International Comparison of Public Sector Performance: The Use of Anchoring Vignettes to adjust Self-Reported Data

Nigel Rice, Silvana Robone, Peter C. Smith

Centre for Health Economics, University of York

Abstract

International comparison of performance has become an influential lever for change in the provision of public services. For health care, patients' views and opinions are increasingly being recognized as legitimate means for assessing the provision of services, to stimulate quality improvements, and more recently, in evaluating system performance. This has shifted the focus of analyses towards the use of individual-level surveys of performance from the perspective of the user and raises the issue of how to compare appropriately self-reported data across institutional settings and population groups. This represents a major challenge for all public services, the fundamental problem being that comparative evaluation needs to take account of variations in social and cultural expectations and norms when relying on self-reported information. Using data on health systems responsiveness across 18 OECD countries contained within the World Health Survey, this paper outlines the issues that arise in comparative inference that relies on respondent self-reports. The problem of reporting bias is described and illustrated together with potential solutions brought about through the use of anchoring vignettes. The utility of vignettes to aid cross-country analyses and its implications for comparative inference of health system performance are discussed.

Introduction

International comparison has become one of the most influential levers for change in public services. By establishing benchmarks in examining performance, cross-national comparison offers opportunities for countries to assess their place in relation to others; to learn from experience elsewhere; to identify and explore trends in performance both within and across public sectors and to inform realistic expectations on likely rates of growth in different sectors of the economy and the appropriate allocation of resources (O` Mahony and Stevens 2004, Gonzalez Block 1997). Crucially, the ability to benchmark performance against relevant comparative countries enables the public sector to promote accountability to its citizens and, perhaps more importantly, to systematically evaluate performance. Moreover, increasingly, international organisations and national agencies require comparative data across a range of countries to support their specific programmes and policies. In these ways, international comparison plays a fundamental role in promoting the diffusion and uptake of innovative practice aimed at improving the provision of public services and in securing public support (O`Mahony and Stevens 2004, Kumar and Ozdamar 2004, Gonzalez Block 1997). The value placed on comparative information is perhaps best evidenced by the extensive resources invested by national and international organizations in the collection, publication and analyses of cross-national data (Ovretveit, 1998).

International comparisons have been conducted in many areas of the public arena. In respect of health care systems, international comparative work has informed debate and policy across EU

and OEDC countries on, among others, the level of health care spending (Anderson et al. 2007, White 2007, Anell and Willis 2000, Schieber and Poullier 1991), health care performance (OECD, 2000, Andersen, 2001, Reinhardt et al, 2002), access to health care (van Doorslaer and Masseria, 2004), waiting times (Siciliani and Hurst, 2005; Willcox et al. 2007), patients' experiences of care (Coulter and Cleary, 2001) and the configuration and delivery of primary care services (Schoen et al. 2007, Schoen et al. 2006). Perhaps the most scrutinized comparative indicators of health system performance is the World Health Report 2000 (World Health Organization, 2000) which provides country rankings across measures of health systems goals defined as improving health, responding to individual's expectations and fairness of financial contributions. Across other sectors, perhaps the best known comparative analysis is the Programme for International Student Assessment (PISA) which compares the educational achievement of 15 year olds across 43 OECD countries (OECD, 2004).

A fundamental concern of cross-national surveys is the comparability of the data collected. This refers to the usefulness of data in drawing comparison across different populations, including the ability to draw meaningful inference in relation to each other or against a common benchmark (Verma, 2002). Many of the data that international comparisons have traditionally relied upon have been reported at broad aggregate country level, for example, the number of hospitals or doctors or the level of health care expenditure. A reliance on aggregate data, however, often makes it difficult to successfully disentangle the many possible reasons for observed variations between countries and accordingly, the results of such analyses are often highly contested and inconclusive. This might be due to a lack of accepted definitions across countries for the outcomes of interest, even where these are thought to be objectively measured (for example, common definitions for health care providers, such as hospitals, physicians, and nursing homes etc.) (Ovretveit 1998, Yepes, 1991, Gravelle 1987, Scheiber 1991). Even where standard agreed definitions exist, problems may arise if reporting practices differ across countries for example, due to national statistical offices adopting different accounting principles and systems (Arah et al. 2003, Bullinger 2003, Schieber 1991).

The design of cross-national surveys presents challenges beyond those required for single country surveys. A crucial consideration is the identification of items and concepts that form the survey question to ensure they are meaningful across the countries sampled (Lynn 2005). In addition to concerns about the psychometric statistical properties relating to the reliability and validity of survey instruments, the design of the survey needs to be mindful of the requirement for cross-cultural equivalence of instruments. Social scientists distinguish between interpretative and procedural equivalence (Johnson 1998). The former refers mainly to the subjective cross-cultural comparability of the meaning of items and concepts, and the latter is concerned with the objective development of comparable survey measures across cultural groups. There are a number of methods used to establish or assess the cross-cultural equivalence of a survey. These include involving country experts to ensure the meaning of the questions is as intended (Okasaki and Sue, 1995); using "good" question wording practices (using short and simple sentences, making reference to concepts relevant for everyday existence, and avoiding slang terms, Brislin 1986, Scheuch 1993 and McKay et al. 1996); and following, for example, back-translation procedures (Johnson 1998).[1] Other techniques include the use of follow up questions at a pre-testing phase to identify difficulties with question interpretation (Johnson et al. 1997, Krause and Jay 1994) and the use of multiple indicators for the construct to be measured to avoid losing data due to poor performance of individual indicators

[1] The basic procedure of back-translation "calls for a bilingual person to translate a source questionnaire into a target language. A second bilingual person is then asked to translate this version back into the source language without knowledge of the original instrument. The initial and revised versions of the source language version are then compared, discrepancies are identified, and appropriate revisions are made" (Johnson 1998, pp 18).

(Okazaki and Sue 1995). Securing comparability between countries requires careful planning to ensure the data collected are not beset by country-specific differences in the methods used for data collection or cultural differences in the meaning and reporting of survey questions.

In an attempt to enable more comprehensive investigation of the reasons for observed differences in comparative performance, recently the focus of research has shifted towards the use of data measured at an individual level, often derived from administrative records or cross-country surveys. Moreover, measures of performance are becoming increasingly reliant on the perspective of the user where patients' views and opinions have long been recognized as a legitimate means for assessing the provision of health services (Coulter and Magee, 2003). Measures of performance reliant on respondent self-reported data contain additional challenges. In particular, a critical issue for international comparison is variation across individuals both within and between countries regarding both their perception of survey questions and the concepts the questions are attempting to elicit (Lynn 2005). For example, in the absence of an appropriate benchmark, ratings of health status on a scale of very poor to excellent are likely to imply different underlying `true health' for an active 20 year old, than a sedentary 80 year old. Drawing on survey data on health systems responsiveness, this paper describes the issues that arise when using self-reported data to compare cross-country health systems performance. The use of anchoring vignettes to adjust self-reports for systematic reporting behaviour is described and illustrated through the analysis of performance across 18 OECD countries.

