

Are different measures of self-rated health comparable? An assessment in five European countries

Hendrik Jürges · Mauricio Avendano ·
Johan P. Mackenbach

Received: 14 April 2008 / Accepted: 3 September 2008 / Published online: 24 September 2008
© The Author(s) 2008. This article is published with open access at Springerlink.com

Abstract *Objective:* Self-rated health (SRH) is widely used to compare population health across countries, but comparability is often hampered by the use of different versions of this item. This study compares the WHO recommended version (ranging from ‘very good’ to ‘very bad’) with the US version (ranging from ‘excellent’ to ‘poor’) in European countries. *Methods:* Data came from the Survey of Health, Ageing and Retirement in Europe (SHARE). Both the WHO and US versions of SRH were measured in representative samples of Europeans aged 50+ ($n = 11,643$) in five countries. Concordance between the two SRH versions and differences in their associations with demographics, chronic diseases, functioning and depression were assessed using ordered probit regression. *Results:* The US version has a more symmetric distribution and larger variance than the WHO version. Although the WHO version discriminates better at the positive end, the US version shows better discrimination at the positive end of the scale. Sixty-nine percent of respondents provided literally concordant answers, while only about one-third provided relatively concordant answers. Overall, however, less than 10% of respondents were discordant in either sense. The two versions were strongly correlated (polychoric correlation = 0.88), had similar associations with demographics and health indicators, and showed a similar pattern of international variation. *Conclusion:* Health levels

based on different measurements of SRH are not directly comparable and require rescaling of items. However, both versions represent parallel assessments of the same latent health variable. We did not find evidence that the WHO version is preferable to the US version as standard measure of SRH in European countries.

Keywords Self-rated health · World health · International comparisons · Research design · Europe

Introduction

Self-rated health is an independent predictor of mortality [1–8], and it is the most widely used comprehensive health measurement [9] recommended by the World Health Organization (WHO) [5, 10]. Although differences have been observed between countries in self-rated health levels [10, 11], measurements vary in wording and scale across surveys [7, 12]. It is not known whether self-rated health variations across countries are due to true health differences or to the use of different measurements of self-rated health.

Two-five-point scale versions of self-rated health have been used in international surveys: The first one comprises answer categories ranging from ‘very good’ to ‘very poor’, and has been recommended by WHO-Europe and the European community health monitoring programme [7, 8, 13]. The second version ranges from ‘excellent’ to ‘poor’ and has been primarily applied in the US. It is not known whether both versions are directly comparable, which hampers international comparisons across surveys that use different measurements [7]. As opposed to the US version, the WHO version has been hypothesised to comprise a balanced set of two positive categories (*very good, good*), one neutral

H. Jürges
Mannheim Research Institute for the Economics of Aging
(MEA), University of Mannheim, Mannheim, Germany

M. Avendano (✉) · J. P. Mackenbach
Department of Public Health, Erasmus MC, University Medical
Center Rotterdam, PO Box 2040, 3000 CA Rotterdam,
The Netherlands
e-mail: m.avendanopabon@erasmusmc.nl

category (*fair*), and two negative categories (*bad*, *very bad*) [7]. However, no studies have empirically examined these advantages of the WHO version, and the scientific evidence for recommending this version remains scarce.

This study compares the WHO and the US versions of self-rated health across five different European countries. We applied both measurements in a sample of over 11,000 respondents to the Survey of Health, Ageing and Retirement in Europe (SHARE) in five European countries. To our knowledge, this is the first study to assess differences in the distribution of different versions of self-rated health, and in their association with demographic and health variables across countries.

Methods

Study population and data collection

SHARE

Details on the SHARE study in Europe have been described elsewhere [14, 15]. Briefly, in 2004, a survey was

conducted in representative samples of the non-institutionalised population aged 50 + in Sweden, Denmark, Germany, the Netherlands, France, Switzerland, Austria, Italy, Spain and Greece ($n = 22,777$). Interviews were face-to-face and took place in the household. Trained interviewers conducted interviews using a computer assisted personal interviewing program. The set-up allowed each country to use exactly the same underlying structure and questionnaire [14, 15].

