J. R. Statist. Soc. A (2012)
175, *Part* 2, *pp.* 1–21

Vignettes and health systems responsiveness in cross-country comparative analyses

Nigel Rice and Silvana Robone

University of York, UK

and Peter C. Smith

Imperial College Business School, London, UK

[*Read before* The Royal Statistical Society *at the 58th World Statistics Congress of the International Statistical Institute in Dublin on Tuesday, August 23rd, 2011, the President*, Professor V. S. Isham, *in the Chair*]

Summary. The paper explores the use of anchoring vignettes as a means to adjust survey reports of health system performance for differential reporting behaviour by using data contained within the World Health Survey. Survey respondents are asked to rate their experiences of health systems across a number of domains on a five-point categorical scale. Using data provided through a set of vignettes we investigate variations in reporting of interactions with health services across both sociodemographic groups and countries. We show how the method of anchoring vignettes can be used to enhance cross-country comparability of performance. Our results show large changes in the rankings of country performance once adjustment for systematic country level reporting behaviour has been undertaken compared with a ranking based on raw unadjusted data.

Keywords: Anchoring vignettes; Cross-country comparison; Healthcare responsiveness; Health system performance; Hierarchical ordered probit

1. Introduction

Increasingly patients' views and opinions are being recognized as an essential means for assessing the provision of health services, to stimulate quality improvements and, more recently, in measuring health systems performance (Coulter and Magee, 2003). Although traditionally patients' views have been sought on the quality of care provided and satisfaction with health services, the World Health Organization has proposed the concept of responsiveness as a more desirable measure by which health systems can be judged (Valentine *et al.*, 2003a). Responsiveness relates to a system's ability to respond to the legitimate expectations of potential users about non-health aspects of care and, together with health and fairness of financial contribution, has been suggested as an intrinsic goal of health system performance (Murray and Frenk, 2000). In broad terms, health system responsiveness has been defined as the way in which individuals are treated and the environment in which they are treated. Importantly it encompasses the notion of an individual's experience of contact with the health system (Valentine *et al.*, 2003a).

A central purpose for measuring outcomes, such as health system responsiveness, is to enable institutions to compare and contrast their performance with that of others, including performance

Address for correspondence: Nigel Rice, Centre for Health Economics, Alcuin Block A, University of York, Heslington, York, YO10 5DD, UK.
E-mail: nigel.rice@york.ac.uk

 0964–1998/12/175000

obtained in other countries. By establishing relevant benchmarks, a cross-national comparison offers the opportunity for countries to assess their place in relation to others, to learn from experience elsewhere and to identify and explore trends in performance (O'Mahony and Stevens, 2004; Gonzalez Block, 1997). The question, however, of how appropriately to compare across countries with different institutional settings and populations is a central challenge for comparative work across all public services. Studies that are aimed at comparative inference have rarely taken into consideration possible variations in cultural expectations that might influence the reporting behaviour of surveyed respondents (Blendon *et al.*, 2003). Attempts to enhance cross-country comparison have tended to focus on defining objective measures of desired outcomes and developing survey instruments that are relevant and understandable across cultural settings (e.g. Lynn *et al.* (2006), Okazaki and Sue (1995), Brislin (1986) and Murray *et al.*, (2003)). In itself this is, however, unlikely to ensure comparability of response if individuals in different populations or subgroups when faced with survey questions about the functioning of health systems systematically differ in their interpretation of the available response categories, such as 'poor' or 'good' performance (Sadana *et al.*, 2002). Where this is so then a fixed level of underlying performance is unlikely to be rated equally across populations of interest (see Tandon *et al.* (2003)) and accordingly cross-population comparison may produce misleading assessments of relative performance. This differential mapping from the underlying latent construct of interest (objective performance) to the available survey response categories is a source of reporting heterogeneity and has been variously described as state-dependent bias (Kerkhofs and Lindeboom, 1995), scale-of-reference bias (de Groot, 2000), response category cut point shift (Sadana *et al.*, 2002) and differential item functioning (King *et al.*, 2004; Kapteyn *et al.*, 2007).

The degree to which self-reported survey data are comparable across individuals, socioeconomic groups or populations has been debated extensively, usually with regard to measures of health status (e.g. Jürges (2007), Bago d'Uva *et al.* (2008), Lindeboom and van Doorslaer (2004), Iburg *et al.* (2002), Manderbacka (1998), Kempen *et al.* (1996), Kerkhofs and Lindeboom (1995) and Idler and Kasl (1995)) and health-related disability (Kapteyn *et al.*, 2007). Similar concerns extend to self-reported survey data on aspects of health system performance, e.g. the responsiveness of the system, where the characteristics of survey respondents and cultural norms regarding the use and experiences of public services are likely to be influential in shaping an individual's responses.

Recently, the method of anchoring vignettes has been promoted as a means for controlling for systematic differences in preferences and norms when responding to survey questions (for example, see King *et al.* (2004)). Vignettes represent hypothetical descriptions of fixed levels of a latent construct, such as responsiveness. If we consider a categorical reporting scale varying from 'very bad' to 'very good', then reporting behaviour results from individuals applying different response thresholds, when mapping underlying performance on a latent scale to the ordinal response categories. Since the vignettes are fixed and predetermined, any systematic variation across individuals in the rating of the vignettes can be attributed to differences in reporting behaviour. Accordingly, responses to the vignette questions allow the response thresholds, or cut points, to be modelled as a function of the sociodemographic characteristics of respondents. Since individuals are asked to evaluate the vignettes in the same way as they evaluate their own experiences, this information can then be used subsequently to adjust the self-reported data of a respondent's own contact with health services. For within-country analyses, by applying the thresholds that are observed for a typical respondent (e.g. the average) as a benchmark, responses of other individuals can be rescaled, or anchored, to provide adjusted comparable data. Similarly for cross-country analyses, responses can be rescaled to a chosen benchmark country to aid comparison.

Various studies have applied the vignette approach and made use of what has been termed the hierarchical ordered probit (HOPIT) model to adjust self-reported data for systematic differences in respondents' use of threshold values. The method has mostly been applied to self-reported data on health status (for example see Iburg *et al.* (2002), Tandon *et al.* (2003), Murray *et al.* (2003), King *et al.* (2004) and Bago d'Uva *et al.* (2008)). More recently, there have been attempts to extend the methodology to health systems performance, e.g. Valentine *et al.* (2003b) using the World Health Organisation Multi-country Survey responsiveness module, Sirven *et al.* (2008) using data from the Survey of Health, Ageing and Retirement in Europe and Puentes Rosas *et al.* (2006) using a survey of user satisfaction in Mexico. Rice *et al.* (2010a) illustrated the issues of cross-country comparison of public sector performance by using a less robust specification of the HOPIT model than in this paper. Our paper complements and extends this literature in considering the performance comparison issue, why it is important and how information that is extracted from vignettes can be used to enhance both within- and across-country comparability of health system performance. We describe both non-parametric and parametric approaches to adjusting self-reported data and apply the latter to an analysis of performance across 54 countries by drawing on data from the World Health Survey (WHS). By benchmarking reporting behaviour to that observed within a selected country, we evaluate whether differential reporting behaviour affects cross-country rankings of health system responsiveness. This is undertaken for countries stratified by levels of income as defined by the United Nation's human development index (HDI). Our findings suggest that reporting of health system responsiveness varies both within and across countries, and our estimation exercise illustrates how reporting heterogeneity affects cross-country rankings of responsiveness.

2. Health system responsiveness

The concept of responsiveness as a measure of health systems performance was developed and promoted by the World Health Organization. The concept covers a set of non-clinical and non-financial dimensions of quality of care that reflect respect for human dignity and interpersonal aspects of the care process (Valentine *et al.*, 2009). Human rights include concepts such as respecting patients' autonomy and dignity, whereas interpersonal aspects of care, or client orientation, focus on aspects that are commonly expressed as hotel facilities, e.g. the quality of basic amenities. These are measured across eight domains chosen to reflect the goals for healthcare processes and systems valued highly by individuals in their contact with health systems. The domains are *autonomy*, *choice*, *clarity of communication*, *confidentiality of personal information*, *dignity*, *prompt attention*, *quality of basic amenities* and *access to family and community support*. Definitions of these domains together with examples of the questions that are asked to survey respondents are provided in Table 1.

Increasingly patients' views and opinions are being recognized as the appropriate source of information on non-technical aspects of the healthcare process and accordingly measurement of health system responsiveness is based on surveys of users' views. In principle, the concept covers both interactions with health services together with broader experiences and interactions with health systems, including, for example, health promotion campaigns and public health interventions (Valentine *et al.*, 2009). Respondents are asked to rate their most recent (in the previous year) experience of contact with the health system within each of the eight domains. The response categories available are 'very good', 'good', 'moderate', 'bad' and 'very bad'. Responsiveness is viewed as a multi-dimensional concept, with each domain measured as a categorical variable for which there is an assumed underlying latent scale.