Anchoring vignettes

Efforts to enhance cross-cultural equivalence of survey instruments will not ensure comparability if individuals, when faced with an instrument involving ordinal response categories, interpret the meaning of the response categories in a way that systematically differs across populations or sub-populations stratified by socio-economic and or demographic status (Sadana et al., 2002). As an example, it is natural to think of good or poor system performance to mean different things to different people. Differential mapping from the underlying latent construct of interest to the available 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) and response category cut-point shift (Sadana, 2000). Systematic variation in reporting behaviour can be examined in relation to measured attributes of individuals such as their socio-economic characteristics. For example, income has been shown to be a determinant of reporting heterogeneity in self-reported general health status (Bago d'Uva et al. 2007). Differential reporting behaviour can be shown diagrammatically with the example as in Figure 1. Assume individuals in country A and country B are asked to rate the performance of their health systems according to a scale ranging from "very bad" to "very good". Reporting heterogeneity results in respondents in country A applying a different set of thresholds to the underlying latent construct of performance compared to respondents in country B. In country A, respondents rate the performance of their health system "good", whereas respondents in country B rate their system "very good". A casual inspection of the ratings in the two countries would suggest that individuals in country A face poorer health system performance compared to individuals in country B. However, both groups face the same underlying level of performance depicted by the solid vertical line. Correcting for differences in the use of the thresholds is fundamental to achieving comparative analysis across the two countries. The challenge is to model the positioning of the thresholds as functions of observed characteristics of the relevant populations and use this information to benchmark a comparison to a chosen threshold scale.

The method of anchoring vignettes has been promoted as a means for adjusting for systematic differences in preferences and norms across population groups responding to survey questions (for

example, see Salomon et al. (2004)). Vignettes represent hypothetical descriptions of fixed levels of a latent construct such as health system performance. Since the vignettes are fixed and pre-determined, any systematic variation across individuals in the rating of the vignettes can be attributed to differences in reporting behaviour. The response categories available to respondents when rating the vignettes are the same as those for the rating of own experiences of system performance.

Responses to vignettes allow the researcher to model the response scales, or cut-point thresholds as a function of the characteristics of respondents. Because 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 obtained from respondents on system performance. For cross-country analyses, applying the thresholds observed in one of the countries as a benchmark, responses in other countries can be re-scaled to the benchmark to provide adjusted comparable data.

In recent years anchoring vignettes have been increasingly included in surveys (for example, see http://gking.harvard.edu/vign/) and methodology for determining the form of vignette questions has been a focus of research attention. The use of many different vignettes for each self-assessment question may increase significantly the cost of data gathering and, if poorly defined, may not be informative in correcting for differential reporting behaviour. King and Wand (2007) have developed techniques based on the use of entropy measures to evaluate the information available in a set of anchoring vignettes and to detect vignettes that might not be useful for identifying reporting heterogeneity.

A number of studies have promoted the vignette approach and made use of what has been termed the hierarchical ordered probit model (HOPIT) model to adjust self-reports and make it possible to partition observed differences in self-reported responses into differences due to reporting behaviour and genuine differences in the underlying latent construct under scrutiny, in our study performance. To date, these studies have, on the whole, been interested in adjusting self-reported data on health status (for example see, Kapteyn et al. 2007, Bago d'Uva et al. 2007). In contrast we are interested in performance measured as health system responsiveness and developed by the WHO. Responsiveness is defined as aspects of the way individuals are treated and the environment in which they are treated during health system interactions (Valentine, 2003). The concept covers a set of non-clinical and non-financial dimensions of quality of care, assessed through eight domains, which reflect respect for human dignity and interpersonal aspects of the care process.

The World Health Survey (WHS)

The most ambitious attempt to date to measure and compare health systems responsiveness is the World Health Survey (WHS). The WHS is an initiative launched by the WHO 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., 2003b). The basic survey mode was an in-person interview, consisting of either: 90-minute in-household interviews (53 countries), 30-minute face-to-face interviews (13 countries) or computer assisted telephone interviews (4 countries). In total, seventy countries participated in the WHS 2002-2003. 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 surveyed. Data collection was on a modular basis covering different aspects of health and health systems, including information on health state valuation, health system responsiveness and health system goals. Samples have undergone extensive quality assurance procedures, including the testing of the psychometric properties of the responsiveness instrument. In addition to concerns about the psychometric

properties of the instruments, the WHS paid close attention to the issue of comparability in the development of the instrument. (Ustun et al., 2003b)

The WHS contains a module on health system responsiveness. This has been developed from an extensive consultation process aimed at gathering information on the aspects of the delivery of health care that individuals valued most. The resulting instrument was fielded in the WHO Multi-Country Survey Study on Health and Responsiveness (2000-2001) (MCSS - see Üstün, et al., 2003a) and a refined version of the MCSS module was incorporated in the WHS. The WHS responsiveness module gathers basic information on health care utilization for both inpatient and outpatient services. In the analysis that follows, we focus on inpatient data.[2] The measurement of responsiveness was obtained by asking respondents to rate their most recent experience of contact with the health system within a set of eight domains by responding to set questions. The domains consist of "autonomy" (involved in decisions), "choice" (of health care provider), "clarity of communication" (of health care personnel), "confidentiality" (of personal information), "dignity" (respectful treatment and communication), "prompt attention" (e.g. waiting times), "quality of basic amenities" and "access to family and community support". We use data from the short-form questionnaire within the WHS which contains one item questions per domain.[3] The following five response categories were available to respondents when rating their experience of health systems: "very good", "good", "moderate", "bad", and "very bad". The responsiveness module further includes a set of anchoring vignettes. Vignettes represent hypothetical descriptions of fixed levels of a construct, such as responsiveness, and individuals are asked to evaluate these in the same way as they evaluate their own experiences of the health system. The vignettes provide a source of external variation from which information on systematic reporting behaviour can be obtained. Valentine et al., 2003 provide descriptions of other vignettes and domains.

We describe the phenomenon of reporting behaviour and how adjusting for its determinants impacts on the ranking of country level performance using data from the WHS on 18 OECD countries that provided in household or face-to-face interviews.[4] To explore responses to questions on health system responsiveness we make use of respondent characteristics. These include age, gender, level of education and income. Level of education is measured both as a categorical variable containing 7 categories representing stage of schooling, for example, 'primary school completed', 'secondary school completed', to 'post graduate degree completed' and as a continues variable measuring the number of years of education completed. For the descriptive results we use the categorical measure of education, whilst for the ordered probit and HOPIT models, we use the number of years of education. Gender is coded 1 for women and 0 for men. Income is an indirect measure of household permanent income based on a series of questions based on ownership of physical assets. This was deemed a more appropriate comparative measure of income across the countries covered by the WHS. Details of the approach to its measurement can be found in Ferguson et al., 2003. We make us of a categorisation of the permanent income variable which indicates the quintile of the income distribution in which the measure of permanent income falls with 1 indicating the lowest quintile and 5 the highest quintile. These variables are specified as

[2] Inpatients data offer more information about the responsiveness of health care systems compared to outpatient data, as the former makes reference to eight domains while the latter only to seven.

[3] The long-form questionnaire uses two question items per domain but is not available for all the countries considered here.

[4] The 18 OECD countries are Austria, Belgium, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Mexico, Netherlands, Portugal, Slovak Republic, Spain, Sweden, United Kingdom. Australia, Luxemburg and Norway are omitted due to data potentially being of lower quality being obtained though computer assisted telephone interviews (CATI).

regressors in models to explain responsiveness and for modelling reporting behaviour through the cut-points of the HOPIT model. These variables have been used in studies investigating the determinants of reporting behaviour for self-reported measures of health (Bago d'Uva et al., 2007; Banks et al., 2006). Summary statistics for the set of variables are provided in Table 1.

Empirical methods

Responsiveness can be viewed as a multidimensional concept, with each domain measured as a categorical variable, for which there is an assumed underlying latent scale. The ordered probit model is appropriate for modelling categorical outcomes which makes use of a set of constant cut-points thresholds applicable to all individuals to map responses on a latent scale (estimated via a conditional mean function) to observed outcomes. Accordingly, the model assumes homogeneous reporting across individuals. Where this assumption does not hold, estimates of the impact of explanatory variables on outcomes of interest will be biased and, in part, will reflect differential reporting behaviour across individuals.