The present analysis is based on data for Austria, Germany, the Netherlands, Spain and Greece, because overlapping answer categories for the two versions of self-rated health were verbally identical in these countries (Table 1). Other countries were excluded because translated answers were not verbally equivalent for both versions. All participants were drawn from probability samples of the underlying 50+ population in each country. Samples for Germany and the Netherlands were drawn from regional registries; the sample for Spain was drawn from a national population registry; and samples for Austria and Greece were drawn using telephone directories as sampling frames and pre-screening in the field of eligible sample participants [14, 15]. Household response rate was

Table 1 Original language answer categories for self-rated health using the European (WHO) and the US versions in five European countries: The SHARE study

Language	Countries	Self-rated health (WHO)	Self-rated health (US)
Generic (English) version		1 Very good	1 Excellent
		2 Good	2 Very good
		3 Fair	3 Good
		4 Bad	4 Fair
		5 Very bad	5 Poor
German	Austria, Germany	1 Sehr gut	1 Ausgezeichnet
		2 Gut	2 Sehr gut
		3 Mittelmäßig	3 Gut
		4 Schlecht	4 Mittelmäßig
		5 Sehr schlecht	5 Schlecht
Spanish	Spain	1 Muy Buena	1 Excelente
		2 Buena	2 Muy buena
		3 Pasable	3 Buena
		4 Mala	4 Pasable
		5 Muy mala	5 Mala
Greek	Greece	1. Πολύ καλή	1. Αριστη
		2. Καλή	2. Πολύ καλή
		3. Μέτρια	3. Καλή
		4. Κακή	4. Μέτρια
		5. Πολύ κακή	5. Κακή
Dutch	Netherlands, Belgium	1 Heel goed	1 Uitstekend
		2 Goed	2 Heel goed
		3 Redelijk	3 Goed
		4 Slecht	4 Redelijk
		5 Heel slecht	5 Slecht

55.6% in Austria ($n = 1,893$); 63.4% in Germany ($n = 2,866$); 61.6% in the Netherlands ($n = 2,731$); 53.0% in Spain ($n = 2,252$); and 63.1% in Greece ($n = 1,901$) [14]. Response rates are comparable to those observed in other European surveys [14]. The total final sample for analysis comprised 11,643 participants.

Calibrated sampling weights were designed to adjust for the complex sampling design and non-response in each country [14]. However, due to the fact that the present study does not compare population parameters, we did not apply sampling weights. Because we examine intra-individual consistency of responses to both version of self-rated health, applying weights would not alter our results.

Self-rated health

Individuals were asked to rate their health separately using the WHO version (very good, good, fair, bad, or very bad) and the US version (excellent, very good, good, fair, or poor) of self-rated health. Half of the sample was randomised to receive one of the two versions at the beginning or at the end of the health survey. Table 1 summarises the original categories used in each country.

Demographic and health covariates

The following variables were assessed: (1) *Age* and *sex*; (2) *Highest level of education*, reclassified into three levels using the UNESCO International classification of education (ISCED-97) [16]: “low” (ISCED 0–2), “medium” (ISCED 3,4), and “high” (ISCED 5,6). (3) *Chronic diseases* ever diagnosed by a doctor, including heart disease, stroke, hypertension, hypercholesterolemia, diabetes, lung disease, asthma, arthritis, osteoporosis, cancer, ulcer, Parkinson disease, cataracts, and hip fracture. Information on these diagnoses was based on self-reported information only. Individuals’ answers were summarised in three categories: no condition, one or two conditions, and three or more conditions. (4) *Symptoms* as measured by self-report of back or joint pain, angina or chest pain, breathlessness, persistent cough, swollen legs, sleeping problems, fall and fear of falling, dizziness, stomach or intestine problems, and incontinence. Answers were summarised in three categories: no symptom, one or two symptoms, and three or more symptoms. (5) Limitations with ADL (activities of daily living), measured by a validated scale of limitations individuals have with basic activities, namely dressing, walking, bathing, eating, getting in and out of bed, and using the toilet [17]. (6) Limitations with IADL (instrumental activities of daily living), measured by a validated scale of limitations with the following activities:

using a map, cooking, shopping, telephoning, taking medications, working in the house, and managing money. Limitations with ADLs and IADLs were summarised in three categories: no limitation, one or two limitations, and three or more limitations. (7) *Depression* as measured by the Euro-Depression (Euro-D), a scale of depression symptoms validated for the European population. A EURO-D score higher than three is indicative of a depressive symptomatology and was used to dichotomise this variable [18].

