Soc Indic Res (2012) 105:255–271
DOI 10.1007/s11205-011-9880-z

Comparability of Health Care Responsiveness in Europe

Nicolas Sirven · Brigitte Santos-Eggimann · Jacques Spagnoli

Accepted: 27 May 2011 / Published online: 10 June 2011

Abstract The aim of this paper is to measure and to correct for the potential incomparability of responses to the SHARE survey on health care responsiveness. A parametric approach based on the use of anchoring vignettes is applied to cross-sectional data (2006–2007) in eleven European countries. More than 7,000 respondents aged 50 years old and over were asked to assess the quality of health care responsiveness in three domains: waiting time for medical treatment, quality of the conditions in visited health facilities, and communication and involvement in decisions about the treatment. Our results suggest that there is reporting heterogeneity across countries and across individuals within countries, and the degree of heterogeneity varies with the health care domain. Although leading countries in terms of health care responsiveness remain among the most successful even after correction for reporting heterogeneity, one may acknowledge many shifts in the ranking of the other countries.

Keywords Anchoring vignettes · Cross-country comparison · Chopit model

1 Introduction

Cross-national evaluations of health care systems have long relied essentially on indicators of expenditure (e.g. proportion of Gross National Product invested in health) and health (e.g. life expectancy at birth, level of avoidable mortality, subjective health), with health considered as the main, if not only, outcome. More recently, health care responsiveness, or

N. Sirven (✉)
IRDES Institute for Research and Information on Health Economics,
10, Rue Vauvenargues, 75018 Paris, France
e-mail: sirven@irdes.fr

B. Santos-Eggimann · J. Spagnoli
IUMSP Institute of Social and Preventive Medicine University of Lausanne,
52, Route de Berne, 1010 Lausanne, Switzerland

the extent to which the process of care delivery matches patients' expectations (Murray and Frenk 2000), was added as an important criterion for evaluating health care systems and specific indicators were integrated in the WHO World Health Report 2000 (WHO 2000). Health survey respondents are now increasingly asked about their experience of access to care. A range of dimensions have been identified for the responsiveness concept: respect of autonomy, confidentiality, dignity, prompt attention, communication, social consideration, as well as the quality of basic amenities, the choice and continuity of care (WHO 2002).

The evidence of differences in responsiveness of national health care systems (Coulter and Jenkinson 2005; Schoen et al. 2004) points at a potential for improvements in the quality of health care delivery. Geographic variation must, however, be interpreted with caution due to the subjectivity of questions on access to and quality of health care. Like many other subjective variables, responsiveness evaluations may be influenced by expectations, and expectations and other factors driving satisfaction with access or responsiveness may be influenced by a range of country specific factors, necessitating some adjustment in order to be able to draw conclusions about genuine differences in health care responsiveness. Responses to subjective questions are also likely to reflect different 'response styles' across regions or countries, reflecting historical, cultural or environmental circumstances (Hausdorf et al. 2008).

The anchoring vignettes method, designed for correcting subjective responses and thus avoiding the effects of a response style bias, is currently applied in WHO surveys (King et al. 2004). In the domain of health systems' responsiveness however, experience with vignettes is still scarce, with the exception of studies by Rice et al. (2008, 2009, 2010). Using WHO data and including all age categories, these authors analyzed the performance of health systems in nine culturally diverse countries including India, African countries, Malaysia, Mexico and one European country (Spain) (Rice et al. 2009). Compared with raw data, they found differences in the ranking of countries when the vignettes' methodology was applied. Very recently, they also analyzed 17 European countries (Rice et al. 2010). Results pointed to large country ranking differences after correction for response style.

In parallel to methodological work of vignette validation, there is a need to explore the application of this methodology for the purpose of comparing health system responsiveness in other databases produced by international surveys. In Europe, the anchoring vignettes methods was used to compare the level of subjective health in countries participating in the first wave of the Survey of Health, Ageing and Retirement in Europe—SHARE—taking variations in response style into account (Jurges 2007). The second wave of this survey integrated additional vignettes dedicated to a range of other dimensions, including expectations towards the healthcare system. SHARE is a survey focusing on the population aged 50 years and over. Owing to the high prevalence of chronic diseases in the second half of the life, this age category is particularly exposed to experiences with the healthcare system. As shown in the first wave of SHARE, only a small proportion had no medical contact in the year preceding the survey (Santos-Eggimann et al. 2005). Assessing the performance of health care systems in the age range covered by SHARE is thus important.

The aim of this exploratory study is to apply the vignette methodology in the domain of health systems' performance using data from the second wave of SHARE. We compare health care responsiveness, as assessed by individuals in the second half of their life, adjusting for a range of socio-demographic and health factors and describe the effect of correction for response style looking at data collected in 11 European countries.

2 Data

2.1 Sources

This study uses data from SHARE. This survey has been developed on the basis of prior successful socioeconomic surveys covering the older part of the population: the Health and Retirement Survey (HRS) in the United States, and the English Longitudinal Survey of Ageing (ELSA). This European bi-annual longitudinal survey aimed at carrying out international comparisons and analysis of economic and social problems related to ageing.[1] The COMPARE project collects survey data in a subset of all SHARE countries based on the 'anchoring vignettes' method (King et al. 2004) in order to create internationally comparable measures of several dimensions of the quality of life, of which health care responsiveness is one dimension. COMPARE is part of the family of research projects linked to SHARE. Data collection is in parallel to the SHARE data collection in waves 2004 and 2006–2007 and follows the same procedures. The sample covers respondents of age 50 and older and their spouses in 11 EU countries: Denmark, Sweden, Germany, Poland, Netherlands, Belgium, France, Czech Republic, Spain, Italy and Greece. Each participating country delivered a random sample of approximately 780 respondents (about 520 households) drawn from the SHARE main sample. The vignettes are administered along with the SHARE questionnaire. Notice that this study is based on release 2.3.0 of SHARE wave 2 data. Analyses were conducted on all subjects aged 50 year old and over in 2007 who participated in SHARE wave 2 and its vignettes supplement in eleven countries, including questions on health care responsiveness and corresponding vignettes.