The ordered probit model can be extended such that the thresholds can be modelled as functions of individual characteristics. This can be achieved using, what has been termed, the hierarchical ordered probit model (HOPIT) developed by Tandon et al. (2003) (also see Terza, 1985). The method draws on the use of the anchoring vignettes to provide a source of external information that enables the identification of the thresholds as functions of relevant covariates. The model is specified in two parts: the first to identify the cut-points as a function of relevant covariates using information from the vignettes (*reporting behaviour equation*), and the second to map individual socio-economic and other characteristics to underlying self-reported health system responsiveness while controlling for reporting heterogeneity (*responsiveness equation*).

Reporting behaviour equation

To identify the cut-points as a function of respondent covariates, let R_{ik}^{v*} represent the underlying health system responsiveness for vignette k, rated by individual i such that:

$$R_{ik}^{v*} = K_{ik}\eta_k + \varepsilon_{ik}^{v}, \qquad \varepsilon_{ik}^{v} \mid K_i \sim N(0,1) \tag{1}$$

where K_{ik} is the vector of vignettes, η_k is a conformably dimensioned vector of parameters 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} (ranging from 'very bad' to 'very good'). We assume that the mapping from R_{ik}^{v*} to the observed category, r_{ik}^{v}, is through the following mechanism:

$$r_{ik}^{v} = j \quad if \quad \mu_i^{j-1} \le R_{ik}^{v*} < \mu_i^{j} \tag{2}$$

$$\text{for } \mu_i^0 = -\infty, \mu_i^5 = \infty, \forall i,k; \quad j = 1,\ldots,5$$

Where we allow the cut-point thresholds to be functions of covariates, X such that:

$$\mu_i^{j} = X_i\gamma^{j} \tag{3}$$

where μ_i^j, $j=1,\ldots 5$ are parameters to be estimated along with η_k and $\mu_i^1 < \mu_i^2 < \ldots < \mu_i^5$. If the covariates affect all thresholds by the same magnitude then we observe parallel cut-point shift. Where covariates impact differentially on the thresholds, this is refers to non-parallel shift (Jones et al. 2007).

Conditional on the thresholds, underlying health system responsiveness faced by individual i can be expressed as:

$$R_i^{s*} = Z_i\beta + \varepsilon_i^s, \qquad \varepsilon_i^s \mid Z_i \sim N\left(0, \sigma^2\right) \tag{4}$$

where Z_i represents a set of regressors predictive of responsiveness. We assume that the observed categorical response, r_i^s, relates to R_i^{s*} in the following way:

$$r_i^s = j \quad if \quad \mu_i^{j-1} \le R_i^{s*} < \mu_i^j \tag{5}$$

$$\text{for } \mu_i^0 = -\infty, \mu_i^5 = \infty, \forall i; \quad j = 1,\ldots,5$$

where μ_i^j are defined by (3) with γ^j fixed and it is assumed that R_{ik}^{v*} and R_i^{s*} are independent for all $i=1,\ldots,N$ and $k=1,\ldots,V$. Note that σ^2 in (4) is identified due to the cut-point thresholds being fixed through the reporting behaviour equation. The probabilities associated with each of the 5 observed response categories are given by:

$$\Pr(r_i = j) = \Phi(\mu_i^j - X_i\beta) - \Phi(\mu_i^{j-1} - X_i\beta), \qquad j = 1,\ldots,5 \tag{6}$$

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

The use of vignettes to identify reporting heterogeneity relies on two assumptions. Firstly, individuals classify the vignettes in a way that is consistent with the rating of their own experiences of health system responsiveness. This is termed response consistency and implies that the mapping from underlying latent responsiveness to available response categories is the same for the vignettes as that used for reporting of own experiences. Secondly, it is assumed that conditional on the socio-economic characteristics that determine reporting behaviour, the level of responsiveness faced by an individual does not influence the way s/he reports the responsiveness of the hypothetical scenarios. This is referred to as the irrelevance of own provider responsiveness or vignette equivalence.

For the set of OECD countries contained in the WHS we investigate evidence for reporting behaviour and whether this is related to socio-demographic and country characteristics. We then adjust for systematic reporting across the set of countries using the HOPIT model and compare the results to those obtained from an ordered probit model. The difference between the ordered probit and HOPIT model lies in the modelling of the cut-point thresholds so that these are assumed to be a function of individual socio-demographic and country level effects for the latter, and constant in the former. We show that modelling the thresholds using the HOPIT model reflects better the predictions of responsiveness observed in the raw data than those from the ordered probit model. By benchmarking reporting behaviour to a selected country, we then compare the ranking of performance across the 18 OECD countries with those observed in the raw data. Benchmarking to a single country enables the removal of the influence of differential country-specific reporting behaviour and hence provides a more comparable basis on which to evaluate country performance.

Results

Determinants of reporting behaviour

Figure 2 presents a summary of the proportion of respondents reporting each of the five categories of responsiveness for both own experience of contact with health services and the set of five vignettes for the domain "communication". Example data are provided for the three countries of Mexico, Spain and the UK. Since the vignettes provide descriptions of a fixed level of responsiveness, variation in reporting across respondents is assumed to be due to reporting heterogeneity. Figure 2 provides evidence of heterogeneous reporting across the three countries illustrated. For example, while the majority of respondents in Mexico rate the first vignette "good", a sizable proportion rate the vignette "very good" while others use the categories "moderate", "very bad" and "bad". Reporting for the fifth vignette shows greatest heterogeneity with all but "very good" being used by more than 10% of respondents. Similar patterns can be observed across Spain and the UK. Clearly, respondents, when faced with a vignette description apply different thresholds between response categories, resulting in a lack of consistency in reported responsiveness. Moreover, the distributions of responses for a given vignette differ across the three figures indicating heterogeneous reporting across countries. This can be seen, for example, in the reporting of the first vignette. In Mexico and Spain the majority of respondents rate the vignette "good" with "very good" being the second most frequently used category. The pattern is reversed in the UK where respondents make greatest use of the response category "very good" and less use of the category "good". Differences in the reporting of the other vignettes across the three countries are also observed.

Figure 2 further shows the reporting of respondents own experiences of contact with inpatient health services. As expected these show heterogeneity both across respondents within each country and across the three countries. While the pattern of reporting shows some similarity between Mexico and Spain, the pattern of reporting in the UK is very different. These observed differences will, in part, reflect differences in actual underlying responsiveness faced by individuals in each country, but will also reflect differences in reporting behaviour. While not shown, similar patterns to those observed within and across Mexico, Spain and the UK are observed across other countries contained in the WHS, and across the other domains of responsiveness.

We investigate the determinants of reporting behaviour by considering the differential rating of the vignettes by socio-demographic characteristics of respondents. This is shown graphically in Figure 3 for the domain "communication" where responses to the second vignette are summarised by income quintile, educational attainment, age and gender for data for Mexico. If we consider educational attainment, there is a clear gradient in the use of response categories as we move from completion of primary school to completion of a degree. In general, better educated respondents make greater us of the category "very good" and less use of the categories "good", "moderate" and "bad". Similarly, we observe a gradient across income quintiles. Richer respondents, in general, make greater use of the response category "very good" and less use of the remaining categories. Interestingly there appears to be little difference in reporting by gender with equal use made of the available response categories by both men and women. There is some indication of reporting behaviour differing by age with older respondents tending to make less use of the category "very good" and greater use of the alternative categories compared to younger respondents. Overall, Figure 3 indicates reporting behaviour to be related to educational attainment and income and to a lesser extent age and gender. While the degree of reporting heterogeneity and how this relates to socio-demographic characteristics varies across countries and domains of responsiveness, it can be expected to extend across all the OECD countries considered here.