Methods of analysis

We assessed the distribution and cross-tabulations of self-rated health ratings, and examined concordance between the WHO and US measurements in three ways:

- (1) *Concordance measures*. *Literal concordance* occurs when an individual’s response to both versions is verbally consistent regardless of the self-rated health version (e.g., respondent answers “very good” to both the US and WHO version). Combinations of either the two highest positive or the two highest negative ratings possible in both scales were also classified as concordant. *Relative concordance* occurs when an individual’s responses to both versions are consistent in terms of their position in the self-rated health scale. This assumes that individuals use the scale midpoint as an anchor or population average [19].
- (2) *Polychoric correlations* were calculated by maximum likelihood [20] using R 2.7.0, and assuming that general health is a normally distributed continuous latent variable divided into ordered levels [20]. A correlation close to one indicates that both scales measure the same concept. We used both Chi-squared tests and root mean square errors of approximation (RMSEA) to test the assumption of normality of latent health [21].
- (3) *Ordered probit regressions* [22, 23] were used to assess whether the associations of self-rated health with demographic and health variables differed for the WHO and US versions. The latent continuous variable ‘general health’ is modelled as a linear function of covariates. Coefficients summarise the effect of a one-unit increase in the explanatory variables on the continuous (latent) outcome variable. Country effects were measured by effect coding (effects are measured relative to the grand mean). Cross-equation tests (based on a seemingly unrelated estimation of the two ordered probit equations) were used to assess whether effect sizes differ significantly between the two versions. Analyses were conducted using Stata 9.2.

Results

Differences in distributions

Table 2 shows the distribution of answers to both self-rated health items. The WHO version had a more skewed distribution than the US version. In the WHO version, only 1.6% of participants rated their health as “very poor” (the bottom category), whereas more than 15% selected the top category “very good”. In contrast, about the same proportion of individuals selected the top and bottom categories in the US version.

Individuals appear to be in better health when confronted with the US version. Whereas 27.3% reported to be in very good or excellent health in response to the latter, only 15.5% reported ‘very good health’ (the top category) in response to the WHO version (Table 2). Similarly, whereas about 7% of respondents reported that their health was poor when presented with the US version, about 9.7% reported their health was poor or very poor when presented with the WHO version. Thus, the same verbal presentations elicited different assessments in the WHO and US versions.

Cross-tabulations in Table 3 show higher levels of literal than relative concordance. For instance, among those who reported that they were in good health when confronted with the WHO version, only 24.9% reported to be in very good health (relative concordance), whereas 65.7% reported to be in good health (literal concordance) when presented with the US version. Nevertheless, only about 10% of these participants reported that they were in excellent, fair or poor health, which were discordant ratings. The total percentage of concordant ratings is shown in Table 4. Percentages add up to more than 100%, because cases at the scale endpoints can be concordant both relatively and literally. Overall, 69.0% of participants provided literally concordant answers, whereas only 30.1% provided relatively concordant answers. Responses were discordant for only 8.1% of participants.

Table 3 Cross-tabulation of SRH (Self-rated health) between the WHO and US versions (row percentages) among men and women aged 50 years and over in five European countries: The SHARE study

SRH-WHO	SRH-US		Good	Fair	Poor	Total (col. %)
	Excellent	Very good				
Very good	37.9	51.3	10.5	0.3	0.0	15.5
Good	3.6	<i>24.9</i>	65.7	5.8	0.1	45.0
Fair	0.2	2.3	27.8	66.6	3.1	29.8
Poor	0.0	0.0	4.6	<i>41.4</i>	54.1	8.1
Very poor	0.5	0.5	0.0	11.1	87.8	1.6
Total (row %)	7.5	19.8	39.8	26.0	6.8	100.0

Notes: Numbers in italics indicate relative concordance; Numbers in boldface indicate literal concordance; Numbers in bold italics indicate both relative and literal concordance

Cross-country differences in concordance and discordance rates are statistically significant as suggested by the chi-squared test statistic. This results holds also if all covariates discussed in the next section are held constant. The overall polychoric correlation between the two versions was 0.882 (Table 4). Correlations were highest in Germany, the Netherlands and Greece, and lowest in Spain. Although Chi-squared tests reject the assumption of normality latent health, root mean square errors of approximation (RMSEA) indicate a good to acceptable fit, overall and in each country separately.