Table 1. Domains of responsiveness†

Autonomy: respect of patients' views of what is appropriate and allowing the patient to make informed choices
Choice: an individual's right or opportunity to choose a health care institution and health provider and to secure a second opinion and access specialist services when required
Clarity of communication: clear explanation to patients and family the nature of the illness, details of treatment and available options
Confidentiality of personal information: privacy in the environment in which consultations are conducted and the concept of privileged communication and confidentiality of medical records
Dignity: the ability of patients to receive care in a respectful, caring and non-discriminatory setting
Prompt attention: the ability to access care rapidly in the case of emergencies, or readily with short waiting times for non-emergencies
Quality of basic amenities: the physical environment and services often referred to as 'hotel facilities', including clean surroundings, regular maintenance, adequate furniture, sufficient ventilation and enough space in waiting rooms
Access to family and community support: the extent to which patients have access to their family and friends when receiving care and the maintenance of regular activities (e.g. the opportunity to carry out religious and cultural practices)
Example questions used in the WHS to measure responsiveness
Autonomy: how would you rate your experience of being involved in making decisions about your healthcare of treatment?
Choice: how would you rate the freedom you had to choose the healthcare providers that attended to you?
Communication: how would you rate your experience of how clearly healthcare providers explained things to you?
Confidentiality: how would you rate the way your personal information was kept confidential?
Dignity: how would you rate the way your privacy was respected during physical examinations and treatments?
Quality of basic amenities: how would you rate the cleanliness of the rooms inside the facility, including toilets?
Prompt attention: how would you rate the amount of time you waited before being attended to?
Access to family and friends: how would you rate the ease of having family and friends visit you?

†The eight domains of responsiveness defined by the World Health Organization (see Valentine *et al.* (2003a) for a full exposition of these domains). The table provides examples only and not an exhaustive list of questions for each domain. The response categories that were available to respondents were 'very good', 'good', 'moderate', 'bad' and 'very bad'.

3. World Health Survey

The most ambitious attempt to date to measure and compare health systems responsiveness is the WHS. The WHS is an initiative that was launched by the World Health Organization in 2001 aimed at strengthening national capacity to monitor critical health outputs and outcomes through the fielding of a valid, reliable and comparable household survey instrument (see Üstün *et al.* (2003)). 70 countries participated in the WHS 2002–2003, consisting of a combination of 90-min in household interviews (53 countries), 30-min face-to-face interviews (13 countries) and computer-assisted telephone interviews (four countries). All surveys were drawn from nationally representative frames with known probability resulting in sample sizes of between 600 and 10 000 respondents across the countries that were surveyed. Samples have undergone extensive quality assurance procedures, including the testing of the psychometric properties of the responsiveness instrument (for example, see Valentine *et al.* (2009)).

The WHS responsiveness module has been developed from an extensive consultation process aimed at gathering information on the aspects of the delivery of healthcare that individuals value most. The resulting instrument was field tested in the World Health Organization's Multi-country Survey Study on health and responsiveness (2000–2001) (see Üstün *et al.* (2003)) and a refined version of the study's module was incorporated in the WHS. The WHS responsiveness module gathers basic information on healthcare utilization for both in-patient and out-patient services. Here we focus exclusively on in-patient services. The data contain information on the

importance that respondents place on each of the eight domains in the responsiveness section of the WHS. For brevity we present analyses for the following four domains: dignity, confidentiality, quality of facilities and clarity of communication. These domains are considered most important by respondents in Mexico, which is the country that is used to illustrate reporting behaviour. Two items are rated by respondents for each of the domains.

The WHS contains information on individual characteristics and we make use of age, gender, level of education and income. Level of education is measured as both a categorical variable containing seven categories representing, for example, 'primary school completed' and 'secondary school completed' to 'postgraduate degree completed' and a continuous variable measuring the number of years in education. Income is derived from a measure of permanent income based on information on the physical assets that are owned by households. The approach to its measurement has been described by Ferguson *et al.* (2003). In our analysis we construct dummy variables to indicate the tertiles of the within-country distribution of household permanent income to which individuals belong, with the first income tertile considered as the base category. Accordingly, reporting behaviour is assumed to be influenced by the relative position within a country's income distribution rather than its absolute level. The above variables have been extensively used in studies investigating reporting bias in self-reported measure of health (Bago d'Uva *et al.*, 2008; Iburg *et al.*, 2002; Murray *et al.*, 2003; Valentine *et al.*, 2003b) and health-related disability (Kapteyn *et al.*, 2007) and are similarly likely to influence the reporting of health service responsiveness.

The WHS also contains vignettes describing the experiences of hypothetical individuals within each of the eight domains of responsiveness. The vignettes have been divided into four sets (sets A–D) with each set containing five vignettes for each item present within two domains. For example, set A contains five vignettes for each of the two items in the domain of dignity (the items represent respect and privacy) and five vignettes for each of the two items in the domain prompt attention (items representing travelling time and waiting time). Owing to constraints of length of interview, each respondent in the survey was asked to rate the vignettes contained in one of the sets only. Accordingly, each set (and hence each vignette) was rated by approximately 25% of survey respondents. The response scale that was available to respondents answering the vignettes was the same as the scale that was available when responding to their own experiences of health system responsiveness. Examples of the vignettes are provided in Table 2. The fact that not all respondents answer all vignettes does not present a problem for the modelling approach that we adopt.

4. Empirical models

The reporting of responsiveness is via an ordered categorical variable that is assumed to be a discrete representation of some underlying latent scale. If it can be assumed that individuals map the latent scale to the response categories in a consistent way, irrespective of their characteristics or circumstances, then we observe homogeneous reporting behaviour. In these circumstances the standard ordered probit estimator, which assumes a set of constant thresholds in the mapping of the latent scale to the response categories, would provide an appropriate method to model the data. In contrast, reporting heterogeneity, or differential reporting behaviour, arises when individuals differ in the positioning of thresholds when mapping the latent construct to the response categories available. To aid comparison across individuals methods to adjust the data to reflect the differential positioning of the thresholds are required. In this section we begin by briefly reviewing non-parametric approaches to adjusting for differential reporting behaviour before describing the parametric approach that is adopted in our empirical analysis.

Table 2. Examples of vignette questions used in the WHS†

Respectful treatment
'[Anya] took her baby for a vaccination. The nurse said hello but did not ask for [Anya's] or the baby's name. The nurse also examined [Anya] and made her remove her shirt in the waiting room.
Q1: How would you rate her experience of being greeted and talked to respectfully?
Q2; How would you rate the way her privacy was respected during physical examinations and treatments?'
Communication
'[Rose] cannot write or read. She went to the doctor because she was feeling dizzy. The doctor didn't have time to answer her questions or to explain anything. He sent her away with a piece of paper without telling her what it said.
Q1: How would you rate her experience of how clearly health care providers explained things to her?
Q2: How would you rate her experience of getting enough time to ask questions about her health problem of treatment?'
Confidentiality
'[Simon] was speaking to his doctor about an embarrassing problem. There was a friend and a neighbour of his in the crowded waiting room and because of the noise the doctor had to shout when telling [Simon] the treatment he needed.
Q1: How would you rate the way the health services ensured [Simon] could talk privately to health care providers?
Q2: How would you rate the way [Simon's] personal information was kept confidential?'
Quality of basic amenities
'[Wing] had his own room in the hospital and shared a bathroom with two others. The room and bathroom were cleaned frequently and had fresh air.
Q1: How would you rate the cleanliness of the rooms inside the facility, including toilets?
Q2: How would you rate the amount of space [Wing] had?'

†The table provide examples only and not an exhaustive list of possible vignettes for each domain. The response categories that were available to respondents were 'very good', 'good', 'moderate', 'bad' and 'very bad'.