Figures 2 and 3 provide prima facia evidence of differential reporting behaviour across countries and across socio-demographic groups within countries. Using information on respondent ratings of the vignettes, we model reporting behaviour more formally by incorporating the socio-demographic characteristics as covariates when modelling the thresholds in the first-stage of the HOPIT model. A tests of the significance of the resulting estimated coefficients, $\hat{\gamma}^j$, under the null hypothesis of homogenous reporting, provides evidence of reporting behaviour varying by respondent characteristics. This information can subsequently be included in the second-stage of the HOPIT model to adjust self-reported experiences of health service responsiveness for systematic reporting behaviour. The HOPIT model is estimated separately for each of the eight domains and for each of the three countries considered. Age, gender, years of education and income quintiles are specified as regressors in both the index function for responsiveness (Z) and in the modelling of the cut-point thresholds (X). The null hypothesis of homogeneous reporting (indicating that reporting is not a function of the characteristic) is supported by a lack of significance attached to the respective estimated coefficient when included as a regressor in the thresholds.

For both Spain and Mexico a joint test across all characteristics rejects the null in favour of reporting heterogeneity being related to socio-demographic characteristics. For the UK, the null is rejected for only four of the eight domains. The major determinant of reporting behaviour for Spain appears to be income where the majority of coefficients are significantly different to zero. Gender appears to contribute significantly for the domain "confidentiality" only and age for the domains of "quality of basic amenities" and "access to family and community support". For the UK, income appears to be a less prominent determinant of reporting behaviour and is significant for the domain "autonomy" alone. Gender, age and years of education appear significant across two of the eight domains. Mexico exhibits the strongest relationship between socio-demographic characteristics and reporting behaviour with significant effects observed for the majority of domains and socio-demographic characteristics. It should be noted, however, that the sample size available for Mexico is larger than that for other countries, enhancing the ability to observe effects as statistically significant.

Adjusting for reporting behaviour

The results above indicate that differential reporting behaviour is observed in the WHS responsiveness module and that it is a function of socio-demographic characteristics of respondents. In particular, reporting appears to be related mostly to income and education. The exact nature of reporting, however, varies by country. For Spain, we observe reporting behaviour to be related to income quintile for the majority of domains. However, reporting behaviour is a function of income for only a single domain in the UK. For Mexico, reporting behaviour is found to be statistically related to all socio-demographic characteristics for the majority of domains.

The existence of differential reporting behaviour renders the comparison of health system performance problematic due to self-reported information conflating 'true' underlying levels of responsiveness with reporting behaviour. Within a specific country, the latter appears to vary by socio-demographic characteristics of respondents. Reporting behaviour is likely to be more pronounced, however, when comparing across countries due to diverse cultural differences, expectations and norms. In an analysis pooled across the 18 OECD countries we attempt to capture country-level differences in reporting by extending the specification of the thresholds to be a function of country-specific dummy variables and their interaction with the income quintiles.[5] These are in addition to the set of socio-demographic characteristics used in the modelling of country-

[5] Models where the specification of the thresholds include interaction terms between country dummies and other socio-demographic characteristics failed to achieve convergence.

specific responsiveness. The predictive index function, $Z\beta$, of the responsiveness model is specified as a function of the individual specific characteristics, country dummy variables and interactions between the country dummies and the set of individual characteristics.[6] Table 2 shows the coefficients and standard errors for the full set of country dummy variables used to model the cut-point thresholds for "prompt attention". Compared to the baseline country of Mexico, the vast majority of coefficients for the first threshold are positive (Austria being the exception), and the majority of coefficients for the fourth threshold are negative (Italy and Portugal are exceptions). Most coefficients are statistically significant at conventional levels. Taken together these results imply that compared to Mexico, countries are likely to make more use of the extremes of the available reporting categories and rate a given level of responsiveness either as "very bad" or "very good" (this situation is illustrated in Figure 1 for country B compared to country A). The results illustrate that reporting behaviour varies across countries, albeit captured in a fairly rudimentary way by means of country dummy variables.[7]

Cross-country comparison

We investigate the effect of adjusting the data for reporting behaviour by comparing predictions of reporting "very good" responsiveness from an ordered probit and the HOPIT model. Table 3 compares observed and predicted levels of reporting "very good" responsiveness for the three domains "prompt attention", "dignity" and "communication". Predictions are derived from an ordered probit model which makes no adjustment for differential reporting heterogeneity, the HOPIT model with correction for country-level reporting and the HOPIT model benchmarked to the reporting behaviour observed in Mexico.[8] Countries are ranked in order of decreasing frequency of reporting "very good" responsiveness. For each domain, the first column reports the raw frequencies from respondent ratings observed in the data. These vary substantially. For example, for "prompt attention", 56% of respondents in Denmark compared to 13% of respondents in Portugal report "very good" responsiveness. This variation, in part, will reflect differences in true underlying health system responsiveness faced by individuals, but will also partly reflect systematic reporting behaviour that differs across countries. The challenge for comparative analysis is to isolate the impact of the former, from the latter.

The second column of Table 3 reports predicted frequencies from an ordered probit model. This model includes the socio-demographic variables together with a set of country dummy variables and interaction terms between country dummies and the socio-demographic variables. Estimation is undertaken on data pooled across all countries and assumes a set of fixed cut-point thresholds common to all countries.[9] If we consider the ranking of the countries, overall there is little change between the observed frequencies derived from the raw data and the predictions from the ordered probit model. Austria and Denmark are ranked highly; countries such as Ireland, France and Finland tend to be placed in the middle of the rankings; and Mexico, Italy and Portugal tend to be toward the

[6] Due to the enhanced precision with which estimates of responsiveness can be estimated, Mexico is used as the baseline country.

[7] Individual socio-demographic variables were significant in the 1st, 2nd and 4th threshold. The significance of the interaction terms between country dummy variables and income quintiles varied by cut-point and country.

[8] An alternative to predicting a response category such as 'Very good' is to report the latent value of responsiveness, R^*, derived from the respective linear index of the responsiveness equation for the ordered probit and HOPIT models. This will lead to the same ranking of countries as the ranking of the proportion reporting 'very good' responsiveness, but as it is measured on a latent scale is more difficult to interpret intuitively.

[9] That is, the cut-points are not functions of covariates or country-specific dummy variables.

bottom of the rankings. A test of the independence of the rankings[10] rejects the null in favour of dependence ($p < 0.01$ for all domains), indicating an association between the rankings from the ordered probit model and the observed frequencies. There are, however, some notable differences in the rankings of individual countries. For example, for the domain "confidentiality", Hungary falls four places and Finland three places in the rankings, while Spain is raised three places. Further, while we are primarily interested in the ranking of countries, the predictions of the proportion of individuals reporting "very good" responsiveness does not correspondent particularly closely with the raw observed frequencies. This is due, largely, to the ordered probit applying a common fixed set of cut-point thresholds across all countries.

Columns (3) and (4) of Table 3 present predicted frequencies obtained from the HOPIT model. The model contains the same set of explanatory variables specified in the ordered probit model. In addition, the cut-point thresholds are modelled as functions of age, gender, education and income quintiles, together with the set of country dummy variables and interaction terms between the country dummies and income quintiles. The modelling of the cut-points allows us to control for differential reporting behaviour across individuals within countries (via socio-demographic characteristics) and across countries (country dummy variables and interaction terms). The results presented in column (3) represent country-specific predictions obtained from applying the relevant parameters in the thresholds applicable to a particular country. These results adjust for within-country reporting behaviour but do not correct for cross-country differences in reporting. Due to the use of country-specific thresholds, the predicted frequencies should more closely resemble the frequencies observed in the raw data than the corresponding predicted frequencies from the ordered probit model. This is supported by the results where we observe a similar ranking of countries to those found in the raw data. For "confidentiality", the ranking of Hungary, Finland and Spain are more closely related to the raw frequencies than the ranking from the ordered probit model. The predicted frequencies of reporting "very good" responsiveness resemble closely the observed frequencies and a test of the independence of the rankings is firmly rejected.

