply with diagnostic test recommendations might receive very different advice from one presenting to a provider with the same level of competency operating in a locality with poor compliance

<sup>4</sup>This section is drawn from Birnbaum (1967) and further developments by Hambleton and others (1991) and Hattie (1983) who present a more detailed treatment of the topic.

her response to a set of questions. The theory is based on the assumption that there is an underlying latent random variable,  $\theta$ , and every question in a test maps this latent variable to a response. In the context of this paper, the latent variable  $\theta$  is interpreted as provider competence. Using maximum likelihood techniques both  $\theta$  as well as various characteristics of the test, such as the precision with which this parameter is identified by the administered test, can be recovered.

To present the ideas formally consider a test with  $J$  dichotomously scored items indexed  $j = 1, \dots, J$  and define  $x_j \in \{0, 1\}$  as the response to item  $j$ . Let  $\underline{x}$  be the  $(J \times 1)$  response vector and define a scoring rule,  $s(\underline{x}) : \mathfrak{R}^J \rightarrow \mathfrak{R}^+$  as a mapping from the response vector to the non-negative section of the real number line.<sup>5</sup> We are interested in determining the form of this scoring rule. Of particular interest are scoring rules that minimize the *error of classification* in the following sense: consider the decision rule that allocates competence as ‘high’ if  $s(\underline{x}) > x_o$  and low otherwise. An intuitive requirement for a scoring rule is admissibility in the sense of Bayes. An admissible scoring rule will then allow for decisions that minimize a linear combination of Type I (classifying as high when actually low) and Type II (classifying as low when actually high) errors. Once such a scoring rule has been derived, the “minimized” error of classification provides a direct measure of the internal validity of the test. Specifically, the variance of this error indicates the extent to which the given test can distinguish between providers of different competencies. It is precisely in this sense that the use of item response obviates the need for a separate metric to evaluate the validity of the vignettes.

Thus, for every item  $j$ , define  $P_j : \mathfrak{R} \rightarrow [0, 1]$  s.t.  $\Pr(x_j = 1|\theta) = P_j(\theta)$ . The Item Characteristic Curve,  $P_j(\theta)$ , maps the latent variable to the probability of answering item  $j$  correctly. The following assumptions are required:

1. **Unidimensionality:**  $\theta$  is a real number and not a vector of real numbers.
2. **No Differential Item Functioning:** (No DIF): Let  $y_i$  be an attribute of individual  $i$ . Then,  $\Pr(x_j = 1|\theta_i, y_i) = \Pr(x_j = 1|\theta_i)$
3. **Conditional Independence:**  $\Pr(x_j = 1, x_k = 1|\theta) = \Pr(x_j = 1|\theta) \Pr(x_k = 1|\theta) \forall j, k$ .

In the context of the vignettes, these assumptions are fairly stringent. Unidimensionality requires that the vignettes measure only one trait—say competence—and not a vector of traits such as competence and motivation. Similarly, the second assumption requires that responses are generated only as a function of this latent trait and are unrelated to any other characteristics of the provider. This may be particularly hard to justify if providers work in very different environments. For instance, if the same vignette is administered to those working in environments with very different recourse to medical testing facilities, tests ordered during the vignette may be a reflection of the provider’s circumstances rather than the latent trait.

---

<sup>5</sup>A special scoring rule is the ‘number-right’ score under which  $T = \sum_{j=1}^J x_j$ . To contrast with classical test theory, using  $i$  to denote an individual,

$$T_i = \tau_i + \varepsilon_i$$

where  $\tau_i$  is the ‘true’ test score and  $\varepsilon_i$  is a sampling error—if an individual could take a test repeated times, then the average of the test scores would converge to  $\tau_i$ ; equivalently, we have  $E(T_i) = \tau_i$ . Note that under classical test theory, we cannot separate the test from the true-score—the researcher is intrinsically interested in this score itself.

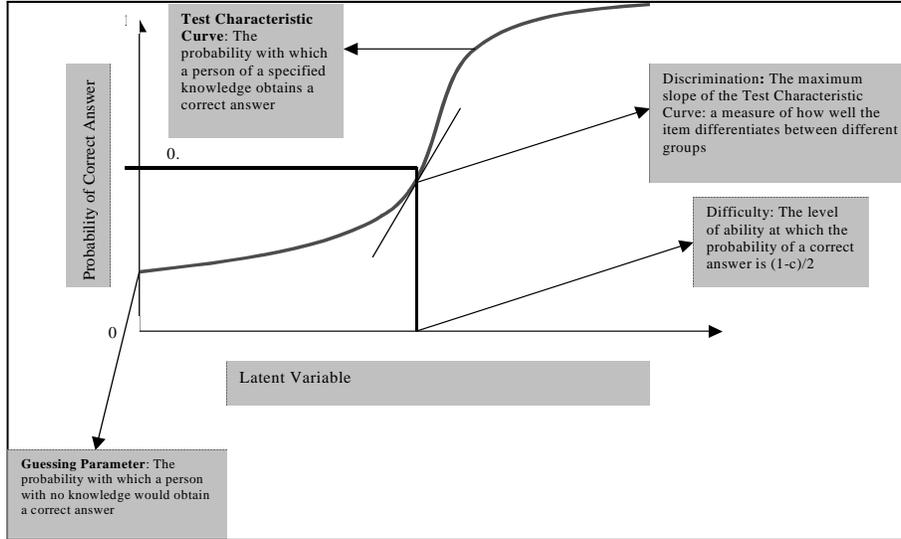


Figure 1: Parameters of the 3-Parameter Logistic

Conditional independence might also be hard to satisfy. Standard procedures followed in the diagnosis and treatment of disease form a decision tree, where successive questioning leads to the elimination of possibilities. Therefore, the response to questions further down particular branches will depend on the answers received previously. Although the construction of the vignettes took into account the assumptions of item response in attempt to minimize this problem, there are cases where these assumptions are probably violated and these are discussed later.

To complete the model, further structure is required on the functions  $P_j(\theta)$ , one for each question  $j$ . Following Birnbaum (1967),  $P_j(\theta)$  is given by the 3-parameter logistic so that

$$P_j(\theta) = c_j + (1 - c_j) \frac{1}{1 + \exp\{-a_j(\theta - b_j)\}} \quad (1)$$

The parameters  $a_j$ ,  $b_j$  and  $c_j$  each have intuitive interpretations (Figure 1).

The "guessing parameter"  $c_j$  represents the lower bound of the characteristic curve—this is the probability that an individual with  $\theta \rightarrow -\infty$  (or no competence at all) will answer item  $j$  correctly. In the context of the vignettes, this could correspond to the provider asking certain obvious questions such as "Does the patient have a fever?". The "difficulty parameter"  $b_j$  is a location parameter and is the value of  $\theta$  s.t.  $P_j(\theta) = \frac{1+c_j}{2}$ , i.e., the point on the latent variable scale where the probability of correctly answering the question is halfway between the floor given by the guessing parameter and the maximum. The greater the difficulty of the item, the lower the probability that someone with a particular value of  $\theta$  will answer the question correctly. Finally, the "discrimination parameter"  $a_j$  is proportional to the slope of  $P_j(\theta)$  at the point where  $\theta = b_j$  and thus measures the ability of the item to distinguish between values of the latent variable that are close to each other in the neighborhood of its difficulty. Due to the characteristics of the logistic function, the slope of  $P_j(\theta)$  reaches its maximum at the level of  $b_j$  though this would not necessarily be true for other functional

forms. Under assumptions [1]-[3] the likelihood function is:

$$L(\theta, a, b, c) = \prod_j \prod_i P_j(\theta_i, a_j, b_j, c_j)^{x_{ij}} \{1 - P_j(\theta_i, a_j, b_j, c_j)\}^{1-x_{ij}} \quad (2)$$

Maximization of the likelihood function then provides the required normal equations. One method of solving this likelihood is to use an iterative process by assuming a distribution over  $\theta$ , solving for the item parameters and then using the item parameters as given for the next iteration. The Joint Maximum Likelihood estimators from this process face problems of convergence and the statistical properties are not well understood (Hambleton and Swaminathan, 1985). The preferred method, Marginal Maximum Likelihood estimation is based on Bock and Lieberman (1970). With a density of the latent variable  $f(\theta)$ , the marginal probability of obtaining the response vector,  $\underline{x} \equiv (x_1, x_2, \dots, x_J)$  is

$$\Pr(\underline{x}|a, b, c) = \int_{-\infty}^{+\infty} \prod_i P_j(\theta_i, a_j, b_j, c_j)^{x_i} \{1 - P_j(\theta_i, a_j, b_j, c_j)\}^{1-x_i} f(\theta) \quad (3)$$

If there are two questions in the vignettes the possible response vectors are  $\{\{0, 0\}, \{0, 1\}, \{1, 0\}, \{1, 1\}\}$ . The expression given by equation (3) is the probability of a *particular* response vector and with two questions four such probabilities are computed—in general with  $J$  questions, the number of possible response vectors will be  $2^J$  when responses are dichotomous. Denoting  $\rho_{\underline{x}}(a, b, c) \equiv \Pr(\underline{x}|a, b, c)$  the marginal likelihood function is given by

$$L(a, b, c|\underline{x}) = \prod_{\underline{x}=1}^{2^J} \rho_{\underline{x}}(a, b, c)^{r_{\underline{x}}} \quad (4)$$

where  $r_{\underline{x}}$  is the frequency of response vector  $\underline{x}$  in the data. Since the latent variable,  $\theta$ , has been removed from the estimation, the estimated item parameters are consistent.<sup>6</sup> Once the item parameters have been estimated, the analog of equation(2) for an individual  $i$  and conditioning on the item parameters provides the required estimates of  $\theta$ . Solutions are determined to an affine transformation so  $\theta$  is assumed to be distributed with mean zero and variance one.<sup>7</sup> To relate to the previous discussion of admissible decision rules and the error of classification, the weights implied by the maximum likelihood procedure are optimal (locally best weights) and lead to admissible decision rules.