4.1. Non-parametric methods

Murray *et al.* (2003), King *et al.* (2004) and King and Wand (2007) have described non-parametric approaches to adjusting for differential reporting behavior. The methods exploit a natural ordering of the vignettes and the relative position of an individual's self-assessment rating within this ordering. In principle, the ordering of the vignettes is set by the researcher but it could be set by a consensus between respondent ratings. It is essential, however, that the ordering chosen is applied consistently to all responses under analysis. The general approach can be illustrated by using the method that was set out in King *et al.* (2004). This consists of recoding the categorical self-assessed response for each individual relative to their ratings on the set of vignettes. Define y_i as the categorical self-assessment for respondent i and $r_{i1}, \ldots, r_{ik}$ the respondent's ratings of the set of K vignettes. The same set of response categories is available to respondents for both the self-assessment and the set of vignettes. If we assume that all respondents order the vignettes in an identical way ($r_{i,k-1} < r_{ik}$, for all i, k), then King *et al.* (2004) suggested defining a recoded response C_i as follows:

$$C_i = \begin{cases} 1 & \text{if } y_i < r_{i1}, \\ 2 & \text{if } y_i < r_{i1}, \\ 3 & \text{if } r_{i1} < y_i < r_{i2}, \\ 4 & \text{if } y_i = r_{i2}, \\ & \vdots \\ 2K+1 & \text{if } y_i > r_{iK}. \end{cases} \tag{1}$$

Accordingly, an individual's original response to the self-assessment is placed on the recoded scale relative to its position with respect to the individual's responses to the ordered vignettes. By rescaling the self-assessed response relative to the responses to the vignettes, this produces a categorical scale with a larger number of possible categories but one where differential reporting behaviour is removed. The recoded variable can then form the basis for direct comparisons across groups of individuals, for example, by comparing the proportion of individuals reporting a particular category of interest. The recoded variable could, however, be subjected to further analysis using parametric methods, e.g. by using the ordered probit model.

King *et al.* (2004) and King and Wand (2007) described how this approach can be extended to situations where respondents provide ties in their ratings of the vignettes or where respondents' ratings are inconsistent with the natural ordering of the vignettes. Where respondents do not uniquely differentiate between vignettes and instead report ties in their assessments, then C can be defined by a vector of values (or range) rather than a scalar. This can be illustrated by supposing that a respondent rates their self-assessment in the same way (i.e. using the same response category) as she or he rates the first two vignettes, such that $y_i = r_{i1} = r_{i2}$. King and Wand (2007) suggested constructing C as a vector of values which range from the minimum to the maximum of conditions that hold true on the right-hand side of expression (1). Accordingly, for $y_i = r_{i1} = r_{i2}$, we would specify the vector of values $C_i = \{2, 3, 4\}$. Information from respondents who rank the vignettes inconsistently can be incorporated in a similar way, again by summarizing C by a vector of values. For example, a respondent with ratings $y_i = r_{i2} < r_{i1}$, which fails to follow the natural ordering $r_1 < r_2$, would have the vector $C_i = \{1, 2, 3, 4\}$. However, these vector values present challenges for characterizing the distributions of C across different groups of individuals and hence summarizing the data post adjustment for differences in reporting behaviour. King and Wand (2007) and Wand *et al.* (2009) suggested ways of combining information from both scalar and vector values of C including a generalization of the ordered probit model which they termed the censored ordered probit model.

Although the above approaches are useful and require no parametric assumptions in the adjustment for reporting behaviour, they rely on having data on both self-assessments and the full set of vignettes for all survey respondents. Requiring each individual to answer all possible vignette questions places an additional cost on survey implementation which might limit the application of the approach in practice. For example, in the WHS the vignettes are grouped into four sets with each individual rating one set only, and accordingly each vignette has been rated by a quarter of survey respondents. The non-parametric approaches are not well suited to such survey designs. A further limitation is that the method requires the ability to order the vignettes from the best to worst scenario and, although this might be feasible when rating vignettes related to domains of health, it appears less straightforward when rating vignettes related to concepts such as health system responsiveness. For these reasons we do not pursue this approach in our analysis of system performance and instead adopt the parametric approach that is described in the next section.

4.2. Parametric methods: the hierarchical ordered probit model

The standard ordered probit model makes use of a set of constant thresholds, μ^j, $j = 1, \ldots, m$, that is applicable to all individuals to map responses on a latent scale, y_i^*, to observed categorical outcomes y_i. The model can be expressed as

$$y_i = j \qquad \text{if } \mu^{j-1} < y_i^* \leqslant \mu^j,\ j = 1, \ldots, m,$$

where the latent variable y_i^* is assumed to be a linear function of a vector of variables Z plus a random error term ε such that

$$y_i^* = Z_i\beta + \varepsilon_i, \qquad \varepsilon_i|Z_i \sim N(0,1),$$

and $\mu_0 = -\infty, \mu^j < \mu^{j+1}$ and $\mu^m = \infty$. The HOPIT model that was developed by Tandon *et al.* (2003) (also see Terza (1985)) is an extension of the ordered probit model that allows the thresholds to vary across individuals. The method draws on the use of the anchoring vignettes to provide a source of external information that facilitates the identification of the thresholds as functions of covariates. For example, income has been shown to be a determinant of differential reporting behaviour in self-reported general health status such that more wealthy individuals have higher expectations of health and hence report lower levels of objectively identical health status compared with less wealthy counterparts (Bago d'Uva *et al.*, 2008). The model can be specified in two parts. The first part utilizes responses to the vignettes to identify the thresholds as a function of individual characteristics (*the reporting behaviour equation*). The second part maps a set of explanatory variables to underlying health system responsiveness while controlling for differences in reporting behaviour obtained through the first step (*the responsiveness equation*). The two parts are outlined more formally below.

4.2.1. Reporting behaviour equation

To identify the thresholds as a function of respondent covariates, let R_{ik}^{v*} represent the underlying health system responsiveness for vignette k, as perceived by individual i. Given that each vignette is fixed and unrelated to a respondent's characteristics, it is assumed that the expected value of the underlying latent scale depends solely on the corresponding vignette, such that

$$R_{ik}^{v*} = \eta_k + \varepsilon_{ik}^v, \qquad \varepsilon_{ik}^v \sim N(0, \sigma_\varepsilon^2), \tag{2}$$

where η_k indicates the mean of the underlying scale for vignette k, and ε_{ik}^v is an idiosyncratic error term. R_{ik}^{v*} is unobservable to the researcher and instead we observe the vignette rating r_{ik}^v on a five-point categorical scale ranging from very bad to very good. We assume that the observation mechanism relating r_{ik}^v to R_{ik}^{v*} is given by

$$r_{ik}^v = j \qquad \text{if } \mu_i^{j-1} \leqslant R_{ik}^{v*} < \mu_i^j, \qquad \text{for } j = 1, \ldots, 5, \tag{3}$$

with $\mu_i^0 = -\infty$ and $\mu_i^5 = \infty$. If the thresholds represent fixed constants, which are common to all individuals, then the above mapping defines the ordered probit model. For the HOPIT model the thresholds are assumed to be functions of covariates, X such that

$$\begin{aligned} \mu_i^1 &= X_i\gamma^1 + u_i, \\ \mu_i^j &= \mu_i^{j-1} + \exp(X_i\gamma^j), \qquad j = 2,3,4, \end{aligned} \tag{4}$$

with $u_i \sim N(0, \sigma_u^2)$, where γ^j and σ_u^2 are parameters to be estimated along with η_k and σ_ε^2. The error u_i represents an unobserved individual specific random effect and is assumed to be independent of X_i and the other error terms in the model. Its inclusion is intended to reflect the correlation across vignette ratings within respondents and the tendency for some individuals to use high or low thresholds consistently. The thresholds are modelled as an exponential rather than a linear function of the covariates (see also Terza (1985) and Pudney and Shields (2000)) to ensure that they are increasing, and hence respect their natural ordering, over all possible values of X_i (see Greene and Hensher (2009), page 81, for a discussion).

4.2.2. Responsiveness equation

Underlying health system responsiveness faced by individual i can be expressed as

$$R_i^{s*} = Z_i\beta + \varepsilon_i^s, \qquad \varepsilon_i^s|Z_i \sim N(0,1), \tag{5}$$

where Z_i represents a set of regressors predictive of responsiveness (Z and X may overlap). As with the vignettes, R_i^{s*} represents an unobserved latent variable and we assume that the observed categorical response r_i^s relates to R_i^{s*} in the following way:

$$r_i^s = j \qquad \text{if } \mu_i^{j-1} \leqslant R_i^{s*} < \mu_i^j, \tag{6}$$

where μ_i^j are defined by expression (4). The variance of the error term in equation (5) is constrained to 1 and the constant to 0 to allow model identification.

It follows that the probabilities that are associated with each of the five response categories can be computed by

$$\Pr(r_i = j) = \Phi(\mu_i^j - Z_i\beta) - \Phi(\mu_i^{j-1} - Z_i\beta), \qquad j = 1, \ldots, 5, \tag{7}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

The use of vignettes to identify reporting heterogeneity relies on the following two assumptions.