To provide rankings comparable across countries we benchmark reporting behaviour to a baseline country, in this case Mexico. We then predict the reporting of "very good" responsiveness for all respondents, irrespective of country of residence, as if they had the reporting behaviour of Mexican respondents. That is, for each country the predicted probability of reporting "very good" responsiveness is computed using the thresholds estimated for Mexico. By removing country-specific reporting behaviour and instead adopt the reporting characteristics for a specific given country we define a comparable basis on which to rank countries. The resulting ranking of countries is provided in column (4) of Table 3 for each of the three domains illustrated. Inspection of these results reveals a different ranking to that observed for the raw frequencies. For example, for waiting times, Denmark falls nine places in the rankings whilst Spain is raised eight places. The predictions are generally smaller for this model compared to the observed frequencies. This is a direct consequence of benchmarking to Mexico, which is a country with a low frequency of reporting "very good" responsiveness relative to other countries. For "prompt attention" and "confidentiality", tests of the independence of the rankings fails to reject the null implying different orderings of the countries before and after adjustment for reporting behaviour. While the same test rejects the null of independence for the domain of "dignity", a visual inspection of the rankings also reveals large differences. For example Hungary falls seven and the Czech Republic four places, while Belgium is raised six places and Finland five places.

[10] Kendall's tau rank correlation (Kendall, 1938) is used as a measure of the degree of correspondence between the rankings. Perfect agreement leads to a coefficient of 1, perfect disagreement -1, and independence 0. The test is under the null hypothesis of independence.

The results presented above relate to three of the eight domains of health system responsiveness. We can assess the overall impact of adjusting for reporting behaviour by aggregating the observed and predicted frequencies across all eight domains of responsiveness. The results of applying a simple aggregation, providing equal weight to each of the eight domains and computing the average, are provided in Table 4.[11] For brevity, we present the observed frequencies with those derived from the HOPIT model benchmarked to Mexico. These results also show considerable movement in the rankings of certain countries when comparing the observed and benchmarked orderings. For example, Greece falls six places, Ireland falls eight places, Hungary falls six places, whereas France, Spain and the Netherlands are raised four places, and Finland six places.

Conclusions

Cross-country comparison provides the opportunity for evaluative analysis of health systems performance and is a fundamental lever for policy development. Increasingly international organisations and national agencies turn to comparative data to shape and assess their strategic priorities and policies. International comparison, however, is beset with methodological difficulties. An increasing reliance on individual-level survey data based on respondent self-reports of system performance presents additional challenges. In particular, self-reported data is likely to suffer from the existence of systematic reporting behaviour. This might be evident both across individuals, stratified by socio-demographic characteristics, within countries and across countries. Systematic reporting behaviour, or 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 by different respondents. For comparable analysis to identify true underlying differences in performance, measures of performance need to be purged of systematic reporting behaviour. Using the methodology of anchoring vignettes we have illustrated the extent of reporting heterogeneity in the WHS and how this varies across socio-demographic characteristics of respondents. Further, reporting behaviour is extended to cross-country comparison where cultural norms and expectations of contact with health systems lead to systematic variation in the rating of system performance. A ranking of OECD countries based on raw data on health system responsiveness leads to marked differences compared to the ranking obtained once the data is adjusted for differential reporting behaviour.

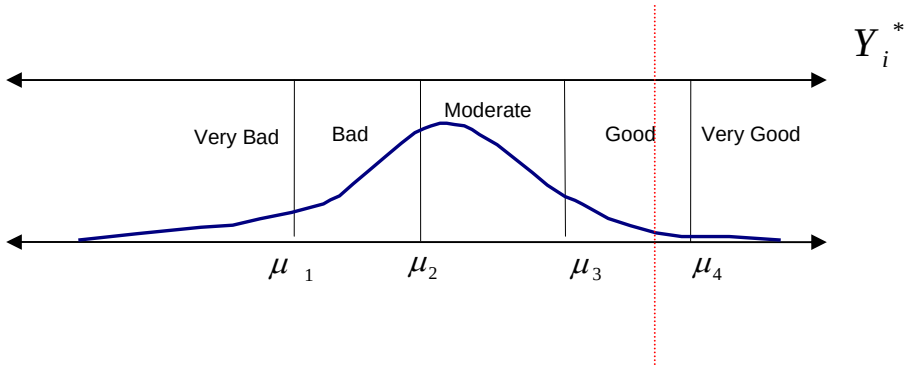
An important consideration when undertaking vignette adjustment is the ability to explain both the underlying latent level of the construct under examination, here health system responsiveness, and any associated reporting behaviour. A failure to correctly describe reporting behaviour (by modelling the thresholds) will lead to biased estimates of the level of responsiveness and biased predictions of response categories. Failure to adequately describe the level of responsiveness (by modelling the mean function) will further result in poor predictions. For the HOPIT approach to work well, the specified model must be capable of adequately predicting the observed country-level raw levels of performance. Only when the model is satisfactory, should benchmarking against a baseline country be performed and comparative analysis undertaken. Changes in the ranking of countries, compared to the ranking obtained from the raw data, might then be appropriately circumscribed to adjustment for systematic reporting behaviour. If the model proves unsatisfactory, then any observed changes in rankings are likely to have resulted both from an adjustment for differential reporting behaviour and an inability of the specified model, prior to benchmarking, to

[11] Affording equal weight to each domain provides a simple and fairly rudimentary means to aggregate across domains. Alternative weights, perhaps making use of respondent's ratings of the relative importance of the domains, might be developed for future use.

predict well the observed raw data. The set of variables extracted from the WHS and used here are arguably better predictors of reporting behaviour (and have been used elsewhere to this effect) than underlying health system responsiveness. However, the HOPIT model appears to predict the raw data sufficiently well to allow benchmarking and comparative analysis. Future research might focus on the appropriate determinants of health system responsiveness to further aid cross-country comparison.

Our results suggest that once differential reporting behaviour has be accounted for and country-specific performance is benchmarked against a single country, in general, health system responsiveness is rated more favourably in OECD countries from Northern and Western Europe, than from Eastern and Southern Europe.[12] While the ranking of Northern European countries (with the exception of Ireland) does not change notably before and after adjustment for reporting behaviour, in general, Western European countries appear to down-report performance and Eastern European countries appear to over-report performance. Accordingly these groups of countries tend to move up and down respectively in the rankings once systematic reporting behaviour has been removed. For Southern European countries, Greece appears to rate health system responsiveness highly but falls down the rankings once benchmarked to Mexcio. In contrast Spain and Portugal rate performance poorly and make modest improvements post adjustment for differential reporting.