Moreover, the precision of the estimate of  $\theta$  (the minimized error of classification) is derived from Fisher's information measure:

$$I(\theta) = \sum_{j=1}^J I_j(\theta) \quad (5)$$

$$I_j(\theta) = \frac{\{P'_j(\theta)\}^2}{P_j(\theta)(1 - P_j(\theta))} \quad (6)$$

<sup>6</sup>The solution of the integral uses the Expected Maximization algorithm developed by Bock and Aitking (1981) for this particular case. The assumption of a density for  $f(\theta)$ , in general the normal, is not restrictive (Hambleton and Swaminathan, 1985).

<sup>7</sup>Solving for the parameters ( $3J$  item parameters and  $N$  doctors) together leaves one degree of freedom (replacing  $\theta' = m\theta + k$ ,  $b' = mb + k$  and  $a' = a/m$  will leave the probability of a correct response unchanged so that  $P_j(\theta) = P_j(\theta')$ )

The standard error of the estimate  $\hat{\theta}$  is then  $\frac{1}{\sqrt{\sum_j I_j(\theta)}}$  and is asymptotically normally distributed.<sup>8</sup> Since this standard error in general is in itself a function of  $\theta$ , the test will display varying levels of precision at different points of the sample range.

Finally, an important transformation of the latent variable is the proportion *true score*, defined as the proportion of questions that an examinee with the latent value  $\theta_i$  is expected to answer correctly (where the expectation is taken over repeated administrations of the same test). Using our notation, the true score,  $\tau = E(\sum_j x_j)$ . Taking expectations inside the brackets and using  $E(x_j) = P_j(\theta)$  for a dichotomous variable

$$\tau = \frac{\sum_j P_j(\theta)}{J} \quad (7)$$

The proportion true score is easily interpretable as the number of questions that would have been answered correctly if the test were taken repeatedly and this transformation is used to benchmark provider competence in Section 6.2 below.

#### 4. Vignettes Construction

The vignettes present a case and a standard-patient to the provider, who is then invited to proceed exactly as she would under normal circumstances. Following the case presentation, the provider is allowed to ask questions and answers are given according to the case history. Finally, the competence index is based on the specific questions asked regarding the history of the case, the examination of the patient, the tests prescribed and the treatment given. Given the specificities of the Indian environment, there were three important considerations.

First, the standardization of the patient was as important as the standardization of the medical problem. Hence, prior to the presentation of the case the team clearly stated the attributes of the patient. It was decided that two attributes, related to compliance and follow-up, were sufficient for the presentation of a standardized patient. In every case it was clearly specified that

1. The patient will comply with the provider's instructions, medications and tests.
2. The patient will follow-up by returning to the provider if necessary.

To check whether providers were unfamiliar with these patient characteristics, the last page of the module specifically asked about the proportion of patients that actually satisfied these two requirements. All providers felt that at least 10 percent of their patients satisfied these attributes while for the median provider

---

<sup>8</sup>As an aside, note that the regression analogue to the IRT framework is the conditional logistic model (McFadden, 1973) where the  $\theta$  are treated as fixed effects to be estimated along with the item parameters. To see this, note that in the two parameter model we have that

$$\begin{aligned} \frac{P_j(\theta)}{1 - P_j(\theta)} &= \exp\{a_j(\theta - b_j)\} \\ \Rightarrow \log \frac{P_j(\theta)}{1 - P_j(\theta)} &= \alpha_j \theta - \beta_j \end{aligned}$$

which is a linear model in the relative probabilities with  $\theta$  as (modified) fixed effects.

this increased to 65 percent, suggesting that in most cases providers had some familiarity with these attributes.

The second issue was the presentation of a consistent case, a problem that arose due to the use of different causal models by providers in the diagnosis and treatment of illnesses. Thus, models of causation could include elements of Ayurveda, Homeopathy, Unani medicine and allopathic practice. Differences in case presentation along any of these dimensions would lead to confounding differences in treatment (for instance an Ayurveda may prescribe different medications depending on the food eaten by the *mother* for diarrhea in a child).<sup>9</sup> To address this issue, during the pilot phase vignettes were administered to providers (not in the sample) from varying medical practices and all questions asked were recorded. Standardized answers were then formulated for these questions and the process was repeated till at least 90 percent of all questions asked regarding the vignettes cases had been incorporated into the design. This process ensured standardization of the entire case, even for practitioners of alternative medical systems although they were not (necessarily) graded on this material. As it turned out, questions specific to Ayurvedic or Unani medicine were almost never asked.

Finally, the cases presented in the vignettes were chosen to adequately represent the wide variety of competence among the providers in the sample. Specifically, following the declaration of Alma Ata (1978) the Indian government introduced a number of medical qualifications, some requiring less than 6 months of training (Registered Medical Practitioners) and others more than 5 years (Bachelor of Medicine and Bachelor of Surgery). This system was ostensibly supposed to result in an efficient system of referral and triage with the less trained providers recognizing and treating certain simple illnesses (diarrhea, viral fevers) and referring others to more specialized providers. Consequently, the vignettes include cases that should be treated at lower levels of the referral system as well as those that should be referred to and treated at higher levels.

The final module has 5 cases. Each case is divided into three sections: History, Examination and Treatment. For the History and the Treatment parts, there is a patient profile (known to the interviewers) that is used to answer questions that the provider asks. Providers are not prompted with any questions in these two sections, and information is given only for questions that are specifically asked by the provider. The treatment section is divided in two: the first part records the treatment given while the second consists of prompted questions asked to elicit treatment options for different types of patients (primarily related to whether the doctor would change her treatment or test recommendations if the patient were poor). The interview starts with the presentation of the case, proceeds through the history, examination and treatment and finally the prompted treatment questions are asked.

The five cases were chosen to resemble the most typical presentations of the overall disease profile in the localities surveyed.

---

<sup>9</sup>Note thought that this study cannot evaluate treatment involving medications that are not allopathic (such as homeopathy or Ayurveda). Interestingly the treatment of the illness remained primarily allopathic—less than 5 percent of all medicines prescribed (both in the vignettes and in clinical observation conducted later) were non-allopathic drugs and in most cases these were accompanied by allopathic medicines that could be evaluated.

1. **Case I:** A mother who brings in a 8 month old child suffering from diarrhea. The child is not severely dehydrated, the diarrhea does not arise from an infection and oral re-hydration therapy is the only recommended action.
2. **Case II:** A man with a one day history of cold and cough, with no fever. The correct diagnosis is viral pharyngitis and the "right" thing to do is to abstain from medicating, except for symptomatic relief.
3. **Case III:** A man with a one month history of weight loss, low grade fever and coughing. If the provider asks, the man has blood in his sputum. The diagnosis is tuberculosis and the correct treatment is either multi-drug therapy or referral.
4. **Case IV:** A 17 year old girl, bought in by her mother, who complains of weakness, lethargy and sudden bouts of crying. The mother says that the child suffers from "Low Blood Pressure". The symptoms in this case were chosen as frequent somatic representations of depression as has been determined from the household survey.
5. **Case V:** A woman complains of a severe headache. The practitioner is told that he/she notices that the woman is pregnant (in the seventh month if asked). The underlying problem is a case of pre-eclampsia, an acute life-threatening condition requiring immediate referral to a hospital.

These cases were chosen so that a different set of "ideal" responses could be elicited regardless of the position of the provider in the medical care system. Thus, Case I and Case II should be treated in a primary care context and not referred to higher treatment levels. Case III should be treated under the government's ongoing Directly Observable Therapy (DOTS) program, but may be treated in a primary care context as well given the perfect follow-up of our standard-patient. Case IV can either be treated or referred, while Case V should be immediately referred to a hospital.<sup>10</sup> Reflecting their position in the triaging system, 80 percent of providers felt that they saw Case I and II almost every day; this decreased to 25 percent for Case III and further to 15 and 8 percent for Cases IV and V, respectively (Table 1).

In the context of the IRT framework it was important that the questions used for grading were such that the answer to any one question would not affect whether some other graded question would be asked. Consider for instance the first case of a child with diarrhea. In this particular case, the provider is recorded as having given a correct response for each of the following questions asked:

1. Does the child have a fever?
2. When did the child last urinate?
3. Is there vomiting?
4. What is the frequency of stools?

---

<sup>10</sup>In three cases (except Case III and possibly Case IV) no medication is required except those given for symptomatic relief. This is a drawback of the current vignettes design—there is no case that requires a specific medication to be given in the primary care setting (such as Albendazole for a worm infestation). In future studies we plan to include a sixth such case.

5. Is there blood and mucous in the stools?

Each of these questions must be asked for a complete history irrespective of the answers to the others both to determine the level of hydration of the child as well as the possibility of a bacterial infection. While clearly these conditions could not be met in all cases (the movement from history and examination to treatment is particularly likely to violate this assumption) the choice of questions was based largely with this conditional independence assumption in mind.

## 5. Sample

The sample is drawn from seven localities in Delhi where a concurrent household survey over two years was conducted by the same team. The attributes of the households in these localities is detailed in Das and Sánchez (2002). The authors find that the sample is not significantly different from Delhi samples of either the National Family Health Survey or the National Sample Survey for most variables, though households in the study are slightly better off in terms of consumption aggregates. There was only a 4% refusal rate to participate in the survey thus making it unlikely that selection into the study biases the sample in terms of health status.

The sample frame for the current study was constructed in two steps. First, a universe of providers was constructed and a short questionnaire with basic questions on training, education and tenure was administered. This universe consisted of two different sets of providers—those who had been visited by households in the survey and those who practiced within a 15 minute walking radius of any household in the sample.<sup>11</sup> The study sample was chosen by drawing 20 to 25 providers from each locality from the set of providers visited and an additional 10 from those who were in the universe, but had never been visited. The first set was drawn via a probability-proportional-to-visits (size) sampling scheme and the second was drawn randomly with equal weights (the scheme for the first set explains the variation across localities—some neighborhoods had fewer total providers visited during the survey period). Thus, the sampling scheme addressed two purposes by simultaneously ensuring that information would be gathered on providers visited by the household as well as ensuring that the overall profile closely resembles the choice set of the average household in the city.