(a) *Response consistency*: it is assumed that individuals classify the vignettes in a way that is consistent with the assessment of their own experiences of health system responsiveness. This implies that the mapping that is used from the latent level of responsiveness shown by the vignettes to the response categories is the same as the mapping that is used to translate latent responsiveness of own experiences to the response categories (hence μ_i^j are assumed to be equivalent in expressions (4) and (6)). King *et al.* (2004) and van Soest *et al.* (2007) have provided some evidence in support of this assumption, whereas evidence that was provided by Bago d'Uva *et al.* (2009), Datta Gupta *et al.* (2010) and Peracchi and Rossetti (2010) is less supportive. Tests of this assumption tend to rely on the availability of objective measures of the concept of interest.

(b) *Vignette equivalence*: it is assumed that

> 'the level of the variable represented by any one vignette is perceived by all respondents in the same way and on the same unidimensional scale, apart from random measurement errors'

(King *et al.* (2004), page 194). This assumption implies that any difference in the way that people perceive the situation represented in each vignette must be random, and hence independent of their country of residence, their sociodemographic characteristics or the level of responsiveness that they face. This is reflected in expression (2) by η_k being the same for all individuals. This assumption might not be tenable in cross-country analyses where, for example, differences in institutional settings might lead to different perceived levels of underlying responsiveness. It has been suggested that comparing across reasonably homogeneous groups of countries and conditioning on country level characteristics will alleviate some of these concerns and we follow this approach in our analysis (Kristensen and Johansson, 2008). The literature investigating the assumption of vignette equivalence is equivocal. Murray *et al.* (2003), King *et al.* (2004) and Kristensen and Johansson (2008) have provided evidence in support of the assumption, largely making use of non-parametric methods, whereas Bago d'Uva *et al.* (2009) and Peracchi and Rossetti (2010) were more sceptical. Rice *et al.* (2010b) have explored the validity of this assumption with reference to the concept of responsiveness and using the WHS. Their results provide some evidence in favour of the assumption of vignette equivalence.

5. Empirical strategy

Our empirical approach is as follows. First, we use Mexico as an illustrative country to establish *prima facie* evidence of differential reporting behaviour and to investigate whether this system-

atically varies by demographic and/or socio-economic characteristics of respondents. Mexico is chosen since the sample size that is available (38 455) is far greater than that for other countries, increasing the scope and precision of analysis. We then make use of the HOPIT model to estimate the relationship between the model thresholds that determine the mapping from the latent level of responsiveness to the observed reporting categories and the set of individual characteristics (4). Conditionally on this relationship, we estimate the responsiveness equation, again, as a function of respondent characteristics (5). The coefficients that are estimated by the HOPIT model are compared with the corresponding estimates derived from a more standard ordered probit model assuming fixed thresholds across all individuals.

The model is then extended to assess differential reporting behaviour across countries. In so doing, we consider a larger set of countries available in the WHS and restrict comparison to countries that are characterized by a high or medium level of the United Nations HDI. The HDI is a composite index of human development which combines indicators of life expectancy, educational attainment and income (United Nations Development Programme, 2006). Analysis within HDI groups imposes a degree of homogeneity across countries in terms of their stage of development which aids comparison. Indeed a criticism of the world health report 2000 which attempted to measure and contrast the performance of healthcare systems was that it failed to stratify countries into defined homogeneous subgroups (Hollingsworth and Wildman, 2003; Williams, 2001). In addition to the demographic and socio-economic characteristics that were outlined above, the models contain country-specific dummy variables. These will reflect, for example, economic and cultural differences across countries within a given HDI group.

Finally, we evaluate whether the ranking of countries in the high and medium HDI groups according to the responsiveness of their health system is affected by the presence of differential reporting behaviour. This is achieved by comparing observed unadjusted raw frequencies of responsiveness with estimated frequencies obtained from predictions from the HOPIT model after fixing reporting behaviour to that observed in a specified benchmark country. For ease of presentation we compare rankings of the proportion of respondents reporting very good responsiveness.

6. Results

6.1. Differential reporting behaviour

Fig. 1 investigates reporting behaviour by sociodemographic position of respondents by presenting the proportion of respondents reporting each of the five categories of responsiveness using the second vignette in the domain clarity of communication for Mexico. This domain is used for illustration since it is rated as being most relevant by Mexican respondents. Results are stratified by educational attainment, income quintiles, gender and age. Evidence of systematic reporting behaviour is provided by observed differences in the reporting of any specific response category (e.g. very good) across the levels of the socio-demographic characteristics being analysed. For example, we observe a clear gradient across educational achievement: in general, better educated respondents are more likely to rate this particular vignette as very good compared with less educated respondents. A gradient is also apparent across income quintiles where individuals who are further along the income distribution are more likely to report very good and less likely to report moderate responsiveness compared with individuals who are at the lower end of the distribution. Although there is some evidence of variation across age groups, in general the figures suggest that reporting behaviour is less influenced by gender or age compared with education and income.

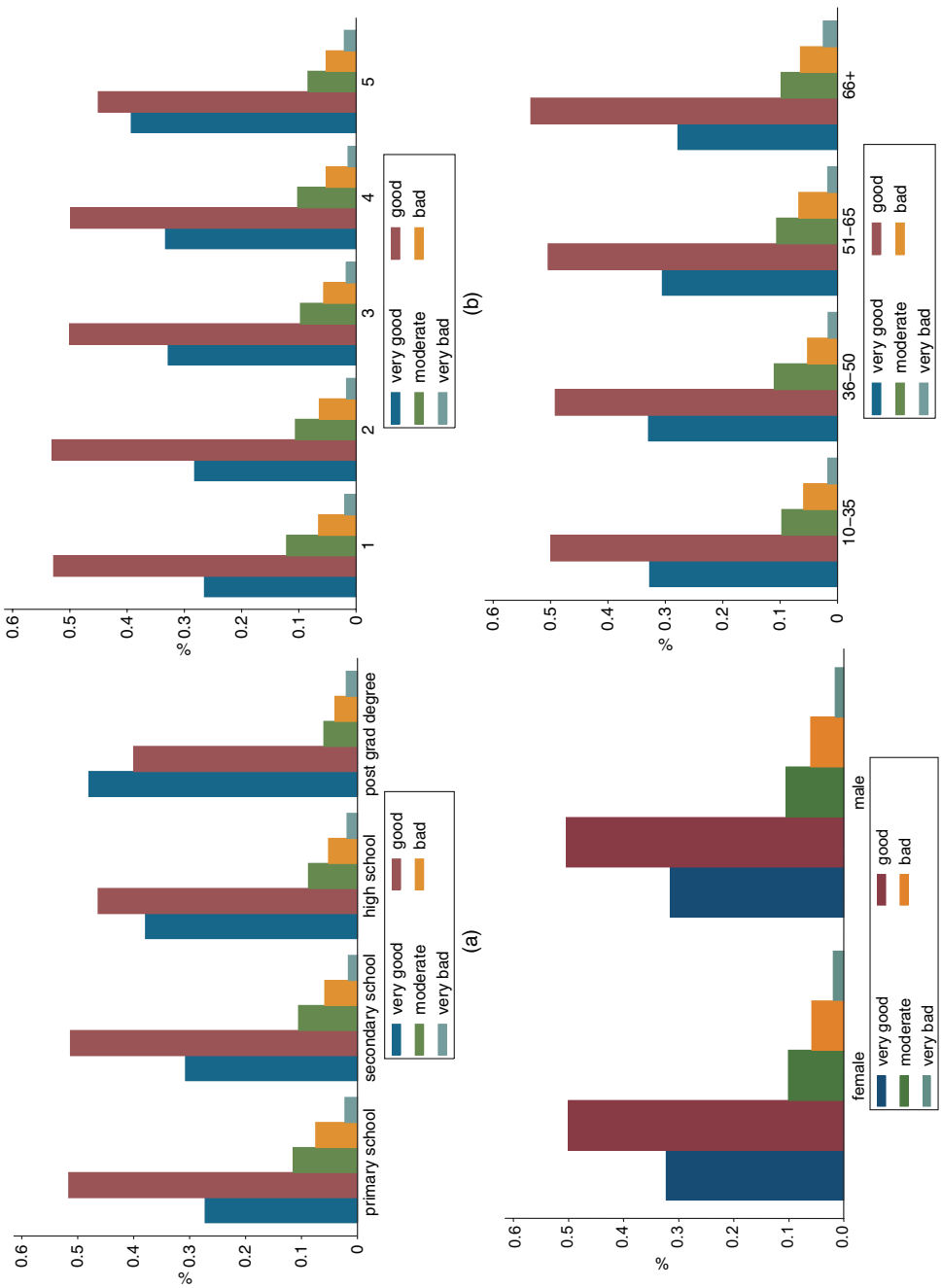

Fig. 1. Vignette ratings for the clarity of communication domian (vignette 2, first item—how clear healthcare providers explained things), Mexico: (a) education; (b) income quintiles; (c) gender; (d) age