2.2 Variables

Self-assessment questions and vignette evaluations on health care responsiveness were derived from three of the eight dimensions defined by the WHO for population health surveys: waiting time for medical treatment, quality of the conditions in visited health facilities, and communication and involvement in decisions about the treatment (cf. van Soest 2008). They were part of a self-administered drop-off questionnaire filled after completion of the SHARE main interview. For each dimension, subjects responded to one question evaluating their own experience[2] and then provided their evaluation of one vignette presenting the specific situation of a hypothetical individual. Self-assessments and vignette reports follow a five items scale from very good (1) to very bad (5) ('conditions of the health facilities' and 'communication about treatment') or from very short (1) to very long (5) for the evaluation of the 'time to wait for medical treatment'. Respondents aged 65 and over also evaluated a second vignette for each of the three dimensions of health care responsiveness.

Socio-demographic and health variables used as controls were extracted from the SHARE core data. They include the usual controls like country dummies and individual characteristics such as gender, age (in years), the level of education (none to primary,

[1] For further details, cf. Börsch-Supan and Jurges (2005) and www.share-project.org.

[2] For instance, "In many countries, it takes time before people can see a specialist and there are waiting lists for certain procedures. Overall, in your situation, how would you rate the amount of time you have to wait for medical treatment?" Or, "Overall, how would you rate the conditions of the health facilities you have visited?" and "Overall, how would you rate how clearly doctors and nurses communicate with you and involve you in decisions about the treatment?".

Table 1 Share of unsatisfied with health care responsiveness

	Time to wait for medical treatment	Conditions of the health facilities	Communication with doctors
Germany	0.196	0.040	0.076
Sweden	0.432	0.074	0.079
Netherlands	0.189	0.027	0.049
Spain	0.469	0.059	0.112
Italy	0.509	0.193	0.163
France	0.270	0.056	0.120
Denmark	0.465	0.040	0.065
Greece	0.489	0.379	0.380
Belgium	0.128	0.019	0.085
Czech rep	0.172	0.063	0.091
Poland	0.702	0.183	0.166
Total	0.360	0.096	0.118

Unsatisfied = reporting bad (long) or very bad (very long). N = 7,189 full rank data matrix. Calibrated individual weights used

secondary, tertiary), and the log of net annual income per capita. Some other more specific variables have been retained in the analysis in order to take into account the influence of health condition on health care satisfaction. They are (1) a dummy based on the Euro-d scale (Prince et al. 1999; Dewey and Prince 2005) which provides a standard measure of the symptoms of depression, and (2) a dummy indicating if the respondent has difficulties in Katz' basic activities of daily living (ADL) or Lawton's instrumental activities of daily living (IADL). We also retained a dummy variable taking the value 1 if the respondent has visited a hospital or a nursing home in the last 12 months.

2.3 Sample and Descriptive Statistics

The complete sample consists of 7,189 individual respondents. The sample size by country goes from 360 respondents in France to 1,108 in Germany. The proportion of men is 45.8% across countries and the median age is 63 years. More detailed descriptive statistics of the sample are displayed in Table 4 in the "Appendix".

A first glance at cross-country means of respondents reporting low levels of satisfaction (i.e. *long* or *very long* for 'waiting time'; and *bad* or *very bad* for the two other domains) indicates that senior Europeans of the COMPARE project are mostly satisfied with the time to wait for medical treatment (Table 1). However, based on national weighted averages, even if only 36% of them would rate waiting times as *long* or *very long*, some important differences between countries are noticeable; especially between Eastern (ex-communist) countries: 70.2% of Polish respondents are dissatisfied, whereas only 17.2% of Czech respondents complain. Other health care responsiveness domains provide less cross-country variation. By and large, seniors rate "conditions of the health facilities" and "communication with doctors" positively since only 9.6% of them report being dissatisfied with the former, and 11.8% for the latter. One common trend is that, Greece, Italy, and Poland, always appear to be among the most dissatisfied countries. To be reliable, these statistics in the raw data should not only control for other covariates, but also be "purged" from potential reporting heterogeneity between countries.

3 Evidence of Reporting Heterogeneity

3.1 Heterogeneity in Vignette Ratings

A simple glance at the distribution of vignettes ratings gives a good idea of potential reporting heterogeneity. If respondents evaluate in a different way the same hypothetical situation (vignette), this is evidence of response scale differences—also referred to as differential item functioning (DIF). Although each vignette describes a given situation for each domain of health care responsiveness, the vignettes ratings show considerable variation, which can be attributed to reporting heterogeneity. For instance, 27.4% of respondents to vignette 1 on "communication with doctors" consider the situation being *good*, while about the same amount of respondents (27.3%) consider the same situation as *bad* (Fig. 1).

Table 2 indicates cross country differences in the evaluation of the vignettes. For instance, the share of people who rated the time to wait for medical treatment in vignette 1 as long or very long goes from 22.1% for respondents in the Netherlands to 52.1% in Denmark. Cross-country differences in vignettes ratings are specific to each dimension, apart from the Czech Republic where respondents give low ratings (inferior to the sample mean) to all vignettes. Respondents from Sweden to Denmark report higher ratings of the two vignettes for "Communication with doctors" while Swedish respondents give low ratings and Danes report higher ratings of the two vignettes for "Time to wait for medical treatment". These differences and similarities in the way respondents perceive each dimension of health care responsiveness will have important consequences in terms of country ranking before and after correction for DIF. Special attention will be given to the Czech Republic, Sweden, and Denmark for the two dimensions mentioned above. In the