Over a period of 12 months vignettes were completed for 85 percent of the original sample. There were two reasons for non-participation, both of which could be correlated to the competence of the provider. The first was turnover whereby the original provider had left the locality or the clinic had shut down since the time of the household survey, a period of between 6 and 12 months. This accounted for 2.5 percent of the original sample. The second was refusal to participate, accounting for 10 percent of providers in the original sample.<sup>12</sup> The direction of bias is hard to determine since refusal to participate could indicate

---

<sup>11</sup>Not all providers who were visited by the households could be contacted. Failure to contact arose due to lack of sufficient information (visited "the doctor in X locality") or the inability of the team (and the household who had visited the provider) to locate the clinic.

<sup>12</sup>The remaining 2.5 percent were all providers who had been visited by households in the past, but could not be located. Typically, these were single visits by household members more than 6 months prior to the survey.

low competence if such providers were worried about the results of the study or high competence if the opportunity cost of time exceeded the incentive to participate.<sup>13</sup> Moreover, disaggregation by the locality of the practice (more competent providers are in richer areas—see below) is not informative since rejections were equally distributed across neighborhoods.

The providers chosen have varying qualifications, institutional affiliations and tenure in the neighborhood. For this paper, qualifications are aggregated into two groups—those who hold Bachelor of Medicine and Bachelor of Surgery (MBBS) degrees and those who do not. The former are the equivalent of MDs in the US; the latter can be practitioners with formal training in alternative medicine, those with no formal training but with degrees recognized by the government, and those with no formal degrees or government recognition.<sup>14</sup> Finally, providers in the public sector are divided in two—those who work in public dispensaries and primary health care centers and those who work in hospitals. Since hospitals in Delhi historically reflected the standard of care in the country, we would expect these two groups to differ substantially.

This sample interviewed, which reflects both actual and available household choice, is slightly different from the universe of providers. In particular, it tends to over-represent MBBS doctors and providers in the public sector. While 45 percent of the universe are MBBS degree holders, this increases to 55 percent in the sample; similarly, public doctors represent 10 percent of providers in the universe, but 20 percent in the sample. This difference could arise due to patterns of household choice (if more competent doctors see more patients) or more rejections by providers with low competence. What this means for our averages is that the results likely represent an *over* estimate of true competence in the population.

## 6. Results

The construction of the competence index uses the history, examination and treatment sections of the vignettes (Appendix 1, Table A lists every item used). For the treatment questions, a single variable is used for every case that summarizes the quality of treatment. This variable is based on ratings on a scale of -3 to 3 by three sets of independent doctors, two from South Asia in similar epidemiological environments and one a team from The Johns Hopkins University School of Medicine. For the present set of results, we treat this variable as dichotomous by collapsing the negative and the positive scores.<sup>15</sup>

From a total of 104 questions, 26 were dropped since they were asked by fewer than 5 percent of providers and the maximum likelihood procedure is not guaranteed to converge with questions that exhibit very low response rates. While this does not affect our ability to classify providers (since there is no variation in the data elicited from these questions), the fact that certain kinds of questions were not asked is interesting for

---

<sup>13</sup>The lack of systematic enforcement of medical regulations has always been an important issue in Delhi and one that results in sporadic action when investigated by newspapers and journalists. For this study almost one year was spent obtaining entry to the field through conversations with providers and repeated interactions to build trust.

<sup>14</sup>Das and Das (2003) discuss these numerous degrees and their origins. Note that there are further subdivisions: even among those with formal training and government recognition, the training period can vary from 1 to 5 years.

<sup>15</sup>An alternative would be to treat this question as a polytomous response and use the functional form of Samejima's (1969) graded response model. Another worry is that the ranking of the provider is sensitive to the rater chosen. With regard to the second, once the treatment score is treated as dichotomous, the kappa measure of agreement averages 90 percent across the two South Asian raters and 80 percent across the South Asian and the US raters. Moreover, since the informational content of treatment questions is very similar to other items in the vignettes, greater subdivisions or the use of different raters do not alter the results.

its own sake. Less than 5 percent ask about the respiratory rate of the patient for any illness including cases II and III where the patient complains of coughing as a main symptom. Apart from this, most omissions arise from a failure to ask what may be considered sensitive questions (sexual behavior, pregnancy for an unmarried woman) or a failure to probe the history of illnesses in the case of the woman with pre-eclampsia. The remaining 73 questions and 5 treatment scores are used to construct the competence index (Appendix 1, Table B lists the questions dropped).

### 6.1. Construction of the Competence Index

Figure 2 presents the distribution of the latent variable  $\theta$  (competence) and the standard-error of the estimates following equation (5). Since the competence index is standardized, individual values can be directly interpreted as standard deviations from the mean. Some interesting characteristics of the vignettes as a measurement tool immediately emerge. The standard error is not symmetrical—classification errors are larger for the least qualified providers compared to the better qualified. The most precise estimates are for providers one standard deviation to the right of the mean. However, the precision on the low end of the spectrum is high enough to distinguish between many interesting subpopulations of providers, such as the type of degree, neighborhood in which they practice and whether they are in the public or private sector.

Figure 3 illustrates in greater detail the relative lack of precision in the region of low competence by examining the difficulty and discrimination ability of each question—the parameters  $a_j$ ,  $b_j$  and  $c_j$  in Equation (1). Although there are a large number of items with difficulties in the mid and upper ranges, there are only 5 items that have difficulties low enough to reduce classification errors for the lower half of the latent variable distribution leading to higher standard errors (recall that the difficulty represents the competence at which the probability of a correct response is halfway between the guessing parameter and the maximum). These include three tests for tuberculosis (blood analysis, sputum and chest X-ray respectively), taking the blood pressure of a woman with pre-eclampsia and an overall treatment score for dealing with depression. Further, the only item that has high discrimination is item 65, examination of blood pressure in the case of pre-eclampsia while the remaining four display fairly flat item characteristic curves.

In terms of provider competence, this has (small) positive and (large) negative repercussions. For questions about tuberculosis related tests, answers corresponded to a case of simple pulmonary tuberculosis with no complications. The test for acid-fast bacilli (AFB) is positive. The profile of the patient includes an actual X-ray of a tuberculosis patient. Any questioning on the right track would lead to an unambiguous diagnosis. Similarly, asking about blood pressure for the pre-eclampsia case also indicates that the questioning is on the right track. The fact that all these questions had lower difficulties indicates that providers on the left half of the competence distribution *were* able to identify (to some extent) the problems that patients were presenting with.

On the negative side, poor measurements on the lower tail of the distribution arose despite our prior knowledge of the distribution of competence. For instance, all the vignettes involve straightforward cases with no complications that are drawn from the most common diseases the providers actually face. That less than 25 percent of the sample examined the patients for any sign of dehydration in a child with diarrhea or

that less than half (49 percent) asked the color of the sputum for a man with a persistent month long cough and weight loss (when asked, the answer would have been "clear with bright red flecks"—a classic sign of TB) is disturbing to say the least.

Figure 4 illustrates three things: (a) the contribution of specific, particularly interesting questions to the estimation of the competence index; (b) what a "good" question is in terms of discrimination ability and difficulty and (c) the tangible meaning of "competence" as estimated by  $\theta$ . The figure plots observed proportions and model predictions of the probability of asking a question against competence. For both proportions, 15 groups were formed, over which responses were averaged. The model predictions are thus our estimate of  $P_j(\theta)$ , the Item Characteristic Curves of item 65, item 39 and item 12.

For examining a woman with pre-eclampsia for high blood pressure, the predictions track observations well, and that the function is fairly steep (has good discrimination power) in the vicinity of the sample mean ( $\theta$  equals zero) implies that it distinguishes well between providers to the left and the right of the mean. The difference between a "bad" provider (say,  $\theta < -1$ ) and a "good" provider (say,  $\theta > 1$ ) could be the difference between life and death—almost all competent providers will discover that the woman has dangerously high blood pressure and more than half of bad providers will not.

Checking to see if a child with diarrhea has a depression in his skull fontanel (indicating severe dehydration) is a difficult question. The discrimination ability of the question is high, but only at the higher ranges of competence. Thus, until the provider reaches a certain competence level, the probability of checking dehydration in this particular manner is very low, though even among our best providers it never exceeds 40 percent. This is revealing of overall competence in our sample, since even though in our vignette the child does not have this depression, failing to check could miss a serious problem.

The easiest question in the vignettes was to ask for an X-ray for a suspected tuberculosis patient. The estimate of the "guessing" parameter is extremely high (just under 80 percent). The figure also illustrates, however, a limitation of the item response technique—the actual data show that providers with low competence do much worse than predicted. The failure of the structural model to pick this up is a consequence of the logistic functional form, which is not flexible enough to capture the actual shape of the data for this item.<sup>16</sup>

This naturally raises the question of how reliable our identifying assumptions of unidimensionality, no differential item functioning and conditional independence are for the data. Two important tests that have been proposed in the literature correspond to model fit (how good is the 3 PL structure in our case for the data) and unidimensionality. The former relies on the analysis of model residuals and a chi-square test to

---

<sup>16</sup>One of the advantages of using item response is for designing vignettes in future such studies. Since information (equation(5)) is additive in items, one can choose the items required for a target information function. In the vignettes the choice of items corresponds to the choice of cases. Disaggregating the information functions case by case shows that most information is provided by the tuberculosis and pre-eclampsia cases. In fact, the average contribution of these two cases to the entire module is 80 percent, with higher contributions at lower competence levels. Thus, to distinguish "bad" from "good" providers or "average" from "good" providers, the use of these two cases is sufficient. However, at high competence levels the informational content drops to below 75 percent, and thus the two cases may not be sufficient to distinguish the "very good" from the "good" providers.

check if the predictions of the model are very different from what is actually observed. Thus, we can define  $r_{ij} = P_{ij} - E(P_{ij})$  for each item  $j$  and a subgroup of providers  $i$ . The chi-square statistic (Yen, 1981)

$$Q_j = \sum_{i=1}^m \frac{N_i [P_{ij} - E(P_{ij})]^2}{E(P_{ij}) [1 - E(P_{ij})]} \quad (8)$$

is then distributed with  $(m - 3)$  degrees of freedom where  $m$  is the number of subgroups (Figure 4 presented exactly this comparison graphically for three items, where  $m = 15$ ). The chi-square test does not reject that the item characteristic curve fits the data for any of our items at the 5 percent level, although the null is rejected at the 1 percent level for items 69 and 75.<sup>17</sup>

A second assumption that can be examined is unidimensionality, that is, whether the responses to the vignettes arise from a single underlying latent variable. A heuristic method to validate the assumption of unidimensionality is to check the eigenvalues from factor analysis of the items in the test (Drasgow and Lissak, 1983). Figure 5 presents the eigenvalue plot from this analysis. The first eigenvalue is almost four times as large as the second, and apart from the first eigenvalue, the rest are indistinguishable. This suggests that the unidimensionality assumption is valid with first component referring to the competence of the provider.