Table 3. Tests of homogeneous reporting, Mexico†

	All		*D.Inc2*		*D.Inc3*		*Women*		*Age*		*Education*	
	$\chi^2(20)$	p	$\chi^2(4)$	p	$\chi^2(4)$	p	$\chi^2(4)$	p	$\chi^2(4)$	p	$\chi^2(4)$	p
Dignity												
respect	197.9	*0.00*	16.7	*0.00*	43.3	*0.00*	28.2	*0.00*	0.3	0.99	32.3	*0.00*
privacy	179.2	*0.00*	15.2	*0.00*	21.6	*0.00*	36.2	*0.00*	1.7	0.79	44.4	*0.00*
Communication												
clear explanations	268.0	*0.00*	10.5	*0.03*	39.4	*0.00*	5.8	0.21	15.9	*0.00*	73.3	*0.00*
time for questions	222.2	*0.00*	17.2	*0.00*	24.4	*0.00*	12.8	*0.01*	8.0	0.09	68.3	*0.00*
Confidentiality												
talk privately	269.1	*0.00*	5.8	0.22	27.2	*0.00*	16.2	*0.00*	15.2	*0.00*	102.1	*0.00*
confidential information	266.6	*0.00*	10.6	*0.03*	40.3	*0.00*	17.9	*0.00*	17.4	*0.00*	81.8	*0.00*
Facilities												
cleanliness	502.8	*0.00*	69.9	*0.00*	153.6	*0.00*	3.3	0.52	8.8	0.07	77.4	*0.00*
space	222.2	*0.00*	17.2	*0.00*	24.4	*0.00*	12.8	*0.01*	8.0	0.09	68.3	*0.00*

†$\chi^2(\cdot)$represent χ^2-statistics (the number in parentheses are the degrees of freedom of the null distribution). p-values are derived for tests of homogeneity in reporting. Figures in italics indicate significance at the 5% level. D.Inc2 represents a dummy variable for the second income fertile. Similarly, D.Inc3 represents a dummy variable for the third fertile.

6.2. Within-country analyses

6.2.1. Homogeneity in reporting behaviour

Table 3 presents results of tests for homogeneity in reporting behaviour for Mexico. For each of the sociodemographic characteristics that were considered, χ^2-statistics and p-values from a Wald test of the joint significance of the estimated coefficients across the four thresholds of the model are reported. Rejection of the null hypothesis indicates that the thresholds are functions of the respective sociodemographic characteristic. Results are shown by age, gender, educational attainment (in years) and two dummy variables representing the second and third income tertiles. In addition to separate tests for each variable, the second column reports a joint test across all socio-demographic characteristics. Results for it show that, for all domain and item combinations, the null hypothesis of homogeneous reporting can be rejected. Consistent with the descriptive analysis, the results indicate greater reporting heterogeneity by income and education, compared with age and gender.

6.2.2. Adjusting for reporting heterogeneity

The effect of adjusting for differential reporting behaviour can be investigated by using data on the self-assessments of respondents' own experiences of health service contact. This can be assessed by comparing the estimated coefficients $\hat{\beta}$ in the responsiveness equation (5) with and without adjustment for reporting behaviour by using the ordered probit (unadjusted, but biased in the presence of systematic reporting behaviour) and the HOPIT model (adjusted). To identify the parameters of an ordered probit model it is customary to fix the constant and variance to 0 and 1 respectively (for example, see Greene (2003)). We follow a similar identification strategy in the HOPIT model and hence the coefficients from the two models are comparable.

Table 4. Coefficients of permanent income (second and third tertile) and education in the ordered probit and HOPIT model, Mexico†

Item	*Results for 2nd income tertile*		*Results for 3rd income tertile*		*Results for education*	
	Ordered probit	*HOPIT*	*Ordered probit*	*HOPIT*	*Ordered probit*	*HOPIT*
Dignity						
respect	0.026	0.025	*0.116*	0.031	*0.013*	0.005
	0.032	0.038	0.036	0.043	0.003	0.004
privacy	0.030	0.041	*0.158*	*0.097*	*0.013*	0.003
	0.032	0.040	0.037	0.046	0.003	0.004
Clarity of communication						
clear explanation	0.013	−0.012	*0.112*	0.067	*0.011*	3×10^{-4}
	0.032	0.037	0.037	0.042	0.003	0.004
time for questions	−0.004	−0.039	*0.091*	*0.085*	*0.010*	4E-04
	0.032	0.037	0.036	0.042	0.003	0.004
Confidentiality						
talk privately	−0.053	0.024	0.017	0.035	*0.011*	0.006
	0.031	0.042	0.036	0.048	0.003	0.004
confidential information	0.039	4×10^{-5}	*0.077*	*0.108*	*0.010*	0.004
	0.032	0.043	0.036	0.049	0.003	0.004
Quality of facilities						
cleanliness	−0.060	−0.048	−0.033	−0.027	*0.009*	0.002
	0.031	0.040	0.036	0.045	0.003	0.004
space	−0.004	−0.039	*0.092*	*0.085*	*0.010*	4E-04
	0.032	0.037	0.036	0.042	0.003	0.004

†Coefficients and standard errors are presented for each domain and item combination. Figures in italics indicate significance at the 5% level.

We compare the estimated coefficients for the second and third tertile of income and for education. Results for Mexico are reported in Table 4. For education, for all items and domains the coefficients from the ordered probit model indicate a positive and significant education effect, implying that higher responsiveness is enjoyed by more educated individuals compared with their less educated counterparts. The coefficients and standard errors from the HOPIT model, however, are smaller than those obtained by using the ordered probit model and suggest that the order probit model overestimates the education effect. A similar result is observed for the income tertiles where again, in general, the ordered probit model appears to overestimate the influence of income on responsiveness, particularly for the third tertile, where the effects are statistically significant. Again, the positive income effects imply that higher responsiveness is enjoyed by wealthier individuals compared with their less wealthy counterparts.

6.3. Cross-country analyses

We now consider the effect of adjusting for reporting behaviour in cross-country analyses by modelling a wider set of countries. This is achieved by extending the model that was presented in the previous section by specifying the thresholds (4) as a function of the set of individual socio-demographic characteristics and country-specific dummy variables. The responsiveness equation (5) also adopts this specification. We have tried specifying further models that include interactions terms between the sociodemographic and country dummy variables, but the effects

Table 5. Coefficients and standard errors of cut points as a function of country dummy variables across high HDI countries: results for the item respect in the domain dignity†

Country	μ_1		μ_2		μ_3	
	Coefficient	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*
United Arab Emirates	−0.016	0.062	0.103	0.056	*−0.472*	0.044
Austria	*−0.281*	0.096	0.085	0.088	*−0.452*	0.056
Belgium	*0.812*	0.102	−0.201	0.103	*−0.411*	0.057
Bosnia	−0.083	0.064	0.082	0.059	*−0.376*	0.040
Czech Republic	−0.028	0.077	*0.186*	0.069	*−0.451*	0.048
Germany	0.062	0.071	0.113	0.063	*−0.308*	0.039
Denmark	*0.945*	0.087	*−0.270*	0.097	*−0.539*	0.062
Spain	*0.089*	0.033	*−0.164*	0.035	*−0.105*	0.017
Estonia	0.071	0.073	*0.201*	0.063	*−0.283*	0.041
Finland	*0.373*	0.070	*0.327*	0.057	*−0.284*	0.042
France	*0.383*	0.104	0.160	0.091	*−0.394*	0.066
UK	*0.400*	0.072	0.018	0.069	*−0.480*	0.046
Greece	0.108	0.074	0.093	0.067	*−0.595*	0.051
Croatia	*0.464*	0.075	*0.264*	0.058	*−0.764*	0.055
Hungary	−0.072	0.057	0.053	0.053	*−0.468*	0.036
Ireland	*0.429*	0.091	−0.128	0.088	*−0.552*	0.059
Italy	−0.055	0.143	−0.052	0.137	*−0.353*	0.085
Latvia	*0.324*	0.079	0.038	0.070	*−0.503*	0.049
Mauritius	*0.653*	0.037	*−0.171*	0.037	*−0.374*	0.022
Malaysia	−0.026	0.035	0.000	0.034	*−0.090*	0.018
Netherlands	−0.001	0.077	*0.319*	0.063	*−0.125*	0.043
Portugal	−0.043	0.081	*0.262*	0.066	*−0.089*	0.041
Slovakia	*−0.192*	0.056	0.094	0.051	*−0.471*	0.037
Slovenia	*0.400*	0.084	*−0.207*	0.091	*−0.356*	0.053
Sweden	*0.811*	0.079	−0.003	0.076	*−0.522*	0.054
Uruguay	0.085	0.046	−0.025	0.047	*−0.071*	0.024

†Mexico is the baseline country. Figures in italics indicate significance at the 5% level. μ_1–μ_3 refer to thresholds 1–3. The model is estimated without interactions.

were largely non-significant and the models produced results that are similar to those presented for the more parsimonious model. To enhance comparability of results, models are estimated across countries within each of the medium and high HDI groups. The models are highly computationally intensive and to aid convergence we have aggregated the response categories very bad and bad into a single category. Accordingly we collapse the five-point categorical scale to a four-point scale. Given the extremely small percentage of respondents who rate responsiveness as very bad (the average across high and medium HDI countries is 1.1% and 1.4 % respectively), the effects of the aggregation of these categories on the estimation results are likely to be negligible.