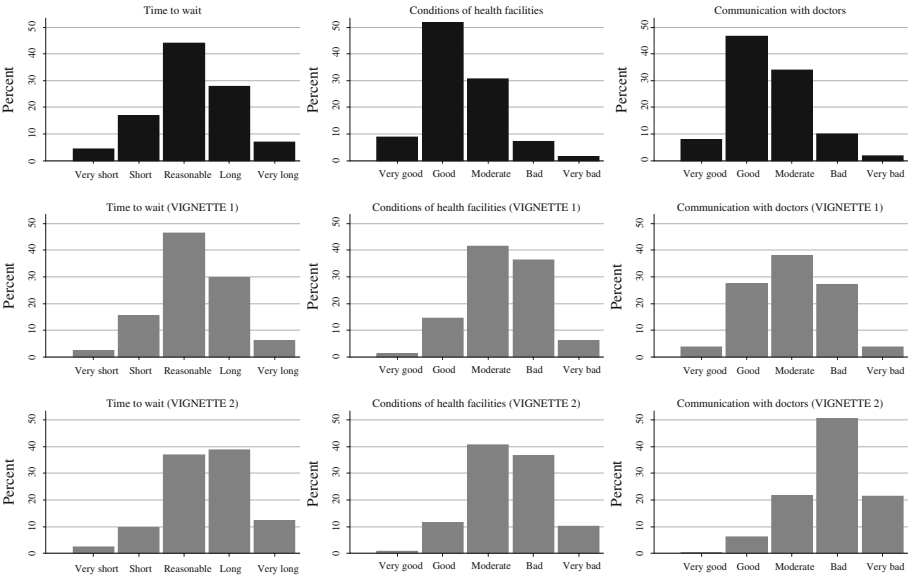

Fig. 1 Distribution of self-assessments and vignette evaluations of three dimensions of health care responsiveness

Table 2 Vignettes ratings—share of unsatisfied by country

Time to wait for medical treatment		Conditions of the health facilities		Communication with doctors		
Vignette 1	Vignette 2 (65 + only)	Vignette 1	Vignette 2 (65 + only)	Vignette 1		Vignette 2 (65 + only)
Germany	0.397	0.133	0.332	0.139	0.295	0.225
Sweden	0.306	0.148	0.519	0.170	0.448	0.377
Netherlands	0.221	0.191	0.272	0.134	0.277	0.282
Spain	0.395	0.191	0.732	0.202	0.315	0.190
Italy	0.300	0.268	0.375	0.182	0.301	0.346
France	0.272	0.241	0.479	0.227	0.191	0.290
Denmark	0.521	0.238	0.472	0.178	0.377	0.326
Greece	0.444	0.453	0.407	0.285	0.396	0.373
Belgium	0.267	0.316	0.583	0.267	0.324	0.374
Czech rep	0.331	0.088	0.262	0.170	0.239	0.269
Poland	0.388	0.266	0.318	0.162	0.277	0.257
Total	0.343	0.211	0.415	0.179	0.287	0.277

Unsatisfied = reporting bad (long) or very bad (very long). Calibrated individual weights used

case of "Conditions of the health facilities", respondents from France, Belgium and Spain report systematic higher ratings while Polish respondents provide low ratings of the two vignettes for this dimension. Focus on these countries may also be interesting.

3.2 Modelling Heterogeneous Reporting: The CHOPIT Model

In order to use the anchoring vignettes to correct self-assessments for response scale differences, we follow King et al. (2004: 192), who extended the standard ordered probit model (or Oprobit) to a joint compound hierarchical probit model (or Chopit). The main differences are that in the latter model (i) vignettes provide information about a common reference to self-assessed questions, and (ii) thresholds for responses to both self-assessed and vignettes questions may vary by country, with individual characteristics, with health conditions, etc. As King et al.(2004: 197) put it "[i] n broad outline, our model can be thought of as a generalization of the commonly used ordered probit model, where we model DIF via threshold variation, with the vignettes providing the key information." The Chopit model consists of an underlying self-assessment equation and an underlying vignette equation, which are additive in a systematic part and a normally distributed error term. These equations explain unobserved genuine satisfaction with responsiveness on a continuous scale. The observed discrete outcomes are obtained as in an ordered probit model, as the category containing the unobserved continuous outcome. The cut-off points defining the categories are constants in the ordered probit model, but vary with control variables such as country dummies and individual characteristics in the Chopit model. The cut-off points are assumed to be the same for self-assessments and vignette evaluations. See King et al. (2004) for details. The Chopit model thus consists of two components: a reporting behaviour part for the thresholds and a responsiveness equation. These are simultaneously estimated with maximum likelihood (e.g. Rabe-Hesketh et al. 2002), the same covariates are used to both determine self-assessment and thresholds and the

observed responses to health care self-assessment (so that $vi = xi$ in the notation of King et al. (2004). The analysis is carried out separately for the three domains of health care responsiveness presented in the above section. Chopit estimations are based on Jones et al. (2007), and were performed using Stata software (StataCorp., 2005).[3]

3.3 Model Estimates and Statistical Inference

Tables 5, 6 and 7 in the "Appendix" provide the estimates of the threshold equations. By and large, our results indicate the presence of systematic reporting behaviour variation that is linked to the individual variables included in the models. Since ratings for health care responsiveness go from 1 (very good/very short) to 5 (very bad/very long), negative coefficients across the thresholds indicate more critical evaluations of responsiveness of a given (vignette) situation, pointing at, for example, higher expectations of health services. In detail, in any of the three domains of health care responsiveness considered here, respondents with depression symptoms give lower ratings (i.e. they express less dissatisfaction for the same situation), while *ceteris paribus*, respondents who have been in a hospital during the last 12 months seem to give higher ratings (i.e. they express more dissatisfaction for the same situation). The age coefficients imply that older people appear to be more accommodating in the case of waiting times for medical treatment and communication with doctors.