Differences in associations with covariates

The distribution of covariates varied significantly across countries (Table 5). For instance, Spaniards and Greeks had relatively low levels of education as compared to Germans and Austrians. More than two-thirds reported one or more diagnosed conditions or symptoms, and 20% reported at least one limitation with ADL. The prevalence

Table 2 Marginal distributions of self-rated health using the US and WHO versions among men and women aged 50 years and over in five European countries: The SHARE study

	Austria		Germany		Netherlands		Spain		Greece		Total	
	WHO	US	WHO	US	WHO	US	WHO	US	WHO	US	WHO	US
Excellent	N.A.	9.4	N.A.	4.7	N.A.	12.7	N.A.	3.5	N.A.	7.2	N.A.	7.5
Very good	17.8	24.7	11.3	17.2	18.3	18.1	9.9	15.1	21.7	27.0	15.5	19.8
Good	44.0	37.2	44.9	41.2	51.5	43.3	40.8	39.0	42.0	36.4	45.0	39.8
Fair	29.1	22.5	32.0	29.1	24.7	22.0	34.0	31.6	29.5	24.2	29.8	26.0
Poor	7.3	6.2	10.1	7.9	4.8	3.9	12.3	10.8	5.9	5.2	8.1	6.8
Very poor	1.8	N.A.	1.8	N.A.	0.6	N.A.	3.0	N.A.	0.9	N.A.	1.6	N.A.
N	1,893		2,866		2,731		2,252		1,901		11,643	

Note: N.A. indicates not applicable

Table 4 Degree of concordance between the WHO and US version of the self-rated health items among men and women aged 50 years and over in five European countries: The SHARE study

Country	% Literally concordant	% Relatively concordant	% Discordant	Polychoric correlation		
				Rho	Chi-squared (df = 15) ^b	RMSEA ^c
Austria	64.7	36.9	7.4	0.872	119.6	0.061
Germany	70.4	28.1	6.7	0.896	120.5	0.049
Netherlands	71.6	29.3	9.1	0.891	110.1	0.048
Spain	67.3	27.1	10.5	0.848	154.6	0.064
Greece	69.6	31.2	6.8	0.894	73.0	0.045
Total	69.0	30.1	8.1	0.882	558.0	0.056
Chi-squared (df = 4) ^a	31.4	59.7	34.7			

Notes:

^a Test of cross-country differences in the proportions of literally concordant, relatively concordant and discordant answers, respectively

^b Test of null hypothesis that latent variables follow a normal distribution

^c Root mean square error of approximation

Table 5 Description of health covariates (percentages) among men and women aged 50 years and over in five European countries: The SHARE study

	Austria	Germany	Netherlands	Spain	Greece	Total	Chi ² (df) ^a
Age 50–59	31.0	34.3	41.6	31.7	38.6	35.7	
Age 60–69	38.9	38.7	31.6	30.6	29.3	34.0	
Age 70–79	21.2	20.4	19.1	26.5	22.6	21.8	183.9 (12)
Age 80+	8.9	6.6	7.6	11.2	9.5	8.6	<i>P</i> < 0.001
Male	42.2	46.9	46.6	42.1	45.6	44.9	21.2 (4)
Female	57.8	53.1	53.4	57.9	54.4	55.1	<i>P</i> < 0.001
Low education	31.5	17.7	57.4	85.2	63.6	49.8	
Medium education	48.8	56.8	23.1	7.6	22.3	32.4	2,899 (8)
High education	19.7	25.5	19.5	7.1	14.2	17.8	<i>P</i> < 0.001
No diagnosed condition	30.9	27.3	32.1	20.5	27.2	27.7	
One or two conditions	54.1	52.8	52.8	51.2	54.6	53.0	211.7 (8)
Three or more conditions	15.0	20.0	15.0	28.3	18.2	19.3	<i>P</i> < 0.001
No symptom	32.3	29.4	38.7	27.4	36.1	32.8	
One or two symptoms	50.8	50.6	48.2	42.7	46.2	47.8	281.5 (8)
Three or more symptoms	16.9	20.0	13.1	30.0	17.7	19.4	<i>P</i> < 0.001
No (I)ADL limitation	79.1	84.5	83.5	73.9	80.7	80.7	
One or two (I)ADL limitations	14.1	11.0	12.3	17.5	15.1	13.7	132.8 (8)
Three or more (I)ADL limitations	6.8	4.5	4.2	8.7	4.2	5.5	<i>P</i> < 0.001
Depression score 0–3	80.3	81.6	81.0	63.9	75.1	76.8	292.7 (4)
Depression score 4 or higher	19.7	18.4	19.0	36.1	24.9	23.2	<i>P</i> < 0.001