Table 5 reports the coefficients and standard errors for the set of country dummy variables in the threshold equations for the item respect (in the domain dignity). For brevity results for the set of high HDI countries only are shown. The coefficients are contrasted against the baseline country, Mexico. Variation in estimated coefficients illustrates the existence of differential reporting behaviour across countries and we note that these attain statistical significance in at least one of the three thresholds for each country. The majority of the coefficients for the first and the second threshold are positive whereas those for the third threshold are negative. In general, these results imply that compared with Mexico respondents in other countries are likely to make greater use of the extremes of the available reporting categories when rating performance.

Table 6. Observed and estimated frequencies of reporting very good responsiveness for the item respect, high HDI countries

Rank by block (1)	*Observed data frequencies, (1)*		*Frequencies from HOPIT model (country-specific cut points), (2)*		*Frequencies from HOPIT model (Mexico-specific cut points), (3)*		*Rank by block (1), (4)*
1	Austria	61.9%	Austria	57.4%	Denmark	54.2%	2
2	Denmark	61.0%	Denmark	56.9%	Finland	53.4%	7
3	Sweden	55.8%	Sweden	52.8%	Sweden	52.6%	3
4	Czech Republic	52.9%	UK	51.3%	Belgium	45.9%	11
5	UK	51.4%	Czech Republic	51.2%	France	42.7%	9
6	Greece	51.0%	Greece	50.2%	UK	42.0%	5
7	Finland	49.3%	Finland	47.5%	Netherlands	40.8%	17
8	Hungary	47.8%	Hungary	46.9%	Uruguay	38.9%	13
9	France	47.6%	United Arab Emirates	46.6%	Czech Republic	36.3%	4
10	Ireland	45.7%	Belgium	46.4%	Estonia	33.5%	16
11	Belgium	44.9%	Ireland	45.5%	Austria	33.0%	1
12	United Arab Emirates	44.4%	France	45.4%	Ireland	32.1%	10
13	Uruguay	37.9%	Bosnia	41.1%	Greece	31.8%	6
14	Latvia	36.2%	Uruguay	40.9%	Spain	31.3%	20
15	Bosnia	36.1%	Croatia	39.4%	Croatia	30.7%	18
16	Estonia	35.5%	Latvia	39.2%	Mauritius	30.1%	24
17	Netherlands	35.3%	Estonia	39.2%	United Arab Emirates	29.7%	12
18	Croatia	35.1%	Germany	38.4%	Germany	29.4%	19
19	Germany	34.2%	Netherlands	38.3%	Slovenia	28.8%	21
20	Spain	30.9%	Slovenia	37.7%	Latvia	28.6%	14
21	Slovenia	30.4%	Spain	37.5%	Portugal	28.2%	25
22	Slovakia	27.6%	Slovakia	36.7%	Hungary	27.6%	8
23	Italy	26.2%	Mauritius	33.0%	Mexico	26.2%	27
24	Mauritius	24.2%	Italy	30.6%	Bosnia	25.6%	15
25	Portugal	18.5%	Malaysia	28.9%	Malaysia	24.5%	26
26	Malaysia	18.2%	Portugal	27.0%	Slovakia	18.2%	22
27	Mexico	16.3%	Mexico	26.2%	Italy	16.5%	23
Pearson's correlation coefficient ρ			Blocks (2) and (1), 0.986		Blocks (3) and (1), 0.737		
Kendall's τ			Blocks (2) and (1), 0.906		Blocks (3) and (1), 0.547		

The results of Table 5 establish the existence of differential reporting behaviour across countries. We next investigate the effect of adjusting for country-specific reporting behaviour by comparing estimated frequencies of reporting very good responsiveness derived from the results of the HOPIT model. These frequencies are presented in Tables 6 and 7 for the item respect respectively for the high and medium HDI groups of countries. The third column of Tables 6 and 7 reports the raw frequencies from respondent ratings observed in the data. These vary substantially and have been ranked in order of reporting very good responsiveness. For example, in the high HDI group 61.9% of respondents in Austria report very good responsiveness compared with 16.3% of respondents in Mexico. This variation in ratings will reflect differences in true underlying health system responsiveness that are faced by individuals but will also, in part, reflect systematic variations in reporting behaviour that differ across countries. The challenge for comparative analysis is to isolate the effect of the former, abstracting from the effect of the latter. Only then can we make meaningful cross-country comparisons of performance.

Table 7. Observed and estimated frequencies of reporting very good responsiveness for the item respect, medium HDI countries

Rank by block (1)	*Observed data frequencies, (1)*		*Frequencies from HOPIT model (country-specific cut points), (2)*		*Frequencies from HOPIT model (India-specific cut points), (3)*		*Rank by block (1), (4)*
1	Paraguay	53.6%	Paraguay	50.6%	Paraguay	43.1%	1
2	Brazil	38.7%	Brazil	41.6%	Georgia	40.5%	3
3	Georgia	31.4%	Georgia	37.8%	Brazil	37.6%	2
4	Ecuador	31.0%	Ecuador	35.4%	Myanmar	36.1%	21
5	South Africa	27.7%	South Africa	32.5%	Dominican Republic	36.1%	12
6	Ghana	27.1%	Ghana	32.2%	Philippines	35.8%	26
7	Namibia	25.2%	Morocco	30.9%	China	33.4%	17
8	Morocco	25.1%	Namibia	30.7%	Guatemala	33.2%	16
9	Bangladesh	24.6%	Swaziland	28.0%	Ecuador	32.3%	4
10	India	20.5%	Congo	27.0%	Congo	30.1%	13
11	Swaziland	17.6%	Bangladesh	25.7%	Namibia	29.7%	7
12	Dominican Republic	17.1%	India	25.6%	Comoros	29.5%	19
13	Congo	16.6%	Dominican Republic	24.4%	South Africa	28.3%	5
14	Tunisia	16.2%	Kazakhstan	23.8%	India	25.6%	10
15	Lao	15.8%	Lao	23.5%	Kazakhstan	25.6%	20
16	Guatemala	15.0%	Russia	23.4%	Ghana	25.2%	6
17	China	14.6%	Tunisia	23.2%	Lao	24.0%	15
18	Russia	13.1%	China	22.5%	Bangladesh	23.9%	9
19	Comoros	12.8%	Sri Lanka	21.9%	Ukraine	23.8%	22
20	Kazakhstan	12.7%	Comoros	21.4%	Russia	23.5%	18
21	Myanmar	11.6%	Ukraine	19.3%	Swaziland	22.7%	11
22	Ukraine	9.9%	Myanmar	19.0%	Pakistan	22.6%	23
23	Pakistan	9.8%	Vietnam	18.5%	Tunisia	20.2%	14
24	Sri Lanka	9.4%	Pakistan	16.0%	Vietnam	20.2%	27
25	Nepal	9.2%	Philippines	15.5%	Nepal	19.0%	25
26	Philippines	7.6%	Guatemala	15.5%	Sri Lanka	16.7%	24
27	Vietnam	7.4%	Nepal	12.7%	Morocco	15.4%	8
Pearson's correlation coefficient ρ			Blocks (2) and (1), 0.931		Blocks (3) and (1), 0.410		
Kendall's τ			Blocks (2) and (1), 0.803		Blocks (3) and (1), 0.307		

Blocks (2) and (3) present estimated frequencies that were obtained from the HOPIT model. The modelling of the thresholds through equation (4) allows us to control for differential reporting behaviour across individuals within countries (via socio-demographic characteristics) and across countries (via country dummy variables). We use the results in the following two ways. First, block (2) represents the estimated frequencies that were obtained from the model calculated separately for each country, and adjusting for within-country reporting behaviour. Crucially the model does not adjust for differences in reporting across countries. Estimated frequencies are obtained by anchoring the relevant parameters in the thresholds to the characteristics of the 'average' respondent in each of the countries that were considered. We refer to this model as the 'country-specific' HOPIT model. Owing to the use of within-country thresholds, the estimated frequencies resemble quite closely the frequencies that were observed in the raw data and the ranking of countries does not change markedly. The correlation coefficient between the raw frequencies (1) and the estimated frequencies (2) is high (Pearson's correlation is 0.99

and 0.93 for the high and medium HDI countries respectively; Kendall's τ rank correlation is 0.91 and 0.80), indicating a fairly strong association.