## 6.2. Benchmarking the Index: How Good is a Competent Provider?

From equation (2) and the discussion of the likelihood procedure, the competence index is standardized due to an indeterminacy in Equation (1) when competence and item parameters are jointly evaluated. Thus, the classification errors discussed so far refer to the probability of misclassification *within* a pre-determined scale but are not informative about how this scale corresponds to competence in an "objective" sense. Since the design of the vignettes was such that a highly competent provider would ask 90 percent or more of the questions included, one way to understand what competence implies in terms of overall performance is to benchmark the results using the true score (equation (7)) of providers at different levels of the competence index. A second source of information are the item characteristic curves for the treatment questions, which can be directly examined to see how competency affects treatment. This provides information, for instance, on whether a provider with average competence in our sample provides the correct treatment for tuberculosis 50 percent or 90 percent of the time.

Using the true-score, even among the highest quintile of the competence index, performance is poor (Table 2). While the difference between the lowest and highest quintiles is almost 2 standard deviations of the competence index, the percentage of appropriate questions asked increases only from 15 to 48 percent—the most competent providers in the sample (in expectations) ask only half the relevant questions for the cases presented. With only 9 history questions asked over 5 cases, it is impossible to have determined, for example, that a child's diarrhea was, in fact, relatively harmless or that the woman with pre-eclampsia in Case 5 needed immediate attention.

---

<sup>17</sup>The chi-square test is not conclusive evidence of "good" model-fit. First, the expected proportions are computed using the identifying assumption of the structural model and second (especially in this case), small sample sizes will lead to a higher probability of acceptance. A rejection of the chi-square is thus very strong evidence that the correct model was not used for the relevant item and thus suggests re-examination of items 69 and 75 in the vignette design.

What does this imply for treatment? Table 3 shows the percentage of doctors who were graded as having given a treatment that was "not harmful" for the patient in either the short or long run by the raters. Harmful treatment, in the case of diarrhea for instance, implies using antibiotics and/or anticholinergics; in tuberculosis failure to either refer or start the patient on multi-drug therapy and in the case of pre-eclampsia failure to refer the patient to hospital for immediate follow-up.<sup>18</sup> While there is some variation, on average in the highest quintile, only 70 percent treated the patient in a manner that was graded positively by the raters; among the lowest quintile this drops to 30 percent. On average, competence levels have to be between 0.6 and 1.3 standard deviations above the mean for providers to have a better than even chance of not harming the patient (the notable exception is the case of tuberculosis where providers who are 1.34 standard deviations below achieve this level as well). Accounting for the guessing probability, this implies that providers "know" the correct treatment about 40 percent of the time.

Looked at in another way, the average provider's treatment is not harmful only between 25 percent and 50 percent of the time for four out of our five cases (again, for tuberculosis, even less competent providers are rated positively 50 percent of the time and this increases to above 90 percent for the highest quintile). Particularly interesting is the case of diarrhea where any treatment that required the use of an oral rehydration solution without antibiotics and/or anticholinergics was treated as beneficial by our raters. Given the widespread advertising and informational campaigns undertaken by the World Health Organization and other health organizations regarding the treatment of infant diarrhea, it is telling that the majority (note that even in the top quintile beneficial treatment is given only 60 percent of the time) do not treat this case of viral diarrhea according to protocol—in most cases by overusing anti-infective drugs and antidiarrheal drugs. Perhaps equally disturbing is the finding that only 67 percent recommended any fluid intake at all, 46 percent for the lowest quintile and 78 percent for the highest.

### 6.3. *Which Doctor?*

Using this index, the paper turns to questions regarding the distribution of competence, and in particular, inequities in the availability of care. Health care depends critically on the availability of care. For this, distance matters. Both due to transport and time costs as well as less measurable social factors, the set of providers that households can visit is circumscribed by residence—most visits in the household survey are within 15 minutes walking distance, a restriction that is particularly binding for women outside the labor force. Thus, the care available in the neighborhood is a reasonable first approximation of the choice set of households.

Inter alia, the analysis addresses questions regarding the distribution of competence among public and private providers. Debates on the relative competence of the public and private sectors in India have raged for years, particularly since over 85 percent of all visits to health care providers are in the private sector (for the households in the sample, this increases to 90 percent). Some believe that quality in the private sector is higher and that is why the private share is so high, particularly among the relatively educated and affluent. Others note that there are no enforced standards in the private sector and that many providers are

---

<sup>18</sup>These guidelines were emerged at by the raters themselves; the implications are deduced from the type of treatment and the rater scores.

unqualified "quacks". The decomposition of competence shows that both these views are consistent with the data.

Figure 6 plots the distribution of competence for three different categories of providers—MBBS degree holders in the public sector, MBBS degree holders in the private sector and non-MBBS providers (all in the private sector).<sup>19</sup> The first three graphs show the histogram and the overlaid kernel density for each of the three categories while the last shows the kernel densities for all three categories together (histograms are presented since sample sizes are small).

Two characteristics emerge. The first is that private providers belong to two very different groups—the distribution of competence among non-MBBS providers is skewed to the right with the mode (equal to the mean) at -0.7, while that of MBBS providers is skewed to the left with a mode more than one standard deviation higher at 1.3 (mean 0.65). The distribution of all private providers is distinctly bimodal. Public providers, while always bounded by private providers, also exhibit a bimodal distribution with modes at 1.1 and -0.5. An institution specific analysis reveals that one (partial) reason for this bimodal distribution is the aggregation of providers in bigger central hospitals and those in small clinics and dispensaries—the mean competence of the former is close to 0.6 while the latter lies 0.4 standard deviations below at 0.2 (we stress that this is only a partial explanation since some central hospitals have very low means as well). This breakdown of the competence density is interesting because it simultaneously satisfies both notions of the health care system in India. The notion that the public sector is a "worse" performer than the private sector is justified by comparisons of the distribution of public doctors to the distribution of MBBS private doctors. On the other hand, the opposite is also justifiable by comparing the distribution of public doctors to that of their non-MBBS private counterparts.

What does this imply for poor people? The choice set of the poor is clearly worse than that of the rich—moving from low to middle-income areas increases average competence by 0.5 standard deviations, and from low to rich by over 1 standard deviation (Table 4, Column 1). This average difference is driven by a number of factors. In line with predictions from a hedonic model of location, providers with less training are overwhelmingly in poor areas with the proportion of MBBS providers more than doubling moving from poor to rich (Table 4, Column 2). In addition, *within* every qualification, less competent providers are in poor areas. Thus, MBBS providers are almost 0.6 standard deviations less competent when in poor compared to the rich areas and a similar result obtains for those without an MBBS as well (Table 4, Column 3 and 4). To illustrate the kind of treatment these competence differentials imply, if a person in a poor neighborhood randomly picks a provider, the mean competence would be about -0.5. This translates into an increased probability of harmful treatment by close to 20 percent for a number of illnesses including tuberculosis and pre-eclampsia.

One option for poor people is to then use government health facilities, where theoretically, providers are assigned independent of competence. Unfortunately, the data do not support this ideal. Public doctors in poor areas are substantially worse than those in rich neighborhoods (Table 4, Column 5). While this is

---

<sup>19</sup>Recall from our previous discussion that MBBS degrees are the equivalent of MDs in the US and apart from such providers our sample includes providers with a variety of other qualifications as well.

particularly true for public providers not based in hospitals where the difference in competence is almost 0.8 standard deviations moving from poor to rich, it also holds true for hospitals (Table 4, Columns 6 and 7). In fact, non-MBBS providers in rich areas are more competent than public MBBS providers in poor areas ( $p < .05$ ).<sup>20</sup>

A final alternative for the poor is then to use hospitals since they systematically outperform all other providers in poor and middle income areas. However, both in data from the parallel household survey as well as the National Sample Survey visits to public hospitals among the poor are minimal. Using data from the National Sample Survey for instance, Mahal and others (2001) find that the proportion of people from the poorest quintile in urban Haryana (the closest equivalent for Delhi that they present) that go to the private sector for outpatient treatments was higher than the proportion in the richest quintile (88 percent versus 84 percent) and the fraction going to public hospitals is a small subset of the total public visits. For the household survey in the seven localities of this study, similar numbers obtain. Although visits to the private sector are less frequent (70 percent) and higher among the rich, public hospitals still accounted for less than 5 percent of out-patient visits.

The regression results largely duplicate the bivariate comparisons. Table 5 shows the results from the OLS regression. Column 1 uses only the area of residence and the institutional affiliation; Column 2 adds on the qualification of the provider and Column 3 adds in the geographical origin and tenure in the neighborhood.<sup>21</sup> The correlation between poverty and competence remains. From Column 3, the size of the coefficient implies a decrease of 0.67 standard deviation in competence moving from the richest to the poorest neighborhood and 0.28 decline for a one standard deviation increase in the percentage of poor households for the sample. Younger providers in the sample have higher competence controlling for other factors (Column 3), suggesting a vintage effect whereby learning by doing gains are offset by gains in the technology of training. There are also no significant interaction effects between neighborhood and tenure (not reported), consistent with sorting of providers into localities. Including the origin of the provider is not particularly informative—providers from Delhi and Uttar Pradesh (the state neighboring Delhi) are significantly more competent and those from Bihar less so (coefficients not reported). What explains these coefficients is a matter of speculation.