Note: ^a Chi-Squared test of country differences in the distribution of covariates

of chronic diseases, symptoms and limitations was highest in Spain and lowest in the Netherlands. The prevalence of depression is much higher in Southern countries than in Austria, Germany, or the Netherlands.

Table 6 shows ordered probit regression models, which summarise the effect of a one-unit increase in the explanatory variables on latent general health. With the exception of three country effects, all variables were significantly

associated with both versions of self-rated health. Standard errors were marginally smaller for the US version, which reflects its more even distribution as compared to the WHO version. As assessed by cross-equation tests, the associations of self-rated health with most demographic and health variables were statistically indistinguishable for both the US and WHO versions (Table 6), with two exceptions: Firstly, the effect of being older than 80 years old on self-

rated health was significantly larger for the WHO version, which probably reflected its better discriminative power at the negative end of the scale. Secondly, chronic diseases were more strongly associated with the WHO than with the US version, although this difference was small. Overall, the WHO and US versions were very similarly associated with other variables.

Using different versions of self-rated health did not influence the ranking of countries in terms of their self-rated health. For both versions, self-rated health conditioning on covariates was best in Greece and the Netherlands, and worst in Germany (Table 6). The only exceptions were Austria and Spain, where ranks changed depending on the self-rated health version used. For other countries, self-rated health rankings were identical for the two items.

Cross-equation tests of parameter differences for the two versions of self-rated health were also computed separately by country (results not shown). In Germany, the Netherlands and Austria, there were no significant differences between the WHO and US versions in their associations with any of the covariates. In Spain, we found differences only for the number of conditions. In Greece, associations with age and education were different between versions, but associations with other variables did not differ.

Discussion

Although WHO has recommended the WHO version as the standard measurement of self-rated health in the European context [7, 8], our results suggest that this version is not

Table 6 Ordered probit regressions (fully adjusted models) of self-rated health for the WHO and US item versions and cross-equation tests ($N = 11,622$) among men and women aged 50 years and over in five European countries: The SHARE study

Covariate	Self-rated health (WHO)		Self-rated health (US)		Cross-equation test Chi-squared (df)
	Coeff.	SE	Coeff.	SE	
Age 50–59	0.000		0.000		
Age 60–69	0.077**	0.026	0.083**	0.025	
Age 70–79	0.236**	0.029	0.209**	0.029	12.32 (3)
Age 80+	0.254**	0.044	0.145**	0.044	$P = 0.006$
Male	0.000		0.000		0.18 (1)
Female	–0.131**	0.022	–0.123**	0.021	$P = 0.670$
Low education	0.000		0.000		
Medium education	–0.185**	0.028	–0.228**	0.027	4.96 (2)
High education	–0.371**	0.032	–0.418**	0.031	$P = 0.084$
No chronic conditions	0.000		0.000		
One or two chronic conditions	0.758**	0.027	0.691**	0.026	10.86 (2)
Three or more chronic conditions	1.160**	0.037	1.133**	0.037	$P = 0.004$
No symptoms	0.000		0.000		
One or two symptoms	0.457**	0.025	0.445**	0.024	0.48 (2)
Three or more symptoms	0.855**	0.038	0.837**	0.038	$P = 0.787$
No (I)ADL problems	0.000		0.000		
One or two (I)ADL problems	0.440**	0.033	0.445**	0.033	2.24 (2)
Three or more (I)ADL problems	0.952**	0.055	1.017**	0.061	$P = 0.326$
Depression score 0–3	0.000		0.000		0.52 (1)
Depression score 4 or higher	0.458**	0.029	0.442**	0.028	$P = 0.472$
Austria	0.039	0.024	–0.069**	0.024	
Germany	0.293**	0.021	0.312**	0.020	
Netherlands	–0.125**	0.020	–0.107**	0.020	
Spain	0.009	0.023	0.019	0.022	35.61 (4)
Greece	–0.215**	0.023	–0.155**	0.021	$P < 0.001$
Threshold 1	–0.247**	0.033	–0.834**	0.034	
Threshold 2	1.486**	0.036	0.218**	0.032	
Threshold 3	2.956**	0.042	1.653**	0.034	
Threshold 4	4.116**	0.056	3.140**	0.042	