Secondly, to provide rankings that are comparable across countries we benchmark reporting behaviour to that observed in the baseline country chosen in each of the high and medium HDI groups. Again, we then estimate the reporting of very good responsiveness separately for each country assuming that all respondents had the reporting behaviour of the baseline country, i.e., for each country within an HDI group, the estimated probability of reporting very good responsiveness is computed using the thresholds for the baseline country (Mexico for high HDI countries (Table 6) and India for medium HDI countries (Table 7)). Table 5 has shown that in the high HDI group of countries respondents in general are more likely to rate the respect item as very good compared with Mexican respondents (indicated by the negative coefficient on the third cut point). This result is consistent with the observation that the estimated frequency of reporting very good responsiveness decreases for the majority of countries when the cut points of Mexico, instead of the country-specific cut points, are used to compute the estimations. Adopting the reporting behaviour that is observed in the baseline country offers a more comparable basis on which to rank the countries, the results of which are provided in block (3). Inspection of these results reveals a ranking that is different from that observed for the raw frequencies (block (1)). For example, for the high HDI group of countries, Austria falls 11 places and Bosnia falls nine places in the rankings once we benchmark to adjust for differential reporting behaviour. In contrast, the Netherlands moves up 10 places and Mauritius eight places in the rankings post benchmarking. For the countries in the medium HDI group, Bangladesh and Tunisia fall nine places in the rankings whereas the Philippines rises 20 places and China rises 10 places.

If we use the correlation between the raw frequencies and the HOPIT-adjusted estimated frequencies (3) as a measure of association of the results and an indication of the closeness of the rankings then we see that these are lower than their respective values when comparing the raw data with the within-country estimated frequencies (2)—Pearson's coefficient is 0.74 for the set of high HDI countries and 0.41 for the medium HDI countries; Kendall's τ is 0.55 and 0.31 respectively. This notable decrease in the correlations supports a change in the orderings of the countries before and after adjusting for reporting behaviour and confirms the visual inspection of the rankings outlined above that reveals large differences.

Estimated frequencies from the benchmarked HOPIT model allow us to consider the importance of adjusting for differential reporting in explaining cross-country differences in reported rates of responsiveness. For example, if we consider the group of high HDI countries, a naive estimate of the difference in reporting very good responsiveness between the country ranked first (Austria) and the baseline country (Mexico) is 45.6% (61.9% − 16.3%). If we anchor reporting behaviour in Austria to the response scales that were used by Mexican respondents, the difference is reduced to 6.9% (33.0% − 26.2%). Accordingly, approximately 85% of the observed difference in reporting frequencies between the highest and the lowest ranked countries appears to be due to reporting behaviour. Although this is an extreme example and results will vary by the choice of countries compared, it illustrates the potential effect that reporting behaviour may have on cross-country comparisons of performance.

Although the approach described is designed to remove the influence of systematic reporting behaviour across countries, a natural question to ask is whether the ranking that is produced from the adjusted estimated responses provides a more accurate reflection of system performance than the rankings that are observed in the unadjusted raw data. In an attempt to address this question, we follow Datta Gupta *et al.* (2010) and compare the rankings based on the estimated adjusted frequencies with those obtained through a potentially more objective measure of system performance in the form of healthcare spending *per capita* (measured in 2001 US

dollars: source United Nations Development Programme). Clearly finding objective measures of performance is intrinsically difficult—after all why bother to undertake a cross-country comparison based on self-reported data where objective measures exist? It is also debatable whether our choice of measure has a clear link to the responsiveness of a health system. However, all other things being equal, in high income settings more spending is likely to feed through to greater quality and responsiveness of health services. We correlate the ranking of countries that is produced from *per capita* healthcare expenditures to those observed in the raw data (Table 6; block (1)) by using Kendall's τ-statistic. This is compared with the ranking between *per capita* expenditure and the adjusted estimated frequencies. The country rankings are available on request, but in summary we find $\tau = 0.373$ for the former comparison and $\tau = 0.461$ for the latter, suggesting that, although healthcare expenditure *per capita* is itself an imperfect measure of system performance, the adjusted estimated frequencies derived from the HOPIT model appear to reflect better an arguably more objective measure than the raw data frequencies.

7. Conclusions and discussion

A clear purpose for outcome measurement is to enable institutions to compare and contrast their performance with that of others, including at the macrolevel the performance that is secured in other countries. For this international comparison has become one of the most influential levers for change in public services. Increasingly patients' views and opinions obtained through surveys are being recognized as a legitimate and important means for assessing the performance of health systems. A reliance on individual level survey data based on respondents' self-reports of system performance presents challenges for international comparison. In particular, self-reported data are likely to suffer from the existence of systematic variations in reporting behaviour. This might be evident both across individuals, stratified by sociodemographic characteristics, within countries and across countries. Such reporting heterogeneity results from survey respondents applying different thresholds when reporting (using a categorical scale) an underlying latent construct such as health system responsiveness. Accordingly, a given fixed level of performance might be rated differently across survey respondents. To identify true underlying differences in performance, measures of performance need to be purged of systematic variations in reporting behaviour. Using the method of anchoring vignettes this paper has illustrated how the reporting of health system responsiveness might vary both within and across countries. Our results indicate the presence of variation in reporting behaviour that is linked to the sociodemographic characteristics of survey respondents within countries.

Differential reporting behaviour appears to exist across countries. This is evident in the WHS data where country level rankings of responsiveness obtained from the observed raw data vary from the estimated rankings obtained through the HOPIT model when reporting behaviour is anchored to a common scale. Although some caution is merited when interpreting rankings as definitive indications of comparative system performance, the results suggest that cross-country analyses that rely on survey respondents' reports of interactions with public services need to consider the extent of systematic differences in reporting behaviour. For this, the method of anchoring vignettes offers a potentially powerful tool to adjust survey results and to place cross-country comparative analysis on a more consistent footing than that obtained from a simple comparison of observed raw data frequencies.

The use of anchoring vignettes in conjunction with the HOPIT model promises to be an important tool to aid cross-country comparison of health system performance. The use of the approach, however, has limitations. First, the set of sociodemographic variables that were extracted from the WHS used in this work appear to be better predictors of variation in reporting

behaviour (used to model the thresholds: equation (4)) than predictors of underlying health system responsiveness (used in the responsiveness: equation (5)). Future research might focus on the appropriate determinants of health system responsiveness to aid cross-country comparison further. Secondly, the method relies on the assumption of response consistency and vignette equivalence and the validity of these assumptions remains the subject of current research (van Soest *et al.*, 2007; Kristensen and Johansson, 2008; Bago d'Uva *et al.*, 2009; Peracchi and Rossetti, 2010). Thirdly, the inclusion of vignettes necessarily entails a cost for survey implementation and it is, therefore, important to consider their design to ensure that they elicit relevant information efficiently. This is a further area of on-going research activity (King and Wand, 2007). Finally, the HOPIT model is heavily parameterized and non-parametric methods to enhance cross-country comparability of system performance should be investigated where data allow.

Acknowledgements

This research was funded by the Economic and Social Research Council under the 'Public services programme' (RES-166-25-0038) and the 'Large grant scheme' (RES-060-25-0045). We thank the World Health Organization for providing access to the WHS and, in particular, Somnath Chatterji, Amit Prasad, Nicole Valentine and Emese Verdes. We are grateful to five referees and the journal's editors for helpful comments on an earlier draft and the Health, Econometrics and Data Group Seminar Series at the University of York.

References

Bago d'Uva, T., van Doorlsaer, E., Lindeboom, M. and O'Donnell, O. (2008) Does reporting heterogeneity bias the measurement of health disparities? *Hlth Econ.*, **17,** 351–375.

Bago d'Uva, T., Lindeboom, M., O'Donnell, O. and van Doorslaer, E. (2009) Slipping anchor?: testing the vignettes approach to identification and correction of reporting heterogeneity. *Discussion Paper 09-091/3*. Tinbergen Institute, Rotterdam.