Tables 5, 6 and 7 provide some interesting information on cross country differences in the thresholds. For example, dummies for Sweden and the Czech Republic have positive and significant coefficients in all the thresholds equations (mu1 to mu4) for "Time to wait for medical treatment", while Denmark is associated with negative coefficients for the thresholds equations (mu2 to mu4). In other words, for the same situation to be comparable across countries (without DIF; i.e. comparable with the way benchmark respondents (Germans) rate the given situation), ratings from the Swedish and Czech respondents need to be adjusted to a higher degree of dissatisfaction, while ratings from the Danes need to be adjusted to a lower degree of dissatisfaction. This result means that, once the influence of individual covariates have been controlled for, the Swedes and the Czechs in the sample are generally less dissatisfied than Germans with the same situation, while the Danes are on average more dissatisfied.

Different situations appear when it comes to the other dimensions of health care responsiveness considered here. Although once again, ratings from the Czech respondents need to be adjusted to a higher degree of dissatisfaction, French, Belgian, and Spanish respondents' ratings of "Conditions of the health facilities" need to be adjusted to a lower degree of dissatisfaction, in order to be comparable (with the German's way of rating). Threshold equations for "Communication with doctors" (especially mu3 and mu4) indicate that, to be comparable, average ratings for e.g. Sweden, Denmark, Greece, and Belgium need to be adjusted to a higher degree of dissatisfaction while Polish respondents' ratings need to be adjusted to a lower degree of dissatisfaction. These results are generally coherent with vignettes ratings displayed above in Table 2.

Table 3 displays Chopit estimates of health care responsiveness alongside estimates for the same covariates (in the self-assessment equation) obtained from a usual Probit. Notice that these latter coefficients are normalized to the estimate of sigma (the variance of the latent variable in the Chopit model) in order to make results comparable (the variance

[3] Program file (.do) available upon request. In the interests of brevity and to conserve space, standard errors are not reported here. More detailed tables are available from the authors upon request.

Table 3 Comparable Oprobit and Chopit estimates for self-assessment equations

	Time to wait for medical treatment		Conditions of the health facilities		Communication with doctors	
	OPROBIT	CHOPIT	OPROBIT	CHOPIT	OPROBIT	CHOPIT
Age (years)	−0.005***	−0.002	−0.006***	−0.006***	−0.007***	−0.001
Gender						
Woman	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Man	0.033	0.047	0.054**	0.109***	0.014	0.012
Education						
<Primary	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Secondary	−0.032	0.076*	0.018	−0.004	−0.003	−0.008
Tertiary	−0.133***	0.008	0.002	−0.031	−0.058*	−0.091*
Income						
Log(income + 1)	0.005	0.002	−0.013	−0.012	-0.015	−0.016
Health						
Euro-d	0.118***	0.079*	0.166***	0.099*	0.155***	0.096**
ADL/IADL	−0.001	−0.013	0.120***	0.201***	0.067**	0.103**
In hospital/12 m.	−0.012	0.045	−0.155***	−0.014	−0.090*	−0.011
Country						
Germany	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Sweden	0.645***	0.978***	0.369***	0.217**	−0.009	−0.447***
Netherlands	0.215***	0.474***	0.161**	0.181	−0.041	−0.311***
Spain	0.724***	0.660***	0.506***	−0.237**	0.247***	0.126
Italy	0.602***	0.706***	1.098***	1.075***	0.477***	0.444***
France	0.113*	0.158*	0.241***	−0.099	0.196***	0.391***
Denmark	0.713***	0.443***	0.189***	0.128*	−0.029	−0.026***
Greece	0.540***	0.324***	1.234***	1.462***	0.936***	0.807***
Belgium	−0.311**	−0.216***	−0.014	−0.390***	0.138***	−0.010
Czech Rep.	0.032	0.391***	0.595***	0.912***	0.147***	0.195***
Poland	1.095***	1.147***	0.987***	1.270***	0.390***	0.551***
Vignettes						
Dummy vig. 1		0.397**		1.282***		0.516***
Dummy vig. 2		0.756***		1.420***		1.620***
Sigma		1.024***		1.015***		0.926***
Log. L.	−9135.0	−21601.6	−7867.2	−19757.9	−8478.7	−21133.1
H0: "Reporting homogeneity"		709.0***		1213.8***		548.6***
H0: "Parallel cut-point shift"		358.3***		550.7***		277.5***

Rating for health care responsiveness goes from 1 (very good/very short) to 5 (very bad/very long). Ordered probit coefficients normalised for variance

* $P < 0.10$; ** $P < 0.05$; *** $P < 0.01$

would normally be normalised to one, but here it is set equal to the estimate from the Chopit model).

The choice between Oprobit and Chopit depends on the validity of the assumption of homogeneous reporting. We can test for reporting heterogeneity across all cut-points for all covariates by checking joint significance of all variables in all cut-point equations

(Tables 5, 6 and 7). The result of the likelihood ratio test of constant cut-points reported at the bottom of Table 3 indicates that the null hypothesis of homogeneity is strongly rejected for all three domains of health care responsiveness. In addition, the assumption that the covariates have the same effect on all cut-points (parallel cut-point shift) is also strongly rejected. In other words, the Oprobit estimates may be biased since they do not only reflect how responsiveness varies with the covariates but also are affected by differential item functioning.

Results from Chopit models are quite different from Oprobit ones. In the latter, age is significantly associated with better ratings in any of the three domains of health care, suggesting that people are more accommodating with the system as life expectancy is reducing. However, in the Chopit models, age remains significant only for 'conditions of the health facilities'. Health status is also an important determinant of self-assessments in the Oprobit models since respondents with depression symptoms (Euro-d) systematically report worse health care responsiveness in all three domains. Although this result stills hold in Chopit models, the coefficient is smaller (though still significant at <10%) for "time to wait for medical treatment" and "conditions of the health facilities." According to the Chopit estimates, being in a hospital does not affect the way people rate the situations—one interpretation could be that the vignettes help clarifying the concept of health care responsiveness for all respondents, irrespective of whether they have recent experiences with the health care system or not.