The most important correlate of competence is whether the provider holds an MBBS degree or not. A provider with an MBBS degree is at least 0.9 standard deviations more competent in all specifications. Interestingly, these observable features of the provider still leave at least 35 percent of variation in competence unexplained—this result is robust to a number of different specifications that include gender and years of education as well as interactions between locality and other variables.<sup>22</sup> What this implies for the ability

---

<sup>20</sup>It is possible that the observed difference in competence between rich and poor areas reflects differential depreciation in competence across neighborhoods, particularly if providers in poor areas have worse peers and uninformed patients. However, disaggregating providers by tenure (with similar results if age is used instead) shows that there is a worsening of competence levels with tenure in all areas, but the differences between poor and rich neighborhoods remains comparable—among the young this difference is 1.09 standard deviations and it increases marginally to 1.15 standard deviations among the old. These patterns seem more consistent with sorting by providers into localities rather than differential depreciation over time.

<sup>21</sup>Note also that all these variables are observable by both households and surveyors (in fact the origin of the provider is often used as an identifier—"Bengali" doctors in low income neighborhoods are supposedly low quality providers—we find no evidence for this unfortunate assertion). The explained portion of the regression can then also be used a "lower-bound" for the forecasting ability of households.

<sup>22</sup>In addition to the standard  $R^2$ , the table also shows the  $R^2_{MEC}$ , corrected for measurement error, using the estimate of

of households to forecast competence is unclear. This could represent an upper bound if variables like the tenure of the provider are not easily observed; it could also represent a lower bound if communication is an important means for disseminating information.

## 7. Caveats and conclusions

This paper developed a method for measuring clinical competence of medical care providers using Item Response Theory. The method was applied to data collected by the authors on a sample of providers for medical services in Delhi and used to compare rich and poor areas as well as the public and private sectors. The method seems to us (at least) to have promise for future studies of this kind.

The results justify two disparate views on the quality of the public versus the private sector in India. Among providers in Delhi, competence is higher in the private compared to the public sector if the comparison is among those with an MBBS degree (this is consistent with the notion that the private sector is better than the public). Also, competence is lower in the private sector if we restrict private providers to those without an MBBS degree (the private sector is worse than the public). The distribution of competence by neighborhood income confirms that health inequities that arise due to household attributes, such as education, are exacerbated by inequities on the supply side. Private providers in poor neighborhoods are worse than their counterparts in richer areas. Perhaps surprisingly, the same pattern holds among providers in the public sector as well.

The item response method presents some advantages over the (standardized) raw score. First, to the extent that competence is an *estimated* fixed-effect of the provider, the variance of the estimate provides a metric with which to gauge the validity of the instrument. This is useful in contexts where provider behavior may be very different from provider knowledge. Second, the additive form of the information function provides a statistical basis for the choice of cases to be presented. Two cases—tuberculosis and pre-eclampsia—contribute an average of 81 percent of the information contained in the entire module. Thus, unless there is an interest in discriminating between competent and very competent providers (where the informational content of these two cases drops), these two cases are sufficient (with the caveat that easier questions would help reduce errors in the lower end of the distribution). Third, by optimally weighting the responses the competence index provides a more accurate measure of the estimated latent variable.

The distribution of competence obtained by standardizing the raw score is different from that using item response. In particular it is no longer bimodal, the mean score of MBBS providers is lower and the mean score of non-MBBS providers is higher. This is consistent with thinking about the distribution of the standardized

---

the variance of measurement error from the IRT analysis. In a regression context where the dependant variable  $y_i$  is measured with error so that  $y_i = y_i^* + \varepsilon_i$  with  $var(\varepsilon) = \sigma_\varepsilon^2$ ,

$$R_{MEC}^2 = \frac{1 - R^2}{1 + R^2 \frac{\sigma_\varepsilon^2}{\sigma_v^2}} \quad (9)$$

where  $R_{MEC}^2$  is the  $R^2$  corrected for measurement error and we substitute  $\sum \sigma_\varepsilon^2/n$  when, as in our case, the measurement error depends on the  $y_i$ . The  $R_{MEC}^2$  is, as expected, bounded below by  $R^2$  when there is no measurement error and bounded above by 1, when all the residual sum of squares is governed by the measurement error.

raw score as the item response distribution with normally distributed noise added on (recall that the error from all 78 questions will be added to the "true" raw score).

Nevertheless, the use of item response does require some fairly strong identifying assumptions. In particular, the assumptions of IRT include unidimensionality of the underlying measure of competence and no Differential Item Functioning, so that the probability of answering a question in a particular way depends only on the underlying latent measure and not on other variables. While we have evidence that our measure satisfies unidimensionality, we cannot be sure that responses to questions in the index concerning treatment are the same for public and private providers. In some cases (the probability of providing anti-infective drugs for viral pharyngitis) we find differences between treatment patterns after controlling for competence at the 10 percent level (Das and Hammer 2004).<sup>23</sup> It is possible that doctors start thinking the way they practice. Incentives for behavior are certainly different between the sectors. Public doctors on salary face very low-powered incentive schemes while private doctors face very high-powered incentives, depend on retaining clients and perhaps provide more services. These differences in motivation could lead to differences in what providers think is right. In our results, private providers do recommend more proactive treatments than public providers and this could have affected our measure of competence.

---

<sup>23</sup>The statistical test for no DIF uses the invariance of item parameters to compare the difference between the item characteristic curves for two different groups (in our case public and private). Unfortunately, this method does not work with a small sample. Since the test involves a chi-2 test of proportions, even splitting the sample evenly gives only 10 data points for comparison within every group. With such small numbers, the power of the test is small and the null is generically accepted.

## References

- Birnbaum, Allan. 1967. "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In Lord, Frederic M. and M.R. Novick, eds., *Statistical Theories of Mental Test Score*. London: Addison-Wesley Publishing Company.
- Bock, Richard D. and Aitkin, M. 1981. "Marginal Maximum Likelihood Estimation of Item Parameters: An Application of an EM Algorithm". *Psychometrika* 46: 443-459.
- Bock, Richard D. and Lieberman, M.1970. "Fitting a Response Model for n Dichotomously Score Items". *Psychometrika* 35: 179-197.
- Collier, Paul, Stefan Dercon, and John Mackinnon. 2003. "Density versus Quality in Health Care Provision: Using Household Data to Make Budgetary Choices in Ethiopia." *The World Bank Economic Review* 16(3): 425-48.
- Das, Jishnu. 2001. "Do Patients Learn about Doctor Quality? Theory and an Application to India". Ph.D. Dissertation. Harvard University.
- Das, Jishnu, and Carolina Sánchez-Páramo. 2003. "Short but not Sweet: New Evidence on Short Duration Morbidities from India." Policy Research Working Paper 2971. World Bank, Development Research Group, Washington, D.C.
- Das, Jishnu, James Habyarimana, Stefan Dercon and Pramila Krishnan. 2004. "When can School Inputs Improve Test Scores". Policy Research Working Paper 3217. World Bank, Development Research Group. Washington, D.C.
- Das, Jishnu and Jeffrey Hammer. 2004. "Money for Nothing: The Dire Straits of Medical Practice in India". In process.
- Das, Veena, and Ranendra K. Das. 2003. "Pharmaceuticals in urban ecologies: The register of the local." Processed.
- Dragow, F. and R.I. Lissak.1983. "Modified Parallel Analysis: A Procedure for Examining the Latent Dimensionality of Dichotomously Scored Item Responses." *Journal of Applied Psychology*. 68: 363-373.
- Foster, Andrew D. 1995. "Prices, Credit Markets and Child Growth in Low-Income Rural Areas." *The Economic Journal* 105(430): 551-570.
- Hambleton, Ronald K. and H. Swaminathan. 1985. *Item Response Theory: Principles and Applications*. Boston: Kluwer.sdsdsd
- Hambleton, Ronald K., H. Swaminathan and H. Jane Rogers. 1991. *Fundamentals of Item Response Theory*. California: Sage Publications.
- Hattie, J.A. 1985. "Methodological Review: Assessing Unidimensionality of Tests and Items." *Applied Psychological Measurement*. 9: 139-164.

- Jesani, A. ed. *Market Medicine and Malpractice*. Center for Enquiry into Health and Allied Themes and Society for Public Health Awareness and Action. Mumbai. 1997.
- Kakar, D.N. *Primary Health Care and Traditional Medical Practitioners*. New Delhi: Sterling Publishers. 1988.
- Lavy, Victor, and Jean-Marc Germain. 1994. "Quality and Cost in Health Care Choice in Developing Countries." Working Paper 105. Living Standards Measurement Study, World Bank, Washington, D.C.
- Leonard, Kenneth L. 2003. "African Traditional Healers and Outcome-Contingent Contracts in Health Care." *Journal of Development Economics* 71(1).
- Leonard, Kenneth L., and Melkiory C. Masatu. 2003. "Comparing Vignettes and Direct Clinical Observation in a Developing Country Context." University of Maryland, College Park, Maryland. Processed.
- Mahal, Ajay, Abdo Yazbek, David Peters and G.N.V. Ramana. 2001. "The Poor and Health Service Use in India". Processed.
- McFadden, D. 1973. Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka, ed., *Frontiers in Econometrics*. New York: Academic Press.
- John W. Peabody, Jeff Luck, Peter Glassman, Timothy R. Dresselhaus and Martin Lee. 2000. "*Comparison of Vignettes, Standardized Patients, and Chart Abstraction. A Prospective Validation Study of 3 Methods for Measuring Quality.*" *Journal of the American Medical Association*. 283, pp. 1715-1722.
- Reddy, K.N. and V. Selvaraju. 1994. "Health Care Expenditure by Government in India: 1974-75 to 1990-91." National Institute of Public Finance and Policy. New Delhi. Processed.
- Rethans, J. J., F. Sturmans, R. Drop, C. P. M. van der Vleuten, and P. Hobus. 1991. "*Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice.*" *British Medical Journal* 303 (6814): 1377-80.
- Rohde, J.E. and H.Viswanathan. *The Rural Private Practitioner*. Delhi: Oxford University Press. 1995.
- Samejima, F. 1972. "Estimation of Latent Ability using a Response Pattern of Graded Scores." Psychometric Monograph No. 17. Iowa City: Psychometric Society.
- World Bank. 2003. World Development Report 2004: Making Service Work for Poor People. Washington, D.C.: World Bank.
- World Health Organization. 1978. "The Alma-Ata Declaration." Geneva: World Health Organization.
- Yen, W.M. 1981. "Use of the Three-Parameter Logistic Model in the Development of a Standardized Achievement Test." In R.K. Hambleton, ed., *Applications of Item Response Theory*. Vancouver, British Columbia: Educational Research Institute of British Columbia.