Blendon, R. J., Schoen, C., DesRoches, C., Osborn, R. and Zapert, K. (2003) Common concerns amid diverse systems: health care experiences in five countries. *Hlth Aff.*, **22**, no. 3, 106–121.

Brislin, R. W. (1986) The wording and translation of research instruments: field methods in cross-cultural research. In *Field Methods in Cross-cultural Research* (eds W. J. Lonner and J. W. Berry), pp 137–164. Beverly Hills: Sage.

Coulter, A. and Magee, H. (2003) *The European Patient of the Future (State of Health)*. Maidenhead: Open University Press.

Datta Gupta, N., Kristensen, N. and Pozzoli, D. (2010) External Validation of the use of vignettes in cross-country health studies. *Econ. Modllng*, **27**, 854–865.

Ferguson, B. D., Tandon, A., Gakidou, E. and Murray, C. J. L. (2003) Estimating permanent income using indicator variables. In *Health Systems Performance Assessment: Debates, Methods and Empiricism* (eds C. J. L. Murray and D. B. Evans), pp. 748–760. Geneva: World Health Organization.

Gonzalez Block, M. A. (1997) Comparative research and analysis methods for shared learning from health system reforms. *Hlth Poly*, **42**, 187–209.

Greene, W. H. (2003) *Econometric Analysis*. Upper Saddle River: Pearson.

Greene, W. H. and Hensher, D. A. (2009) *Modeling Ordered Choices: a Primer and Recent Developments*. New York: Cambridge University Press.

de Groot, W. (2000) Adaptation and scale of reference bias in self-assessments of quality of life. *J. Hlth Econ.*, **19**, 403–420.

Hollingsworth, B. and Wildman, J. (2003) The efficiency of health production: reestimating the WHO panel data using parametric and non-parametric approaches to provide additional information. *Hlth Econ.*, **12**, 493–504.

Iburg, K. M., Salomon, J., Tandon, A. and Murray, C. J. L. (2002) Cross-country comparability of physician-assessed and self-reported measures of health. In *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications* (eds C. J. Murray, J. A. Salomon, C. D. Mathers and A. D. Lopez), pp. 433–448. Geneva: World Health Organization.

Idler, E. L. and Kasl, S. V. (1995) Self-ratings of health do they also predict change in functional ability? *J. Gerontol. Soc. Sci.*, **B50**, 344–353.

Jürges, H. (2007) True health versus response styles: exploring cross-country differences in self-reported health. *Hlth Econ.*, **16**, 163–178.
Kapteyn, A., Smith, J. and van Soest, A. (2007) Vignettes and self-reports of work disability in the US and the Netherlands. *Am.Econ. Rev.*, **97**, 461–473.
Kempen, G. I., Steverink, N., Ormel, J. and Deeg, D. J. (1996) The assessment of ADL among frail elderly in an interview survey: self-report versus performance-based tests and determinants of discrepancies. *J. Gerontol. Soc. Sci.*, **B51**, 254–260.
Kendall, M. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–89.
Kerkhofs, M. J. M. and Lindeboom, M. (1995) Subjective health measures and state dependent reporting errors. *Hlth Econ.*, **4**, 221–235.
King, G., Murray, C. J. L., Salomon, J. and Tandon, A. (2004) Enhancing the validity and cross-cultural comparability of measurement in survey research. *Am. Polit. Sci. Rev.*, **98**, 184–191.
King, G. and Wand, J. (2007) Comparing incomparable survey responses: new tools for anchoring vignettes. *Polit. Anal.*, **15**, 46–66.
Kristensen, N. and Johansson, E. (2008) New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Lab. Econ.*, **15**, 96–117.
Lindeboom, M. and van Doorslaer, E. (2004) Cut-point shift and index shift in self-reported health. *J. Hlth Econ.*, **23**, 1083–1099.
Lynn, P., Japec, L. and Lyberg, L. (2006) What's so special about cross-national surveys? *Working Paper 2005-16*. Institute for Social and Economic Research, University of Essex, Colchester.
Manderbacka, K. (1998) Examining what self-rated health question is understood to mean by respondents. *Scand. J. Soc. Med.*, **26**, 145–153.
Murray, C. J. L. and Frenk, J. (2000) A framework for assessing the performance of health systems. *Bull. Wrld Hlth Orgnzn.*, **78**, 717–731.
Murray, C. J. L., Ozaltin, E., Tandon, A., Salomon, J., Sadana, R. and Chatterji, S. (2003) Empirical evaluation of the anchoring vignettes approach in health surveys. In *Health Systems Performance Assessment: Debates, Methods and Empiricism* (eds C. J. L. Murray and D. B. Evans), pp. 369–399. Geneva: World Health Organization.
Okazaki, S. and Sue, S. (1995) Methodological issues in assessment research with ethnic minorities. *Psychol. Assessmnt*, **7**, 367–375.
O'Mahony, M. and Stevens, P. A. (2004). *International Comparisons of Performance in the Provision of Public Services: Outcome Based Measures for Education*. London: National Institute of Economic and Social Research.
Peracchi, F. and Rossetti, C. (2010) The heterogeneous thresholds ordered response model: Identification and inference. *Working Paper 10/12*. EIEF.
Pudney, S. and Shields, M. (2000) Gender, race, pay and promotion in the British nursing profession: estimation of a generalized ordered probit model. *J. Appl. Econ.*, **15**, 367–399.
Puentes Rosas, E., Gómez Dantés, O. and Garrido Latorre, F. (2006) Trato a los usuarios en los servicios públicos de salud en México. *Rev. Panam. Salud. Publ.*, **9**, 394–402.
Rice, N., Robone, S. and Smith, P. C. (2010a) International comparison of public sector performance: the use of anchoring vignettes to adjust self-reported data. *Evaluation*, **16**, 81–100.
Rice, N., Robone, S. and Smith, P. C. (2010b) Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness. *Eur. J. Hlth Econ*. to be published.
Sadana, R., Mathers, C. D., Lopez, A. D., Murray, C. J. L. and Iburg, K. M. (2002) Comparative analyses of more than 50 household surveys on health status. In *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications* (eds C. J. Murray, J. A. Salomon, C. D. Mathers and A. D. Lopez), pp. 369–386. Geneva: World Health Organization.
Sirven, N., Santos-Eggimann, B. and Spagnoli, J. (2008) Comparability of health care responsiveness in Europe: Using anchoring vignettes from SHARE. Working Paper DT 15. IRDES, Paris.
van Soest, A., Delaney, L., Harmon, C., Kapteyn, A. and Smith J. P. (2007) *Validating the use of vignettes for subjective threshold scales*. Mimeo. RAND Corporation, Santa Monica.
Tandon, A., Murray, C. J. L., Salomon, J. A. and King, G., (2003) Statistical models for enhancing cross-population comparability. In *Health Systems Performance Assessment: Debates, Methods and Empiricism* (eds C. J. L. Murray and D. B. Evans), pp. 727–746. Geneva: World Health Organization.
Terza, J. V. (1985) Ordinal probit: a generalization. *Communs Statist.*, **14**, 1–11.
United Nation Development Programme (2006) *Human Development Report*. New York: United Nations Development Programe.
Üstün, T. B., Chatterji, S., Mechbal, A., Murray, C. *et al.* (2003) The World Health Surveys. In *Health Systems Performance Assessment: Debates, Methods and Empiricism* (eds C. J. L. Murray and D. B. Evans), pp. 762–796. Geneva: World Health Organisation.
Valentine, N. B., De Silva, A., Kawabata, K., Darby, C., Murray, C. J. L. and Evans, D. (2003a) Health system responsiveness: concepts, domains and operationalization. In *Health Systems Performance Assessment: Debates, Methods and Empiricism* (eds C. J. L. Murray and D. B. Evans), pp. 573–596. Geneva: World Health Organization.

Valentine, N. B., Ortiz, J. P., Tandon, A., Kawabata, K., Evans, D. B. and Murray, C. J. L. (2003b) Patient experiences with health services: population surveys from 16 OECD counties. In *Health Systems Performance Assessment: Debates, Methods and Empiricism* (eds C. J. L. Murray and D. B. Evans), pp. 643–652. Geneva: World Health Organization.

Valentine, N. B., Prasad, A., Rice, N., Robone, S. and Chatterji, S. (2009) Health systems responsiveness—a measure of the acceptability of health care processes and systems In *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects* (eds P. Smith, E. Mossialos and S. Leatherman). London: World Health Organization European Regional Office.

Wand, J., King, G. and Lau, O. (2009). Anchors: software for anchoring vignette data. *J. Statist. Softw.*, to be published.

Williams, A. (2001) Science or marketing at WHO?: a commentary on 'World Health 2000'. *Hlth Econ.*, **10**; 93–100.