Country A



Country B

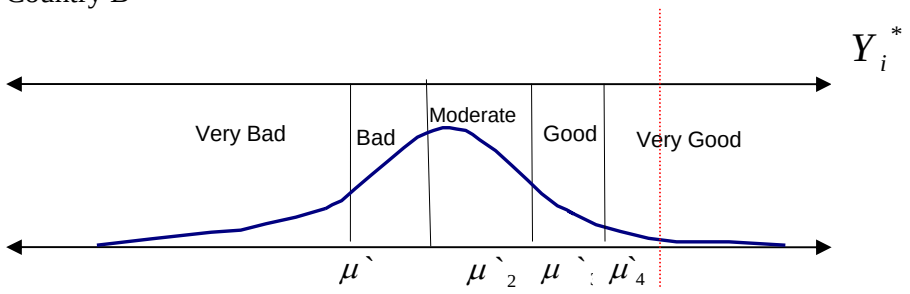
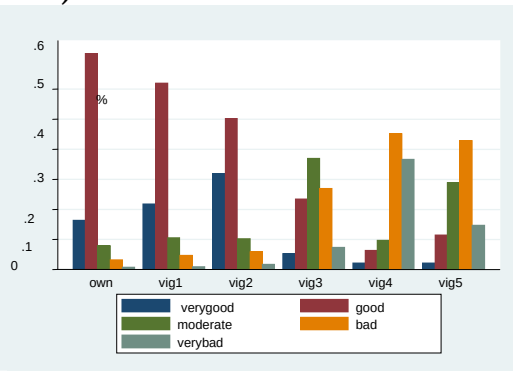


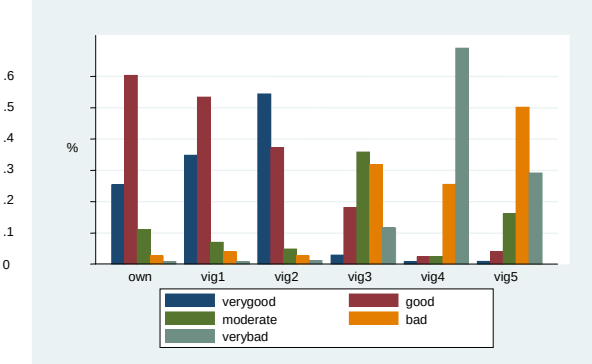
Figure 1:

[12] The main exception is Ireland which is ranked near the bottom of the distribution.

a) Mexico



b) Spain



c) UK

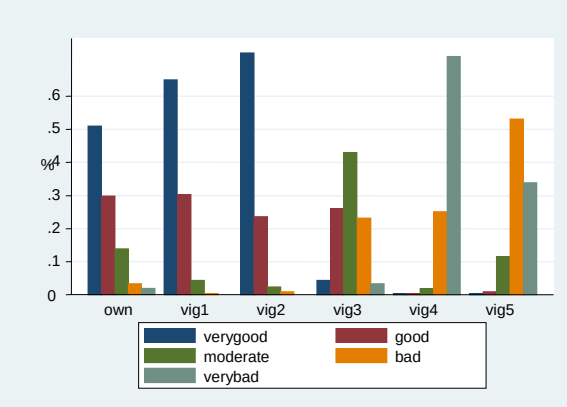
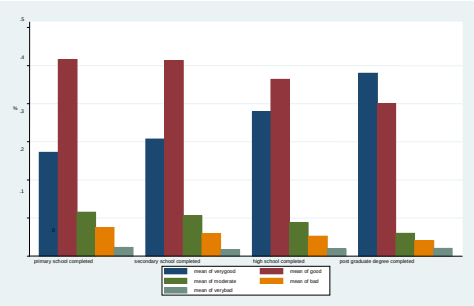
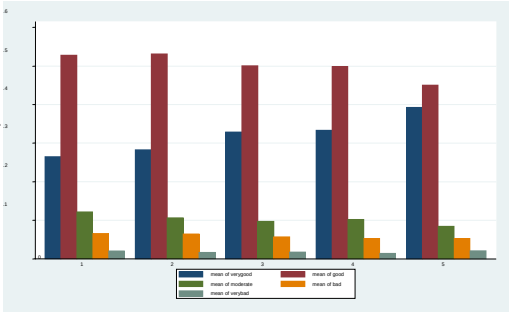


Figure 2: Summary frequencies for the reporting of responsiveness and vignettes for the domain of *Clarity of Communication*

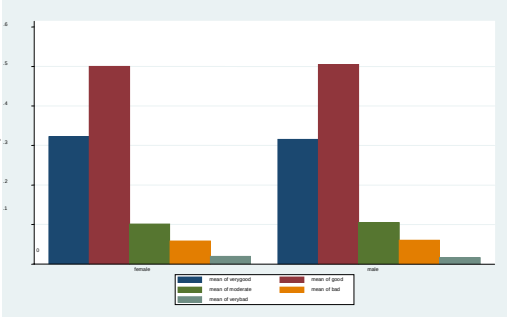
(a) Education



(b) Income quintiles



(c) Gender



(d) Age

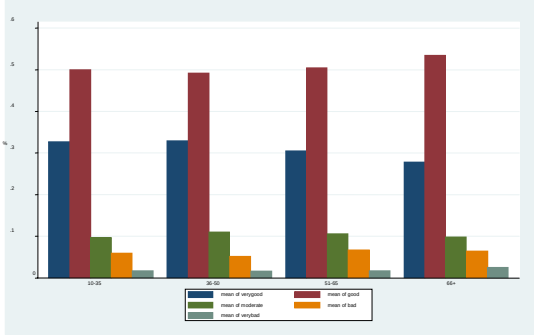


Figure 3: Vignette ratings for *Clarity of Communication* by socio-demographic characteristics of the respondents - Mexico

Country		n	Age Mean (sd)	Gender % female	Education Mean (sd)	Permanent Income†
AUT	Austria	1,055	45 (16)	62	11 (3)	.702
BEL	Belgium	1,012	45 (17)	56	14 (4)	.791
CZE	Czech Republic	1,918	47 (18)	53	12 (3)	.069
DEU	Germany	1,259	50 (18)	60	11 (3)	.327
DNK	Denmark	1,003	51 (17)	53	12 (4)	.908
ESP	Spain	12,023	53 (18)	59	9 (5)	.311
FIN	Finland	1,013	53 (17)	55	12 (4)	.703
FRA	France	1,008	44 (17)	60	14 (4)	.600
UK	United Kingdom	1,200	50 (19)	63	12 (3)	.641
GRC	Greece	1,000	51 (19)	50	10 (4)	.126
HUN	Hungry	1,419	50 (18)	58	12 (4)	-.195
IRL	Ireland	1,014	44 (17)	55	13 (3)	.631
ITA	Italy	1,000	48 (18)	57	11 (5)	.749
MEX	Mexico	40,000	41 (17)	58	7 (5)	-.776
NLD	Netherlands	1,091	44 (18)	67	13 (4)	.713
PRT	Portugal	1,030	51 (19)	62	7 (4)	-.091
SVK	Slovak Republic	2,539	39 (15)	61	13 (3)	.064
SWE	Sweden	1,000	51 (18)	58	12 (4)	.669

Table 1: Descriptive statistics

† Permanent income is a derived variable from household assets measured on a latent scale. We use income quintiles in our analyses

	Cut-Point Thresholds			
Country	1	2	3	4
AUT	-.272 (.181)	-.351 (147)	-.408 (.141)	-.641 (.145)
BEL	.408 (.144)	.017 (.138)	-.067 (.138)	-.297 (.143)
CZE	.329 (.132)	.043 (.122)	.098 (.121)	-.435 (.124)
DEU	.388 (.116)	-.132 (.112)	-.178 (.112)	-.530 (.115)
DNK	.533 (.138)	.394 (.131)	-.014 (.134)	-.822 (.140)
ESP	.392 (.052)	.335 (.048)	.119 (.048)	-.031 (.053)
FIN	.267 (.107)	.149 (.097)	.098 (.097)	-.282 (.102)
FRA	.229 (.164)	.124 (.153)	-.051 (.155)	-.554 (.160)
UK	.342 (.110)	.062 (.102)	-.034 (.102)	-.580 (.106)
GRC	.283 (.105)	-.039 (.098)	-.078 (.100)	-.377 (.105)
HUN	.604 (.082)	.195 (.079)	.002 (.080)	-.539 (.083)
IRL	.258 (.141)	.106 (.124)	-.146 (.124)	-.879 (.127)
ITA	.565 (.227)	.273 (.218)	-.171 (.217)	.070 (.238)
NLD	.356 (.163)	-.055 (.155)	-.065 (.154)	-.322 (.159)
PRT	.254 (.142)	.152 (.133)	.526 (.136)	.588 (.169)
SVK	.313 (.088)	.236 (.081)	.079 (.082)	-.400 (.087)
SWE	.352 (.109)	.056 (.103)	-.067 (.105)	-.557 (.108)

Table 2: Prompt Attention (waiting times): Coefficients and standard errors of cut-points as a function of country dummy variables.