## Tables and Figures

**Table 1: Frequency of Cases in Vignettes Module**

Frequency of Patients Presenting (%)	Vignettes Case				
	Case I: Diarrhea	Case II: Viral Pharyngitis	Case III: Tuberculosis	Case IV: Depression	Case V: Pre-Eclampsia
Almost Every Day	70.79	87.62	26.11	15.42	8.46
Almost Every Week	80.20	97.52	56.65	40.30	26.87
At Least once a month	86.14	99.01	90.15	74.13	59.20
Once a year	88.61	99.01	97.04	94.03	85.57
Never seen such a case	12.4	0.99	2.96	5.97	14.43

*Source:* ISERDD-World Bank (2003). This table shows the frequency with which patients present with the illnesses covered in the vignettes. Percentages are based on reports by the providers in the sample to the question “How often do you see such a case in your clinic?”.

**Table 2: Competence and True Scores**

Quintiles of Competence	Mean Competence	Mean True Score
Least Competent	-1.61	16
Second Quintile	-0.66	19.68
Average Competence	-0.10	23.45
Fourth Quintile	0.73	33.17
Most Competent	1.47	48.82

*Note:* Author's calculations based on World Bank-ISERDD (2003). Competence refers to the value of the latent variable estimated through item response. The true score is a monotonic non-linear transformation of the latent variable that shows the *expected* number of questions that the provider would ask in the vignettes, where the expectation is taken over (hypothetical) repeated administrations of the same test.

**Table 3: Predicted Probability of Non-Harmful Treatment (%)**

Quintiles of Competence	Case Presented				
	Diarrhea	Viral Pharyngitis	Tuberculosis	Depression	Pre-Eclampsia
Least Competent	18	27	56	38	41
Second Quintile	21	37	70	44	49
Average Competence	25	45	77	48	55
Fourth Quintile	34	59	86	58	65
Most Competent	47	71	91	68	74
Guessing Probability	17	14	20	34	28

*Note:* Author's calculations based on World Bank-ISERDD (2003). All numbers are percentages. The predicted probability of non-harmful treatment is based on the item characteristic curves for the treatment grades. The treatment grades were assigned independently on a scale of -3 to +3 by three different raters, two in South Asia and one team from Johns Hopkins University School of Medicine. For this exercise the grades are treated as dichotomous variables. The inter-rater agreement using the kappa measure of agreement exceeds 80 percent for comparisons between the South-Asian doctors and the team from Johns Hopkins and exceeds 90 percent for the South-Asian raters. Non-harmful treatment is defined as the following:

1. Diarrhea: Not advising fluid intake, prescribing antibiotics or antidiarrheals.
2. Viral Pharyngitis: Advising medications other than those for symptomatic relief (such as analgesics)
3. Tuberculosis: Not referring and not prescribing multi-drug therapy.
4. Depression: Not referring and prescribing medications unrelated to the case.
5. Pre-eclampsia: Not referring the patient.

**Table 4: Inequities in the Distribution of Competence**

Income Group	Average Competence	% MBBS Providers	Average Competence (MBBS Only)	Average Competence (Non-MBBS Only)	Average Competence (All public providers including hospitals)	Average Competence (Public, non-hospitals only)	Average Competence (Hospitals Only)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Low-Income	-0.45	35	0.22	-0.82	-0.11	-0.57	0.36
Middle-Income	-0.05	54	0.43	-0.62	0.59	0.62	0.56
High-Income	0.56	78	0.81	-0.32	0.54	0.39	0.61

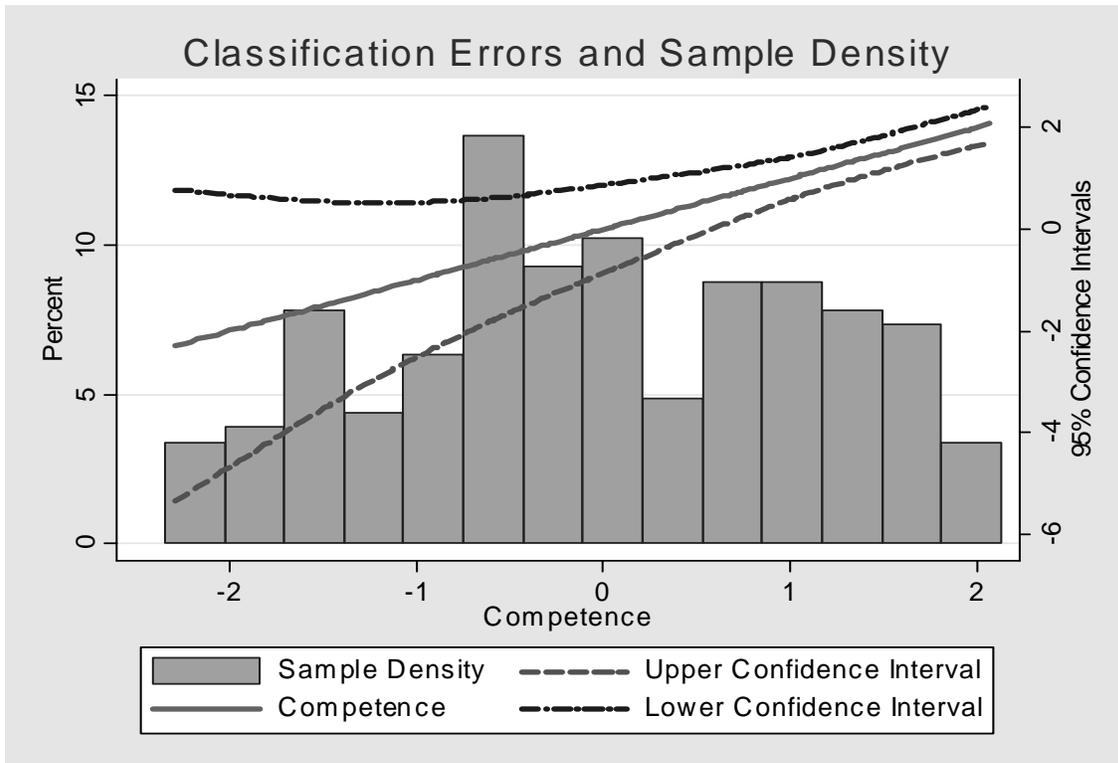
*Note:* Author's calculations based on World Bank-ISERDD (2003). The table shows average competence in different neighborhoods where competence refers to the competence index derived in the text. There are seven neighborhoods in the study and these were assigned as low, middle and high income neighborhoods on the basis of consumption aggregates from the abbreviated NSS consumption schedule administered to the households in the household survey. Of the seven areas in the survey, three were assigned as low-income, two as middle-income and two as high-income. Dividing households into three groups (poor, middle and rich) shows that the percentage of poor (middle) households in the low income areas are 74% (25%), 65% (25%) and 57% (33%); in the middle income areas 15% (73%) and 0% (46%); and in the high income areas 0% (0%) and 8% (21%). Hospitals were assigned to income groups depending on their location.

**Table 5: Correlates of Competence**

	(1)	(2)	(3)
	Income and Institution	Income, Institution and Qualification	All variables
% of poor households in community	-0.013 (0.002)***	-0.007 (0.002)***	-0.009 (0.002)***
% middle income households in community	-0.011 (0.003)***	-0.006 (0.003)*	-0.004 (0.003)
Public Doctor	0.451 (0.172)***	-0.185 (0.175)	-0.174 (0.171)
MBBS Degree Holder		1.132 (0.153)***	0.957 (0.156)***
Tenure in Locality			-0.020 (0.007)***
Controls for Origin of Provider?	NO	NO	YES
Constant	0.616 (0.154)***	-0.194 (0.175)	-0.399 (0.255)
Observations	204	204	194
R-squared	0.19	0.36	0.44
R-squared corrected for measurement error	0.28	0.53	0.64