Nevertheless, some common trends between Chopit and Oprobit can be noticed. For instance, having difficulties in ADL or IADL has a significant impact on the way respondents rate the health care system. Having such limitations increases the probability to be dissatisfied (the coefficient being even more important in the Chopit model than in Oprobit). Gender issues in health care responsiveness only seem to impact how respondents rate "conditions of the health facilities." Men seem more often dissatisfied with this domain of health care than women. It is noticeable that income does not play any role in the way people rate health care situations. Since this result appears to be robust to different model specifications (various sets of socio-economic and health variables have been tested), it may be specific to the sub-population of older citizens.

Table 3 allows discussing the estimates of the coefficients on the country dummies and how they differ between Oprobit and Chopit. Noticeable changes in the case of "time to wait for treatment" concern the coefficient of the dummy for the Czech Republic that is not significant in the Oprobit model (meaning that the Czech Republic holds a similar position to Germany *ceteris paribus*, with regard to its share of unsatisfied respondents) and becomes positive and significant in the Chopit model (meaning that once adjustment for systematic country-level reporting behaviour has been undertaken, Czech respondents appear to be more dissatisfied with their time to wait for medical treatment than Germans). The coefficients for Sweden and the Netherlands is also higher (significant and positive) in the Chopit model for "Time to wait for treatment" while the coefficient for Denmark shrinks (though still significant) from 0.713 in the Oprobit model to 0.443 in the Chopit model. Changes in country dummies also affect the other dimensions of health care responsiveness. For instance, the equations for "Conditions of the health facilities" indicate that the coefficient for France goes from positive (Oprobit) to not significant (Chopit); for Belgium, from not significant (Oprobit) to negative (Chopit); for Spain, from positive (Oprobit) to negative (Chopit). In the case of "Communication with doctors", northern countries (Sweden, the Netherlands, and Denmark) have non-significant coefficient in the Oprobit models while a significant and negative coefficient in the Chopit models.

4 Correcting for DIF

Previous Chopit estimations make possible to illustrate cross-country differences using conditional distributions of health care responsiveness. Predictions form self-assessment equations would give interesting comparable results depending on how the corresponding threshold equations are used. Since the cut-points should be fixed in order to correct for reporting heterogeneity, two counterfactual distributions of self-assessed responses can be simulated from Chopit estimates, using, respectively:

1. The country's own parameters in self-assessment and threshold equations. This distribution of health care responsiveness thus also contains reporting heterogeneity.
2. The country's own parameters in the self-assessment equation, but using the benchmark country's parameters (Germany's) for the thresholds. This means that every respondent in each country is given the thresholds of a similar person in Germany. This distribution is the counterfactual "if everyone in the sample would report in the same way as the Germans." It thus gives people's self-assessment without cross-country reporting heterogeneity.

Notice that it is possible to look at the differences between countries (and rankings) before and after adjustment for DIF. Choosing Germany as the benchmark (since it has the largest sample of people responding to the vignettes) means the two distributions are the same for this country. For other countries, however, the difference between 1 and 2 shows the effect of DIF with reference to Germany (Kapteyn et al. 2007).

Figures 2a–c display cross-country comparisons of the share of dissatisfied respondents (i.e. those who would report the situation being *very bad* or *bad*—resp. *very long* or *long*) depending on the thresholds used. First, DIF correction in the case of "time to wait for treatment" confirms that France, Germany, and Belgium (old "Bismarck" systems) do seem to account for less unsatisfied respondents than other countries. Second, Northern countries—Denmark, the Netherlands, and Sweden—appear to be leading examples in "communication with doctors", once the statistics are purged from reporting heterogeneity. Thirdly, a divide between southern and eastern countries on the one hand and other countries on the other hand, is clear cut since Poland, The Czech Republic, Greece, and Italy have the highest rates of dissatisfaction with "conditions of the health facilities." One would have observed that Spain is also a country where seniors rate this domain of health care as poor, but correction for DIF clearly separates Spain from this group of countries.

Although leading countries in terms of health care responsiveness remain among the most successful after correction for reporting heterogeneity, many shifts happen in the ranking of the other countries. First, northern countries like Sweden and Denmark initially have a similar large share of their respondents unsatisfied with "Time to wait for treatment", unlike the Czech Republic where respondents initially seem less dissatisfied with this dimension. However, Fig. 2a shows that after correction for DIF, Denmark appear to me much closer to The Czech Republic than to Sweden. Second, the rates of unsatisfied respondents with "Conditions of the health facilities" in Fig. 2b initially rank Spain close to the Czech Republic and quite far above France (in terms of dissatisfaction with this dimension), while after correction for reporting heterogeneity, Spain and France have similar low rates of dissatisfaction, and the Czech Republic has even a larger share of dissatisfied respondents than previously. Third, Fig. 2c shows that Belgium initially seemed to hold a similar position to France with regard to the share of respondents