Notes: Larger values = worse health. Source: SHARE 2004, release 1: Austria, Germany, Greece, Netherlands, Spain; * $P < 0.05$; ** $P < 0.01$

clearly superior to the US version. The WHO version discriminates better at the negative end, but the US version is more symmetric and shows better discrimination at the positive end. Individual answers to both items are not fully consistent, and appear to be more concordant in a literal rather than a relative sense. Despite these discrepancies, less than 10% of respondents were discordant in either sense. The US and WHO versions are highly correlated. They show very similar associations with demographic and health indicators, and they show a similar pattern of variation across countries. Overall, although the two measures are not directly comparable, they are in fact different categorizations of latent continuous health.

The strength of this study is the measurement of two self-rated health versions and covariates in several countries. However, some limitations should be considered. Data were only available for individuals aged 50 years and over. As younger individuals are on average healthier, measuring self-rated health in a younger cohort would result in a larger proportion of individuals reporting good health. In younger populations, the US version might be more appropriate because it discriminates better at the positive end. In addition, respondents were presented with both versions of self-rated health along with other health status measurements. The order of presentation (at the beginning or end) may have had an impact on the health ratings [24]. However, we tackled this problem by randomising the order of presentation of both versions, and analyses not shown in this paper indicate that presentation order had little impact on individual's levels of self-rated health.

Comparison with previous studies

The predictive power of subjective global health assessments has been shown in numerous studies [1, 2, 5, 25]. To our knowledge, this is the first study to show that the two most commonly used versions of subjective global health are not directly comparable within and across countries, but relate similarly to other covariates. Consistent with findings from single populations [12], we found that different measures of self-rated health are strongly correlated. Our results confirm findings from previous research suggesting that different measures of self-rated health represent parallel assessments of subjective health [12].

Differences between countries in the level of self-rated health and the association of this variable with socioeconomic and health factors have been reported [10, 11, 26–31]. Our results suggest that even if self-rated health is assessed in all countries using a 5-point scale, bias may yet be present due to differences in the wording of response categories. Thus, cross-country comparisons of population health based on different versions of the self-rated health

item may lead to spurious health variations across populations. On the other hand, the associations of self-rated health with demographic factors such as socioeconomic status were similar for the two self-rated health item versions. Thus, comparisons of how demographic and other factors relate to self-rated health across surveys using a different 5-point self-rated health scale [2, 10, 28–30, 32] are unlikely to be biased.

Interpretation and implications

Most health and social surveys contain only one version of the self-rated health item. This raises the question of whether it is possible to combine data from different surveys that use different versions of this item. Two-thirds of respondents in our study gave literally concordant answers. Thus, one option would be to collapse the two top categories of the US version and the two bottom categories of the WHO version, resulting in a four-point comparable scale. However, although this would minimise differences, this approach would still result in an overestimation of average health in surveys that use the US version. A second alternative is to achieve comparability of different versions of self-rated health by appropriately rescaling items. For instance, two surveys using different self-rated health measures but similar measures for other variables can be made comparable by imputing conditional probabilities obtained from surveys such as SHARE. Consider again the conditional probabilities shown in Table 3. In order to 'convert' the WHO into the US version, a random number, say X , could be drawn for each respondent from a uniform distribution on the zero-to-one interval. A respondent who has answered 'very good' to the WHO version would then be coded as being in 'excellent' health if $X < 0.379$ (thus with a 37.9% probability), as being in 'very good' health if $0.379 \leq X < 0.379 + 0.513 = 0.892$ (thus with a 51.3% probability), as being in 'good' health if $0.892 \leq X < 0.997$, and as being in 'fair' health if $0.997 \leq X \leq 1.000$. A respondent who has answered 'good' to the WHO version would be coded as being in 'excellent' health if $X < 0.036$, and so on. This procedure preserves the marginal distribution of the US version. It could also be repeated several times, yielding multiple imputations [33].