Rank	Prompt Attention (Waiting time)				Dignity (Respect)				Confidentiality (Talk Privately)			
	Ex-ante frequency	Ordered Probit	HOPIT	HOPIT Benchmark MEX	Ex-ante frequency	Ordered Probit	HOPIT	HOPIT Benchmark MEX	Ex-ante frequency	Ordered Probit	HOPIT	HOPIT Benchmark MEX
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	DNK (56)	AUT (52)	AUT (55)	GRC (32)	AUT (62)	AUT (56)	DNK (61)	DNK (55)	DNK (51)	GRC (42)	DNK (51)	GRC (31)
2	AUT (55)	DNK (42)	DNK (52)	AUT (23)	DNK (61)	DNK (56)	AUK (60)	FIN (55)	AUT (48)	AUT (40)	GRC (51)	SWE (26)
3	UK (48)	GRC (40)	UK (48)	NLD (23)	SWE (56)	SWE (49)	SWE (55)	SWE (54)	GRC (48)	DNK (39)	AUT (47)	FRA (25)
4	CZE (46)	UK (40)	CZE (47)	ITA (21)	CZE (52)	CZE (47)	CZE (53)	BEL (43)	SWE (40)	BEL (32)	SWE (44)	UK (21)
5	GRC (44)	CZE (38)	GRC (46)	CZE (20)	UK (51)	UK (47)	UK (53)	FRA (40)	BEL (39)	SWE (31)	IRL (43)	PRT (21)
6	SWE (43)	BEL (35)	SWE (43)	BEL (19)	GRC (51)	FIN (44)	GRC (51)	UK (39)	IRL (39)	UK (28)	UK (41)	DEU (20)
7	BEL (43)	NLD (34)	BEL (42)	UK (18)	FIN (49)	GRC (43)	FIN (47)	NLD (37)	UK (39)	FRA (27)	BEL (39)	NLD (20)
8	NLD (41)	SWE (32)	IRL (41)	ESP (18)	HUN (48)	HUN (39)	HUN (46)	CZE (32)	HUN (36)	IRL (27)	HUN (37)	BEL (19)
9	IRL (38)	SVK (28)	SVK (37)	MEX (17)	FRA (45)	FRA (39)	FRA (44)	AUT (28)	FRA (34)	DEU (24)	CZE (36)	DNK (19)
10	FRA (34)	IRL (27)	NLD (36)	DNK (16)	BEL (44)	BEL (38)	IRL (44)	GRC (26)	CZE (34)	CZE (24)	FRA (35)	AUT (18)
11	FIN (31)	DEU (26)	FRA (34)	SVK (16)	IRL (44)	IRL (34)	BEL (44)	IRL (26)	FIN (31)	ESP (23)	FIN (33)	FIN (18)
12	SVK (31)	FRA (24)	DEU (32)	DEU (16)	DEU (35)	NLD (33)	NLD (35)	ESP (25)	DEU (28)	HUN (23)	DEU (29)	ESP (17)
13	HUN (31)	FIN (23)	HUN (31)	PRT (16)	NLD (35)	ESP (32)	DEU (34)	DEU (23)	NLD (24)	NLD (21)	SVK (28)	MEX (16)
14	DEU (31)	HUN (20)	FIN (29)	SWE (14)	ESP (31)	DEU (29)	ESP (33)	PRT (22)	ESP (22)	FIN (21)	NLD (28)	ITA (13)
15	ITA (22)	ESP (20)	ESP (23)	FRA (13)	SVK (28)	SVK (23)	SVK (32)	HUN (21)	SVK (22)	MEX (19)	ESP (26)	IRL (11)
16	ESP (21)	MEX (19)	ITA (19)	FIN (13)	ITA (26)	MEX (20)	ITA (25)	MEX (18)	ITA (17)	SVK (17)	PRT (17)	CZE (11)
17	MEX (14)	ITA (19)	MEX (17)	HUN (12)	PRT (19)	PRT (20)	PRT (20)	SVK (10)	PRT (15)	PRT (17)	ITA (17)	HUN (9)
18	PRT (13)	PRT (13)	PRT (15)	IRL (11)	MEX (16)	ITA (19)	MEX (18)	ITA (10)	MEX (14)	ITA (11)	MEX (16)	SVK (7)
Kendall's Tau		(2) & (1)	(3) & (1)	(4) & (1)		(2) & (1)	(3) & (1)	(4) & (1)		(2) & (1)	(3) & (1)	(4) & (1)
Ho: Independent												
Test statistic		0.830	0.882	0.216		0.922	0.961	0.529		0.752	0.909	0.242
P-value		<0.01	<0.01	0.225		< 0.01	<0.01	<0.01		< 0.01	<0.01	0.173

Table 3:
Note: For the HOPIT model to ensure the predictions are benchmarked to a constant set of cut-points we apply those of Mexico

Rank	Ex-ante	HOPIT Benchmark MEX
	(1)	(2)
1	AUT (56)	BEL (30)
2	DNK (51)	AUT (27)
3	GRC (46)	DNK (27)
4	BEL (46)	SWE (27)
5	UK (45)	FIN (26)
6	CZE (45)	FRA (26)
7	SWE (43)	UK (24)
8	IRL (42)	GRC (22)
9	HUN (40)	NLD (21)
10	FRA (40)	DEU (20)
11	FIN (35)	CZE (20)
12	DEU (33)	ESP (18)
13	NLD (27)	MEX (17)
14	SVK (23)	PRT (17)
15	ITA (23)	HUN (16)
16	ESP (22)	IRL (16)
17	PRT (16)	ITA (14)
18	MEX (15)	SVK (9)

Kendall's Tau	(2) & (1)
Ho: Independent	
Test statistic	0.4967
P-value	0.0042

Table 4: Rankings of health system responsiveness across all domains

REFERENCES

Anderson, G. F., B.K. Frogner, U.E. Reinhardt (2007). "Health spending in OECD countries in 2004: an update " Health Affairs 26(5): 1481-1489.

Anell, A., M. Willis (2000). "International comaprison of health care systems using resource profiles." Bulletin of the World Health Organization 78(6): 770-778.

Arah, O. A., N.S. Klazinga, D.M.J. Delnoij, A.H.A. Ten Asbroek, T. Custers (2003). "Conceptual frameworks for health systems performance: a quest for effectiveness, quality, and improvement." International Journal for Quality in Health Care 15(5): 377-398.

Bago d'Uva, T. v. D., E., Lindeboom, M., O'Donnell, O., Chatterji, S (2007). "Does reporting heterogeneity bias the measurement of socio-economic inequalities in health?" Health Economics 17(3): 351-375.

Balk, B. (2001). Aggregation methods in international comparisons: what have we learned?, Erasmus University, Rotterdam.

Banks, J., Marmot M., Oldfield Z., Smith J.P (2006). "Disease and Disadvantage in the United States and in England " The Journal of the American Medical Association 295: 2037-2045.

Brislin, R. W. (1986). The wording and traslation of research instruments. Field methods in cross-cultural research. W. J. Lonner, Berry, J.W. Beverly Hills, CA, Sage 137-164.