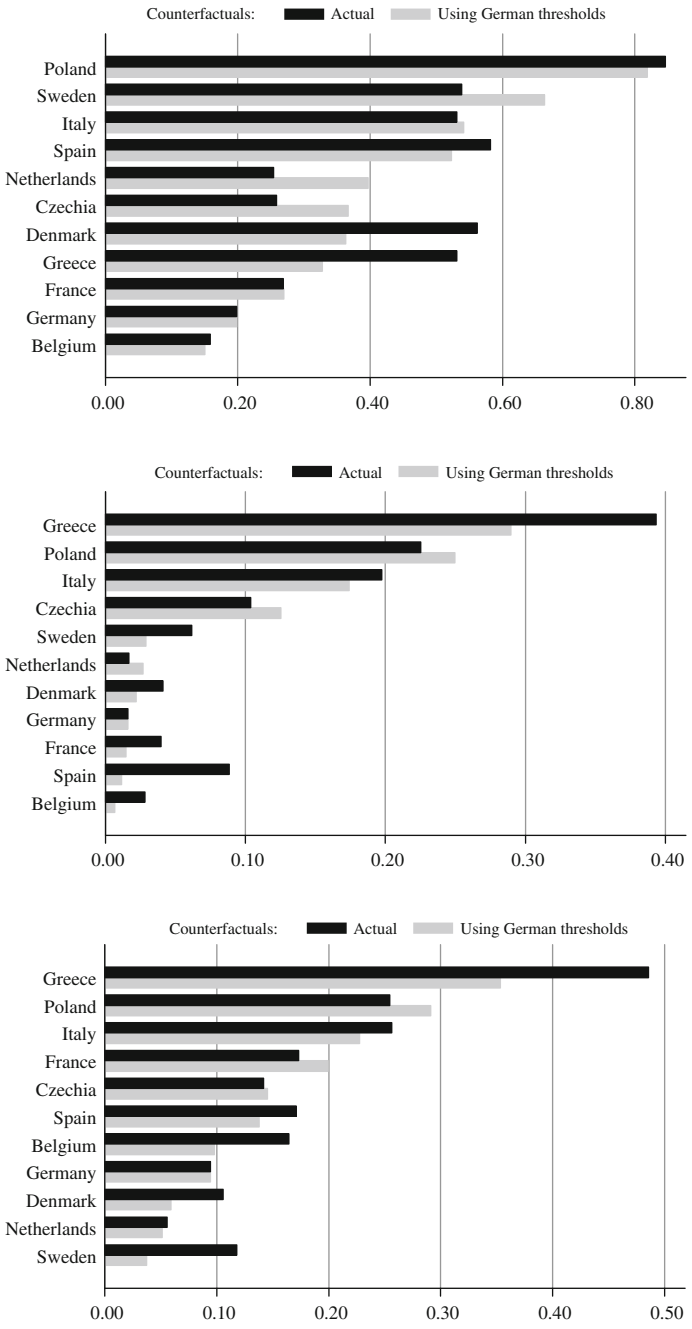

Fig. 2 Cross-country comparisons with DIF correction. **a** Share of unsatisfied with Time to wait, **b** Share of unsatisfied with conditions of health facilities and **c** Share of unsatisfied with communication with doctors

dissatisfied with "Communication with doctors" while after correction for DIF, Belgium has a lower rate of unsatisfied respondents that is then similar to Germany.

5 Conclusions and Discussion

This study used anchoring vignettes to produce internationally comparable data in three domains of individuals' self-assessment of health care responsiveness (the extent to which the process of care delivery matches patients' expectations). Respondents' self-assessments have been re-scaled in each domain—just like if they all had the same understanding of the questions and the same values, cultural beliefs, etc. Chopit models estimates suggest that reporting heterogeneity (1) is more prominent in some countries compared to others in Europe (Germany being the benchmark country), varies across health care domains and across individuals within each country, and (2) can be explained to some extent by both individual (socio-demographic and health variables) and national characteristics. The use of counterfactual distributions of individual's responses to health care responsiveness helped to investigate the genuine difference (i.e. without DIF) between countries. Although leading countries in terms of health care responsiveness remain among the most successful even after correction for reporting heterogeneity, one may acknowledge many shifts in the ranking of the other countries. Some of the usual North–South gradients in Europe are confirmed—for instance, Northern countries (Denmark, the Netherlands, and Sweden) appear to be leading examples in "communication with doctors"—while some important rescaling shed light on potential new patterns of thinking the relationship between health care responsiveness and country specific characteristics—e.g. Sweden and Denmark appear to be less similar when it comes to "Time to wait for treatment".

The empirical literature on health care responsiveness and reporting heterogeneity is still scarce. Nevertheless, recent work by Rice et al. (2009, 2010) provides some interesting findings on the vignettes method applied to health care responsiveness. On the one hand, their results mainly concur with our analysis that correction for DIF modifies the ranking of country performance once adjustment for systematic country-level reporting behaviour has been undertaken. On the other hand, their analysis suggests the differences in ranking before and after correction for DIF are larger. One reason of such divergence could be found in the specific sample we used in this analysis. Could SHARE respondents be seen as a somehow homogeneous population? A first explanation could be that respondents in this sample all have 50 years or over, meaning that a large proportion of them already experienced health care facilities in their life. Interactions between age and country dummies or counterfactual simulations for the respondents aged 50+ using the World Health Survey data (like Rice et al. 2009, 2010) would be a possible way test for this hypothesis.In a broader perspective, there is a need to produce a larger set of references on the application of the vignette methodology, especially using other databases produced by comparable international surveys.

Our results are based on a set of hypotheses that require to be discussed. Crucial to the implementation of the Chopit model are the assumptions of response consistency and vignette equivalence (see for instance Bago d'Uva and Van Doorslaer 2009). Following King et al. (2004), we could for example use objective measures to test the assumption of response consistency. In the context of our particular application, primary evidence suggests that European countries have different waiting times for medical treatment (Björnberg and Uhlir 2008). However, these values are drawn from patients' perspectives and are thus subject to DIF. Or et al. (2010) underline the lack comparability in objective measures

of waiting times since no European standards seem to exist. Nevertheless, Mojon-Azzi and Mojon (2007) provide some interesting comparisons using SHARE data (2004–2005) for waiting times for cataract surgery of 245 respondents in ten countries. They found that waiting times differed significantly between the ten analysed European countries. The waiting time is significantly influenced by the total expenditure on health but not by other country specific health indicators (such as the rate of public expenditure on health, the physician density or the acute bed density The literature on the vignettes application is still scarce and the idea that health care system features could influence response consistency points towards new directions in research. For instance, running respective Chopit models on the different types of health care systems (e.g. Beveridge vs. Bismark) could maybe help reduce response consistency. Additional research exploring the macro-level determinants of health care responsiveness by domain would be a useful step to bring together research issues and public health policies in Europe.