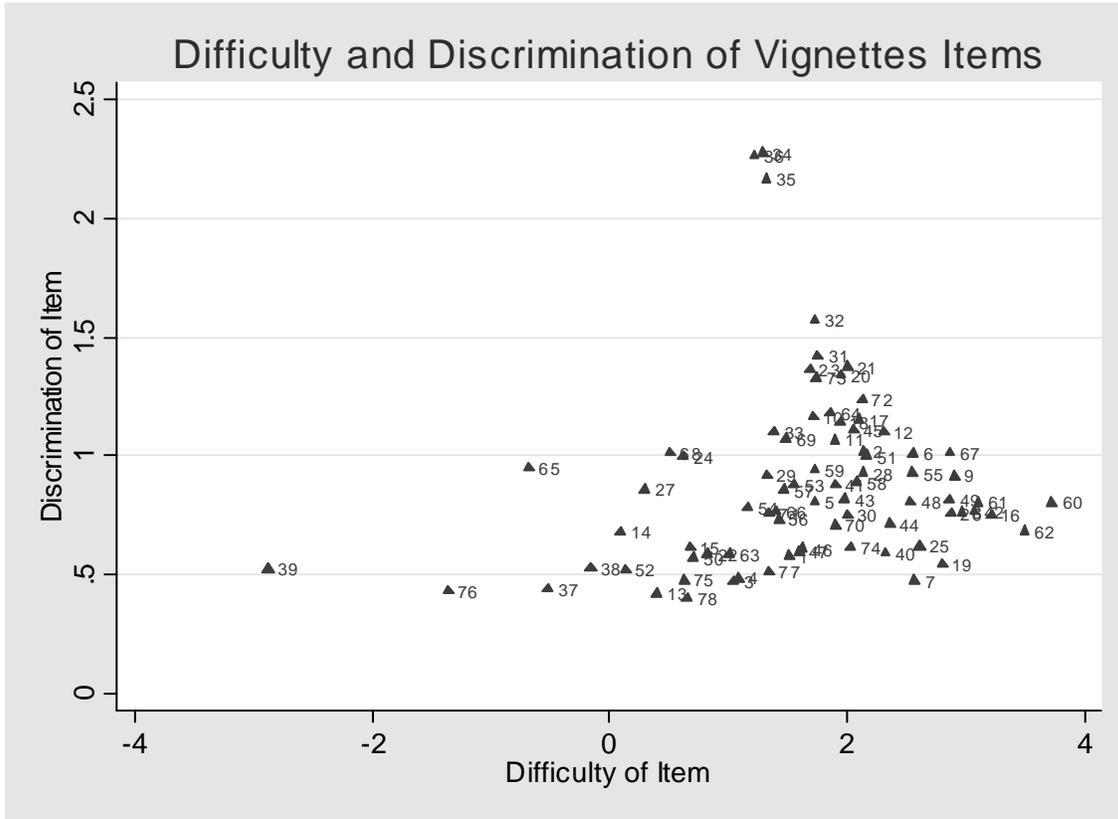
Standard errors in parentheses \* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%. The % poor and % middle income households in the community is computed from a consumption survey of 40 randomly selected households through the parallel household survey. The attributes of these households are detailed in Das and Sánchez (2002) who shows that the households do not differ substantially from the random samples in Delhi of the NSS and the NFHS. For comparability, the consumption questionnaire administered was the abbreviated NSS consumption questionnaire. Public Doctor refers to whether the doctor practices in the public or the private sector. MBBS degree holder is a dummy variable indicating whether the provider holds an MBBS degree. The tenure in the locality is the number of years that the provider has practiced in that particular clinic, from data on the census of providers. The “price charged” is the reported price by the provider when asked about his/her fees. The origin of the provider is based on the census of providers and the R-squared corrected for measurement error accounts for the error of classification of the test. The method is detailed in the text.

**Figure 2: Distribution of Competence and Standard Errors of Competence Index**



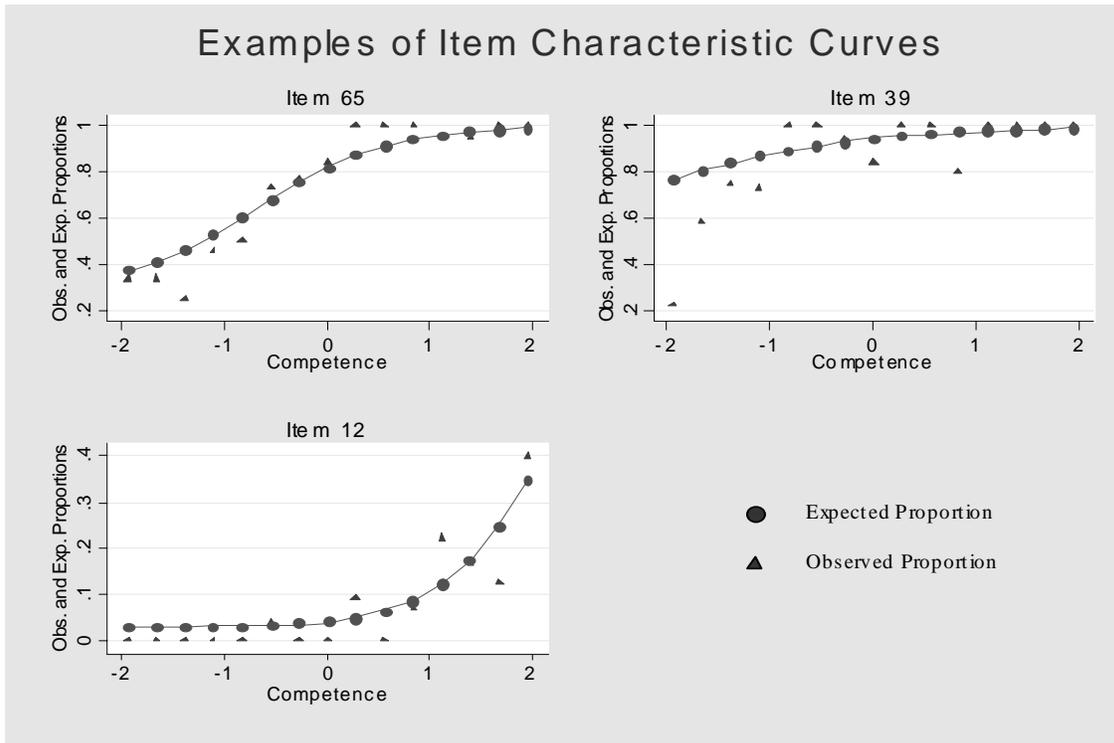
*Source:* Author's calculations based on World Bank-ISERDD (2003). The horizontal axis in the graph is competence, the left vertical axis is the density (in percentages) for the histogram of competence and the right vertical axis shows confidence intervals of competence. The solid line is estimated competence, which is plotted against itself (this would be the 45° line if the scales were the same). The two dashed curves represent the upper and lower confidence intervals at the 95% level of confidence. The histogram underlying the confidence interval curves shows how competence is distributed at values of the index with large and small standard errors.

**Figure 3: Difficulty and Discrimination of Vignettes Items**



*Source:* Author’s calculations based on World Bank-ISERDD (2003). This figure shows the distribution of difficulty and discrimination of all the items used for estimating competence. The item numbers are displayed next to the scatter plot—Appendix 1, Table A details all the items used. The figure shows that there are only 5 items with difficulty below the mean, and of these only one (item 65) has some discriminatory power.

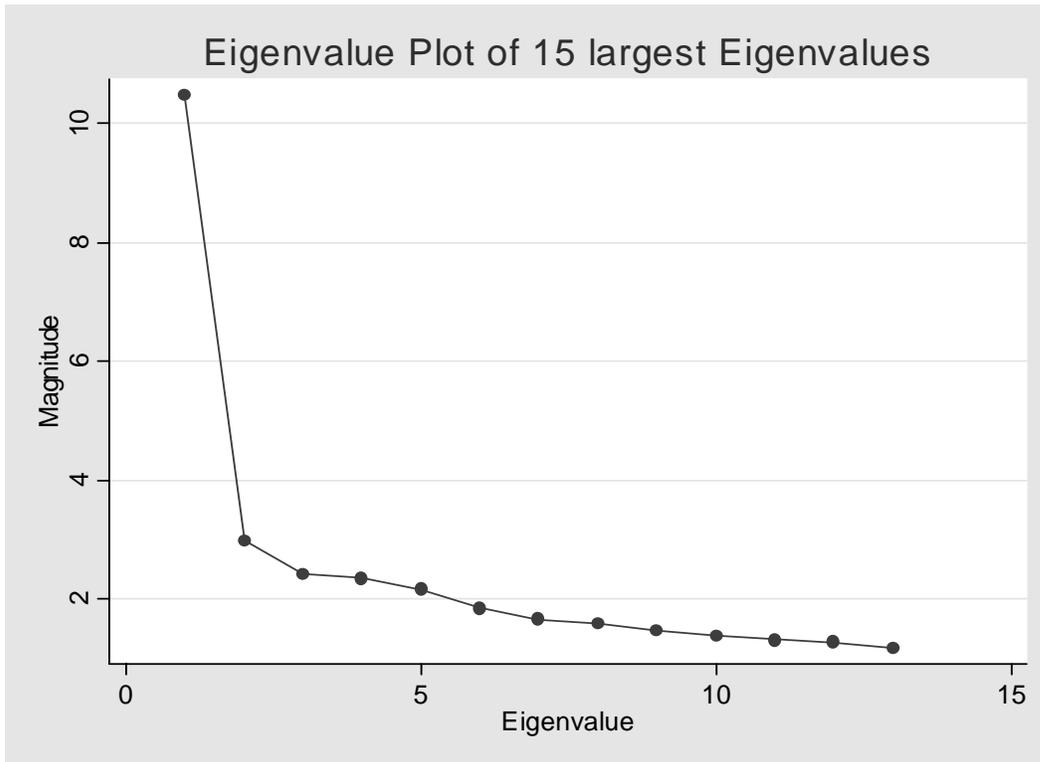
**Figure 4: Specific Item Characteristic Curves**



*Source:* Author's calculations based on World Bank-ISERDD (2003). This figure shows the predicted and observed responses for specific items, plotted against competence. The dots represent expected proportions based on the 3 parameter logistic and triangles represent observed proportions in the data. For this exercise providers were grouped in 15 evenly space groups and averages were taken over provider responses in each group. The items correspond to the following:

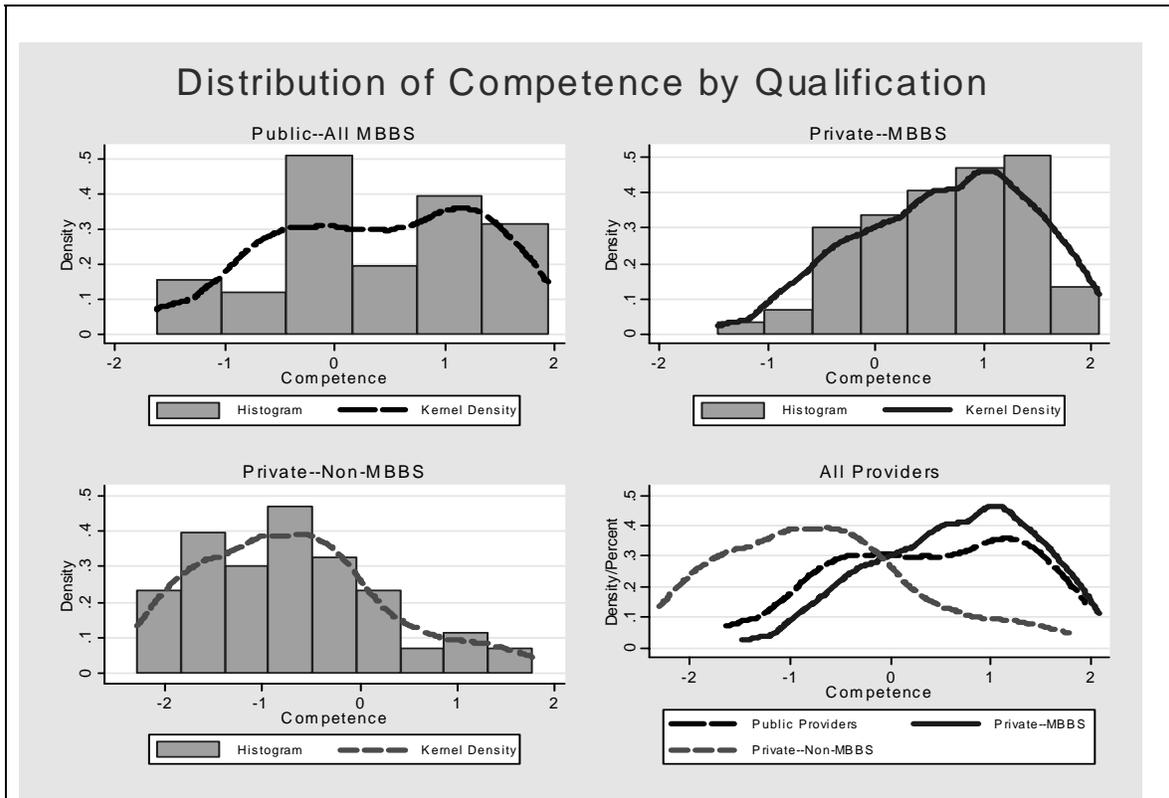
1. Item 65: Positive responses are recorded for item 65 if the provider checks whether a woman with pre-eclampsia has high blood pressure.
2. Item 39: The provider asks a man with tuberculosis to have a chest X-ray
3. Item 12: The provider checks for depression in the skull fontainele in a child with diarrhea.

**Figure 5: Unidimensionality of the Vignettes**



*Source:* Author's calculations based on World Bank-ISERDD (2003). This figure shows the 15 largest eigenvalues to determine whether there is a dominant first factor. The first eigenvalue is almost 5 times larger than the next and the 2<sup>nd</sup> is almost indistinguishable from those following, suggesting that the variation in responses is driven by a single latent variable.

**Figure 6: The Distribution of Competence**



*Source:* Author's calculations based on World Bank-ISERDD (2003). This figure shows the distribution of competence disaggregated by the qualification of the provider. In the first three figures, histograms are presented as well due to the small sample size. For qualifications, the following categories are used, where MBBS refers to Bachelor of Medicine and Bachelor of Surgery, the Indian equivalent of an MD in the United States:

1. **Public-All MBBS:** Providers in the public sector. These providers can be either in public hospitals or in smaller health care centers and dispensaries. All the public providers in this sample hold MBBS degrees, although there are a limited number of public providers who hold other degrees in public hospitals.
2. **Private-MBBS:** All providers in the private sector who hold a MBBS degree.
3. **Private—Non-MBBS:** All providers in the private sector who do not hold a MBBS degree. These could be providers who have formal training and government recognition (such as degrees in alternate medicine such as homeopathy or Ayurveda); providers who have no formal training, but do have government recognition and those who have neither formal training nor government recognition.

## APPENDIX A

**Table A: List of Questions Used**

Item Number	Case	Type of Question	Question
Item 1	Diarrhea	History	Has the child had fever?
Item 2	Diarrhea	History	When did the child last urinate?
Item 3	Diarrhea	History	Has the child had any vomiting?
Item 4	Diarrhea	History	Frequency of Stools
Item 5	Diarrhea	History	Blood/Mucous in Stool
Item 6	Diarrhea	Examination	Pulse Rate
Item 7	Diarrhea	Examination	Temperature
Item 8	Diarrhea	Examination	Mucous membranes for moistness
Item 9	Diarrhea	Examination	Tear Ducts for tears
Item 10	Diarrhea	Examination	Skin Color and Turgor
Item 11	Diarrhea	Examination	Palpation of Abdomen
Item 12	Diarrhea	Examination	Depression of skull fontanelle
Item 13	Viral Pharyngitis	History	Length of Illness
Item 14	Viral Pharyngitis	History	History of fever?
Item 15	Viral Pharyngitis	History	History of cough?
Item 16	Viral Pharyngitis	History	History of chest pain?
Item 17	Viral Pharyngitis	History	History of shortness of breath?
Item 18	Viral Pharyngitis	History	Color of Expectoant
Item 19	Viral Pharyngitis	History	History of headache?
Item 20	Viral Pharyngitis	Examination	Pulse Rate
Item 21	Viral Pharyngitis	Examination	Blood Pressure
Item 22	Viral Pharyngitis	Examination	Temperature
Item 23	Viral Pharyngitis	Examination	Nose and nasal passages
Item 24	Viral Pharyngitis	Examination	Throat Inspection
Item 25	Tuberculosis	History	History of night sweats?
Item 26	Tuberculosis	History	History of chest pain?
Item 27	Tuberculosis	History	History of blood in sputum?
Item 28	Tuberculosis	History	Whether this has happened before?
Item 29	Tuberculosis	History	Has this happened to others in the dwelling?
Item 30	Tuberculosis	History	Patient's profession
Item 31	Tuberculosis	Examination	Pulse Rate
Item 32	Tuberculosis	Examination	Blood Pressure
Item 33	Tuberculosis	Examination	Temperature
Item 34	Tuberculosis	Examination	Chest Inspection
Item 35	Tuberculosis	Examination	Chest Percussion
Item 36	Tuberculosis	Examination	Auscultation

Item Number	Case	Type of Question	Question
Item 37	Tuberculosis	Examination	Blood for TLC/DLC
Item 38	Tuberculosis	Examination	Sputum for AFB
Item 39	Tuberculosis	Examination	Chest X-ray
Item 40	Depression	History	General mood
Item 41	Depression	History	Does patient feel like crying?
Item 42	Depression	History	Why does the patient feel this way?
Item 43	Depression	History	Is there trouble sleeping?
Item 44	Depression	History	Hours of sleep
Item 45	Depression	History	Time of waking up
Item 46	Depression	History	Problems in daily work
Item 47	Depression	History	Whether any deaths/changes in the family
Item 48	Depression	Examination	Pulse Rate
Item 49	Depression	Examination	Blood Pressure
Item 50	Depression	Examination	Menstrual history
Item 51	Depression	Examination	Thyroid Glands
Item 52	Pre-Eclampsia	History	Last Menstrual Period
Item 53	Pre-Eclampsia	History	History of nausea or vomiting?
Item 54	Pre-Eclampsia	History	Whether there has been swelling in the feet
Item 55	Pre-Eclampsia	History	Whether she has felt any foetal movement
Item 56	Pre-Eclampsia	History	Whether an ante-natal check-up was done
Item 57	Pre-Eclampsia	History	Number of Children
Item 58	Pre-Eclampsia	History	Type of birth
Item 59	Pre-Eclampsia	History	Whether any other pregnancies
Item 60	Pre-Eclampsia	History	Whether she has taken any immunizations
Item 61	Pre-Eclampsia	History	Whether she has had an ultrasound
Item 62	Pre-Eclampsia	History	Whether she has a history of hypertension
Item 63	Pre-Eclampsia	History	Severity and Frequency of headaches
Item 64	Pre-Eclampsia	Examination	Pulse Rate
Item 65	Pre-Eclampsia	Examination	Blood Pressure
Item 66	Pre-Eclampsia	Examination	Edema in Feet
Item 67	Pre-Eclampsia	Examination	Examination of eyes and mouth for anemia
Item 68	Pre-Eclampsia	Examination	Proteinuria
Item 69	Pre-Eclampsia	Examination	Urine for glucose
Item 70	Pre-Eclampsia	Examination	Blood Glucose
Item 71	Pre-Eclampsia	Examination	Blood Hb
Item 72	Pre-Eclampsia	Examination	Palpation of fetus
Item 73	Pre-Eclampsia	Examination	Fetal Heart Rate
Item 74	Diarrhea	Overall Treatment Score	Overall treatment score
Item 75	Viral Pharyngitis	Overall Treatment Score	Overall treatment score
Item 76	Tuberculosis	Overall Treatment Score	Overall treatment score

Item Number	Case	Type of Question	Question
Item 77	Depression	Overall Treatment Score	Overall treatment score
Item 78	Pre-Eclampsia	Overall Treatment Score	Overall treatment score

**Table B: Questions not used in analysis due to low responses**

Case Number	Type	Question	Percentage who asked
Diarrhea	Examination	Respiration Rate	2.4
	Examination	Blood for serum electrolytes	0.97
	Examination	Blood for TLC/DLC	3.4
Viral Pharyngitis	History	Has the patient experienced chills?	3.9
	History	Has the patient experienced excessive sweating	2.4
	History	Is there blood in the expectorant?	3.4
	Examination	Respiratory Rate	4.8
Tuberculosis	History	Riskiness of patient's sexual behavior	0.9
	Examination	Respiratory Rate	0.48
		Blood for HIV	3.9
Depression	History	Does the patient feel scared/anxious?	4.8
		Has the patient every thought of suicide?	3.9
		Has the patient every been pregnant/had an abortion?	2.4
		Does the patient take any drugs/alcohol/medication?	1.4
	Examination	Respiratory Rate	0.48
Pre-Eclampsia	History	Weight gain during pregnancy	4.4
		Bleeding or Discharge during pregnancy	1.9
		History of diabetes?	1.4
		History of anemia?	1.4

Case Number	Type	Question	Percentage who asked
		History of heart disease?	0.97
		History of any genetic disease?	0.48
		History of smoking/drinking	0
		Experiences shortness of breath?	0.97
	Examination	Respiratory Rate	1.9
		Auscultation of Chest	4.4