An important finding of this study is that respondents tend to be more concordant in a literal than in a relative sense. This finding might appear to contradict the view that individuals conceive the scale midpoint as the population average health when judging their own health status, independently of the verbal representation [19]. In fact, since two-thirds of our sample selected the equivalent verbal representation in both items, it would seem that respondents try to be consistent in a literal sense, regardless of the relative position of the answer categories. The main

implication is that using a 5-point scale is not enough to ensure comparability, because individuals react differently to various verbal representations when judging their health. As a consequence, comparisons between studies using different verbal answer categories are likely to be biased.

Although levels of self-reported health based on the US and WHO versions are not directly comparable, they are in fact different categorizations of the same latent continuous variable. In particular, both scales have the same properties with respect to demographics and health indicators. Thus, data from surveys using different self-rated health versions could still be used to compare associations of covariates with general health, even though overall health levels cannot be compared. However, this may require the use of appropriate statistical models that interpret self-rated health as different categorisations of an underlying (latent) continuous health variable.

WHO recommends the use of the WHO version as standard measurement of self-rated health in European populations. In our data, we found very little support for this directive. One of the central arguments of the WHO and related reports is that the WHO version comprises a balanced scale of five categories, two of which are positive (*very good, good*), one neutral (*fair*), and two negative (*bad, very bad*) [7, 8]. In our study, however, this balanced set of categories resulted in a skewed distribution of self-rated health. In terms of statistical efficiency, the US version has in fact some advantages. Responses to the US version are more evenly distributed across the 5-point scale, resulting in smaller standard errors of the estimated ordered probit parameters. The fact that both versions are similarly associated with demographic and health determinants further weakens the case for recommending the WHO version. Thus, in studies of older European populations, there does not seem to be a strong argument for preferring the WHO version. Moreover, the choice of a self-rated health version should be based on several considerations, including aspects such as the age distribution of the population studied, because in older populations, the WHO version tends to show a skewed distribution. These results invite a reassessment of WHO recommendations.

Acknowledgements This paper uses data from release 1 of SHARE 2004. The SHARE data collection has been primarily funded by the European Commission through the 5th framework program (project QLK6-CT-2001-00360 in the thematic program Quality of Life). Additional funding came from the US National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01 and OGHA 04-064). Data collection in Austria (through the Austrian Science Fund, FWF), Belgium (through the Belgian Science Policy Office) and Switzerland (through BBW/OFES/UFES) was nationally funded. The SHARE data set is introduced in Börsch-Supan et al. [15]; methodological details are contained in Börsch-Supan and Jürges [14]. Mauricio Avendano was supported by a grant from the Netherlands Organization for Scientific