Acknowledgments This study is part of the COMPARE project. "This paper uses data from SHARE release 2.3.0, as of November 13th 2009. SHARE data collection in 2004-2007 was primarily funded by the European Commission through its 5th and 6th framework programmes (project numbers QLK6-CT-2001-00360; RII-CT-2006-062193; CIT5-CT-2005-028857). Additional funding by the US National Institute on Aging (grant numbers U01 AG09740-13S2; P01 AG005842; P01 AG08291; P30 AG12815; Y1-AG-4553-01; OGHA 04-064; R21 AG025169) as well as by various national sources is gratefully acknowledged (see http://www.share-project.org for a full list of funding institutions)." The authors would like to thank Arthur van Soest, Theresa Bago d'Uva, Silvana Robone, Hendrick Jürges, Renske Kok, and two anonymous referees for useful comments and suggestions on previous versions of this paper.

Appendix

See Tables 4, 5, 6 and 7 are given below.

Table 4 Sample Description

	Obs.	Age (years)	Man (%)	Mean income	Education			Health		
					<Secondary (%)	Secondary (%)	Tertiary (%)	Euro-D (%)	ADL/IADL (%)	Hospital (%)
Germany	1108	65.0	46.6	21448.5	18.5	52.6	28.9	16.8	13.1	17.7
Sweden	456	66.1	46.5	21861.0	51.5	15.4	33.1	15.6	16.9	11.4
Netherlands	485	62.0	48.5	26594.2	51.2	21.4	27.4	14.4	11.8	8.0
Spain	485	64.4	46.8	26430.3	79.6	7.8	12.6	29.9	17.1	12.6
Italy	662	64.9	46.7	25991.1	69.4	21.5	9.1	32.0	18.4	10.9
France	360	65.2	45.8	27216.5	47.3	29.4	23.3	29.4	19.7	14.7
Denmark	937	64.5	45.7	22604.8	22.4	38.8	38.8	14.6	12.2	12.3
Greece	502	64.9	47.8	25001.6	57.4	22.1	20.5	15.5	18.3	8.0
Belgium	819	65.3	46.8	29391.5	51.0	26.0	23.0	25.0	19.2	17.2
Czech rep.	865	64.7	41.0	10049.5	62.2	26.7	11.1	24.3	20.0	18.4
Poland	510	62.6	44.3	19265.4	43.4	38.6	18.0	53.1	34.5	17.1
Total	7189	64.6	45.8	27715.1	47.0	30.0	23.0	23.5	17.6	14.1

Table 5 Thresholds equations for 'Time to wait for medical treatment'

	mu 1	mu 2	mu 3	mu 4
Age (years)	−0.001	0.003**	0.004***	0.004*
Gender				
Woman	Ref.	Ref.	Ref.	Ref.
Man	−0.006	0.034	0.027	−0.037
Education				
<Primary	Ref.	Ref.	Ref.	Ref.
Secondary	0.040	0.029	0.166***	0.145***
Tertiary	0.063	0.075*	0.177***	0.240***
Income				
Log(income + 1)	0.028	−0.001	−0.007	−0.007
Health				
Euro-d	0.019	−0.023	−0.021	−0.111***
ADL/IADL	0.113*	0.010	−0.046	−0.047
In hospital/12 m.	0.116**	0.092**	0.052	−0.058
Country				
Germany	Ref.	Ref.	Ref.	Ref.
Sweden	0.619***	0.438***	0.219***	0.206***
Netherlands	0.050	0.097	0.356***	0.279***
Spain	0.140	−0.139**	−0.115***	−0.081
Italy	0.340***	0.133**	−0.047	0.175**
France	0.361***	−0.054	−0.006	0.035
Denmark	0.132	−0.193***	−0.426***	−0.305***
Greece	0.331***	−0.002	−0.417***	−0.401***
Belgium	0.298***	0.119**	−0.036	−0.011
Czech Rep.	0.577***	0.371***	0.256***	0.291***
Poland	0.267**	0.220***	−0.218***	0.190**
Constant	−2.106	−0.876***	0.525***	1.737***

Unlike King et al. (2004) who use exponential functions, the parameters in these tables relate to the thresholds through linear functions

* $P < 0.10$; ** $P < 0.05$; *** $P < 0.01$

Table 6 Thresholds equations for 'Conditions of the health facilities'

	mu 1	mu 2	mu 3	mu 4
Age (years)	0.004	−0.001	−0.001	0.000
Gender				
Woman				
Man	0.090*	0.046	0.035	0.034
Education				
<Primary				
Secondary	−0.096	−0.026	0.088***	0.106**
Tertiary	−0.086	−0.029	0.058*	0.037

Table 6 continued

	mu 1	mu 2	mu 3	mu 4
Income				
Log (income + 1)	−0.003	−0.001	−0.006	0.023*
Health				
Euro-d	−0.056	−0.075**	−0.076**	−0.081*
ADL/IADL	0.108*	0.104***	0.074**	0.010
In hospital/12 m.	0.230***	0.094**	0.050	0.103*
Country				
Germany				
Sweden	0.281**	−0.441***	−0.359***	−0.143
Netherlands	−0.266**	0.057	0.196***	0.336***
Spain	−0.397***	−0.956***	−0.928***	−0.691***
Italy	0.332***	−0.291***	−0.073	−0.122
France	−0.156	−0.507***	−0.426***	−0.393***
Denmark	0.156*	−0.213***	−0.281***	−0.086
Greece	1.307***	0.097	−0.262***	−0.167**
Belgium	−0.183*	−0.524***	−0.572***	−0.509***
Czech Rep.	0.722***	0.102*	0.099**	0.178**
Poland	0.640***	0.095	0.078	0.034
Constant	−1.892***	0.531***	1.780***	2.678***

Unlike King et al. (2004) who use exponential functions, the parameters in these tables relate to the thresholds through linear functions