Bullinger, M. (2003). International comparability of health interview surveys: An overview of methods and approaches. EUROHIS: Developing Common Instruments for Health Surveys. C. G. A. Nosikov, IOS Press.

Coulter, A., Magee, H (2003) "The European patient of the future (state of health)" Open University Press, Maidenhead.

Coulter, A., Cleary, P.D., (2001) "Patients' experiences with hospital care in five countries" Health Affairs, 20(3): 244-252.

de Groot, W. (2000). "Adaptation and scale of reference bias in self-assessments of quality of life." Journal of Health Economics 19: 403-420.

Ferguson, B. D., Tandon A., Gakidou E., Murray C.J.L. (2003). Estimating Permanent Income using Indicator Variables. Health systems performance assessment: debates, methods and empiricism. E. Murray C.J.L., D.B. Geneva, World Health Organisation: 748-760.

Gonzalez Block, M. A. (1997). "Comparative research and analysis methods for shared learning from health system reforms." Health Policy 42: 187-209.

Gravelle, H., M.E. Backhouse (1987). "International cross-sectional analysis of the determination of mortality " Social Science & Medicine 25(5): 427-441.

Hill, I., B. Courtot, J. Sullivan (2007). "Coping with SCHIP enrollment caps: lessons from seven states` experiences " Health Affairs 26(1): 258-268.

Johnson, T. D. O. R. N. C., S. Sudman, R. Warnecke, L. lacey, J. Horm (1997). Social cognition and responses to surevey questions among culturally diverse populations. Survey Measurement and Process Quality. L. Lyberg, Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz N., Trewim, D. New York, John Wiley & Sons: 87-113.

Johnson, T. J. (1998). "Approaches to Equivalence in Cross-Cultural and Cross-National Survery Research." Zuma Nachrichten Special.

Jones, A. M., Rice, N., Bago d'Uva, T., Balia, S. (2007). Applied Health Economics. Abingdon, Routlegde.

Kapteyn, A. S. J. P., van Soest, A. (2007). "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands." American Economic Review 97(1): 461-473.

Kendall, M. G. (1938). "A new measure of rank correlation." Biometrika 30(12): 81-93.

Kerkhofs, M., Lindeboom, M. (1995). "Subjective health measures and state dependent reporting errors." Health Economics 4: 221-235.

King, G., J. Wand (2007). "Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes." Political Analysis 15: 46-66

Krause, N. M., Jay, G.M. (1994). "What do global self-rated health items measure?" Medical Care 32: 930-942.

Kumar, A., L. Ozdamar (2004). "International Comparison of Health Care Systems." International Journal of The Computer, the Internet and Management 12(3): 81-95.

Lynn, P., L. Japec, L. Lyberg (2005). What`s so special about cross national surveys. 2005 Meeting of the International Workshop on Comparative Survey Design and Implementation

McKay, R. B., Breslow, M.J., Sangster, R.L, Gabbard, S.M., Reynolds, R.W., Nakamoto, J.M., Tarnai, J. (1996). "traslating survey questionnaires: lesson lerned." New Directions and Evaluation 70: 93-104.

Nixon, J., P. Ulmann (2006). "The relationship between health care expenditure and health outcomes." European Journal of Health Economics 7: 7-18.
O` Mahony, M., P.A. Stevens (2004). International comparisons of performance in the provision of public services: outcome based measures for Education, National Institute of Economic and Social Research, London.
Okazaki, S., S. Sue (1995). "Methodological issues in assessment research with ethnic minorities " Psychological Assessment 7: 367-375
Ovretveit, J. (1998). Comparative and cross-cultural health research, Radcliffe Medical Press.
Reinhardt, U.E., Hussey, P.S., Anderson, G.F. (2002) "Cross-national comparisons of health systems using OECD data, 1999. Health Affairs, 21(3): 169-181.
Sadana, R., Mathers C.D., Lopez A.D., Murray C.J.L., Iburg K.M. (2002). Comparative analyses of more than 50 household surveys on health status. Summary measures of population health: concepts, ethics, measurement and applications. M. C. e. al. Geneva, World Health Organization: 369-386.
Salomon, J. A., A. Tandon, C. J. L. Murray (2004). "Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes." BMJ.
Scheuch, E. K. (1993). "The cross-cultural use of sample surveys: Problems of comparability." Historical Social Research 18: 104-138.
Schieber, G. J., J.P. Poullier (1991). "International Health Spending: Issues and Trends." Health Affairs: 106-116.
Schreyogg, J., Tiemann O., Stargardt T., Busse R. (2008). "Cross-country comparisons of costs: the use of episode-specific transitive purchasing power parities with standardised cost categories." Health Economics 17: s95-s103.
Shoen, C., R. Osborn, P.T. Huynh, M. Doty, M. Bishop, J. Peugh, K. Zapert (2006). "On the Front Line of Care: Primary Care Doctors` Office Systems, Experiences, and Views in Seven Countries." Health Affairs 25: w555-w571.
Shoen, C., R. Osborn, M.M. Doty, M. Bishop, J. Peugh, N. Murukutla (2007). "Toward Higher-Performance Health Systems: Adults` Health Care Experiences in Seven Countries, 2007." Health Affairs 26(6): w717-w734.
Siciliani, L., Hurst, J. (2005). "Tackling excessive waiting times for elective surgery: a comparative analysis of policies in 12 OECD countries." Health Policy 75: 201-215.
Tandon, A., Murray, CJL., Saloman, JA., King, G. (2003). Statistical models for enhancing cross-population comparability. Health systems performance assessment: debates, methods and empiricism. E. Murray CJL., DB. Geneva, World Health Organisation: 727-746.
Terza, J. V. (1985). "Ordinal Probit: a generalization." Communication in Statistics 14(1): 1-11.
Ustun, T. B., M. Villanueva, L. Benib, C. Celik, R. Sadana, N. B. Valentine, J. P. Ortiz, A. Tandon, J. Salomon, C. Yang, W. J. Xie, E. Ozaltin, C. D. Mathers and C. J. L. Murray (2003a). WHO Multi-Country Survey Study on Health and Responsiveness 2001-2. Health systems performance assessment: debates, methods and empiricism. E. Murray CJL., DB. Geneva, World Health Organisation: 762-796.
Üstün, T. B., Chatterji, S., Mechbal, A., Murray, CJL., WHS Collaborating Groups (2003b). The World Health Surveys. Health systems performance assessment: debates, methods and empiricism. E. Murray CJL., DB. Geneva, World Health Organisation: 797-808.
Valentine, N.B., A. De Silva, K. Kawabata, C. Darby, CJL Murray., D. Evans, (2003). Health system responsiveness: concepts, domains and operationalization. Health systems performance assessment: debates, methods and empiricism. E. Murray CJL., DB. Geneva, World Health Organisation: 573-596.
Van Doorslaer, E., Masseria, C. (2004) "Income-related inequality in the use of medical care in 21 OECD countries" OECD Health Working Paper No 14, OECD Paris.
White, C. (2007). "Health care spending growth: how different is the United States from the rest of the OECD?" Health Affairs 26(1): 154-161.
Willcox, S., M. Seddon, S. Dunn, R. T. Edwards, J. Pearse, J.V. Tu (2007). "Measuring and Reducing Waiting Times: A Cross-National Comparison of Strategies." Health Affairs 26(4): 1078-1087.
Yepes, F. J. (1991). "Comparative analysis of health systems: Methodological aspects." Salud Publica de Mexico 33: 392-395
Young, F. W. (2001). "An explanation of the persistent doctor-mortality association." J. Epidemiol. Community Health 55: 80-84.