* $P < 0.10$; ** $P < 0.05$; *** $P < 0.01$

Table 7 Thresholds equations for 'Communication with doctors'

	mu 1	mu 2	mu 3	mu 4
Age (years)	0.006***	0.007***	0.005***	0.012***
Gender				
Woman				
Man	−0.038	0.022	−0.013	−0.060*
Education				
<Primary				
Secondary	0.035	-0.038	0.026	0.001
Tertiary	−0.010	−0.041	−0.039	−0.100**
Income				
Log (income + 1)	−0.004	−0.005	0.001	0.025*
Health				
Euro-d	−0.016	−0.039	−0.106***	−0.202***
ADL/IADL	0.055	0.034	0.025	−0.008
In hospital/12 m.	0.173***	0.032	0.076**	0.125**
Country				
Germany				

Table 7 continued

	mu 1	mu 2	mu 3	mu 4
Sweden	−0.098	−0.599***	−0.571***	−0.526***
Netherlands	−0.479***	−0.245***	−0.046	0.033
Spain	0.179*	−0.276***	−0.129**	−0.123
Italy	0.219**	−0.138**	−0.066	−0.156**
France	0.472***	0.102	0.110*	−0.046
Denmark	−0.034	−0.251***	−0.295***	−0.276***
Greece	0.467***	−0.248***	−0.249***	−0.309***
Belgium	0.161**	−0.248***	−0.301***	−0.253***
Czech Rep.	0.358***	−0.106**	0.006	0.094
Poland	0.244**	0.118*	0.070	0.228**
Constant	−1.894***	−0.205	0.857***	1.449***

Unlike King et al. (2004) who use exponential functions, the parameters in these tables relate to the thresholds through linear functions

* $P < 0.10$; ** $P < 0.05$; *** $P < 0.01$

References

Bago d'Uva, T. & Van Doorslaer E. (2009). Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity. Health, Econometrics and Data Group (HEDG) Working Paper 09/30.

Björnberg, A., & Uhlir, M. (2008). *Euro Health Consumer Index 2008 Report*. Winnipeg: Health Consumer Powerhouse.

Börsch-Supan, A., & Jürges, H. (Eds.). (2005). *The survey of health, ageing, and retirement in Europe-met hodology*. Germany: Manheim Research Institute for the Economics of Ageing.

Coulter, A., & Jenkinson, C. (2005). European patients' views on the responsiveness of health systems and health care providers. *European Journal of Public Health, 15*, 355–360.

Dewey, M. E., & Prince, M. J. (2005). Mental health. In A. Boersch-Supan (Ed.), *Health, ageing and retirement in Europe: First results from SHARE* (pp. 108–117). Manheim: Research Institute for the Economics of Ageing.

Hausdorf, K., et al. (2008). Rating access to health care: are there differences according to geographical region? *Australian and New Zealand Journal of Public Health, 32*, 246–249.

Jones, A. M., Rice, N., Bago d'Uva, T., & Balia, S. (2007). *Applied Health Economics*. New York: Routledge.

Jürges, H. (2007). True health differences versuss response style: Exploring cross-country differences in self-reported health. *Health Economics, 16*, 163–178.

Kapteyn, A., Smith, J. P., & van Soest, A. (2007). Vignettes and self-reports of work disability in the United States and The Netherlands. *American Economic Review, 97*, 461–473.

King, G., Murray, C., Salomon, J., & Tandon, A. (2004). Enhancing the validity and crosscultural comparability of measurement in survey research. *American Political Science Review, 98*(1), 567–583.

Kristensen, N., & Johanson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics, 15*(1), 96–117.

Mojon-Azzi, S. M., & Mojon, D. S. (2007). Waiting times for cataract surgery in ten European countries: An analysis using data from the SHARE survey. *The British Journal of Ophthalmology, 91*(3), 282–286.

Murray, C. J. L., & Frenk, J. (2000). A framework for assessing the performance of health systems. *Bulletin of the World Health Organization, 78*, 717–731.

Or, Z., et al. (2010). Are health problems systemic? Politics of access and choice under Beveridge and Bismarck systems. *Health Economics, Policy and Law, 5*, 269–293.

Prince, M. J., et al. (1999). Development of the Euro-d scale–a European Union initiative to compare symptoms of depression in 14 European centres. *British Journal of Psychiatry, 174*, 330–338.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal, 2*, 1–21.

Rice N., Robone S. and Smith P.C. (2008). The measurement and comparison of health system responsiveness. University of York (G.B.), HEDG Working Paper 08/05, mimeo.

Rice N., Robone S. and Smith P.C. (2009).Vignettes and health systems responsiveness in cross-country comparative analyses. University of York (G.B.), HEDG Working Paper 09/29.

Rice, N., Robone, S., & Smith, P. C. (2010). International comparison of public sector performance: The use of anchoring vignettes to adjust self-reported data. *Evaluation, 16*(1), 81–101.

Santos-Eggimann, B., Junod, J., & Cornaz, S. (2005). Health services utilization in older Europeans. In A. Börsch-Supan (Ed.), *Health, ageing and retirement in Europe. First results from the survey of health, ageing and retirement in Europe* (pp. 133–140). Mannheim: Mannheim Resaerch Institute for the Economics of aging (MEA).

Schoen, C., et al. (2004). Primary care and health system performance: Adults' experiences in five countries. *Health Affairs, W4*, 487–503.

StataCorp. (2005). *Stata Statistical Software: Release 9.2*. College Station, TX: StataCorp LP.

van Soest, A. (2008). Enhancing international comparability using anchoring vignettes. In A. Börsch-Supan, et al. (Eds.), *First results from the survey of health, ageing and retirement in Europe (2004–2007)* (pp. 353–357). Germany: Mannheim Research Institute for the Economics of Aging (MEA).

World Health Organization (2000). *The World health Report 2000*. Health systems: Improving performance. Geneva: WHO.

World Health Organization (2001). Background paper for the technical consultation on responsiveness concepts and measurement. Geneva: WHO pp. 13–14 September 2001.