Research (NWO, grant no. 451-07-001) and a Fellowship from the Erasmus University.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Burström B, Fredlund P. Self rated health: is it as good a predictor of subsequent mortality among adults in lower as well as in higher social classes? *J Epidemiol Community Health*. 2001;55(11):836–40.
- Appels A, Bosma H, Grabauskas V, Gostautas A, Sturmans F. Self-rated health and mortality in a Lithuanian and a Dutch population. *Soc Sci Med*. 1996;42(5):681–9. doi:10.1016/0277-9536(95)00195-6.
- Miilunpalo S, Vuori I, Oja P, Pasanen M, Urponen H. Self-rated health status as a health measure: the predictive value of self-reported health status on the use of physician services and on mortality in the working-age population. *J Clin Epidemiol*. 1997;50(5):517–28. doi:10.1016/S0895-4356(97)00045-0.
- Murata C, Kondo T, Tamakoshi K, Yatsuya H, Toyoshima H. Determinants of self-rated health: could health status explain the association between self-rated health and mortality? *Arch Gerontol Geriatr*. 2006;43:369–80.
- DeSalvo KB, Bloser N, Reynolds K, He J, Muntner P. Mortality prediction with a single general self-rated health question. A meta-analysis. *J Gen Intern Med*. 2006;21(3):267–75. doi:10.1111/j.1525-1497.2005.00291.x.
- Frankenberg E, Jones NR. Self-rated health and mortality: does the relationship extend to a low income setting? *J Health Soc Behav*. 2004;45(4):441–52.
- World Health Organization. Statistics Netherlands. Health interview surveys: towards international harmonization of methods and instruments, vol. 58. Copenhagen: WHO Regional Office for Europe, WHO Regional Publications, European; 1996.
- Robine JM, Jagger C, Romieu I. Selection of a coherent set of health indicators for the European Union. Phase II: final report. Montpellier: Euro-REVES; 2002.
- Simon JG, De Boer JB, Joung IM, Bosma H, Mackenbach JP. How is your health in general? A qualitative study on self-assessed health. *Eur J Public Health*. 2005;15(2):200–8. doi:10.1093/eurpub/cki102.
- Bardage C, Plujim SMF, Pedersen N, Deeg DJH, Jylhä M, Noale M, et al. Self-rated health among older adults: a cross-national comparison. *Eur J Ageing*. 2005;2:149–158. doi:10.1007/s10433-005-0032-7.
- Carlson P. Self-perceived health in East and West Europe: another European health divide. *Soc Sci Med*. 1998;46(10):1355–66. doi:10.1016/S0277-9536(97)10093-4.
- Eriksson I, Undén AL, Elofsson S. Self-rated health. Comparisons between three different measures. Results from a population study. *Int J Epidemiol*. 2001;30(2):326–33. doi:10.1093/ije/30.2.326.
- Murray C, Salomon J, Mathers C, Lopez A. Summary measures of population health: concepts, ethics, measurement and applications. Geneva: WHO; 2002.
- Börsch-Supan A, Jürges H. The Survey of health, ageing and retirement in Europe—M, Methodology. Mannheim: MEA; 2005.
- Börsch-Supan A, Brügiavini A, Jürges H, Mackenbach JP, Siegrist J, Weber G. Health, ageing and retirement in Europe. Mannheim: MEA; 2005.

16. Organization for Economic Cooperation and Development. Classifying educational programmes. Manual for ISCED-97 implementation in OECD Countries. Paris: OECD; 1999.
17. Nicholas S, Huppert F, McWilliams B, Melzer D. Physical and cognitive function. In: Marmot M, James B, Blundell R, Lessof C, Nazroo J, editors. Health, wealth and lifestyles of the older population in England: the 2002 English longitudinal study of ageing. London: IFS; 2003.
18. Prince MJ, Reischies F, Beekman AT, Fuhrer R, Jonker C, Kivela SL, et al. Development of the EURO-D scale—a European, Union initiative to compare symptoms of depression in 14 European centres. *Br J Psychiatry*. 1999;174:330–8.
19. Schwarz N. How questions share the answers. *Am Psychol*. 1999;54(2):93–105. doi:[10.1037/0003-066X.54.2.93](https://doi.org/10.1037/0003-066X.54.2.93).
20. Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*. 1979;44(4):443–59. doi:[10.1007/BF02296207](https://doi.org/10.1007/BF02296207).
21. Bollen KA, Long JS. Testing structural equation models. Newbury Park, CA: Sage; 1993.
22. Aitchison J, Silvey SD. The Generalization of probit analysis to the case of multiple responses. *Biometrika*. 1957;44(11/12):131–40.
23. McKelvey RJ, Zavoina W. A statistical model for the analysis of ordinal level, dependent variables. *J Math Sociol*. 1975;4:103–20.
24. Crossley TF, Kennedy S. The reliability of self-assessed health status. *J Health Econ*. 2002;21(4):643–58. doi:[10.1016/S0167-6296\(02\)00007-3](https://doi.org/10.1016/S0167-6296(02)00007-3).
25. Kaplan GA, Goldberg DE, Everson SA, Cohen RD, Salonen R, Tuomilehto J, et al. Perceived health status and morbidity and mortality: evidence from the Kuopio ischaemic heart disease risk factor study. *Int J Epidemiol*. 1996;25(2):259–65. doi:[10.1093/ije/25.2.259](https://doi.org/10.1093/ije/25.2.259).
26. Kunst AE, Bos V, Lahelma E, Bartley M, Lissau I, Regidor E, et al. Trends in socioeconomic inequalities in self-assessed health in 10 European countries. *Int J Epidemiol*. 2005;34(2):295–305. doi:[10.1093/ije/dyh342](https://doi.org/10.1093/ije/dyh342).
27. Mackenbach JP, Martikainen P, Looman CW, Dalstra JA, Kunst AE, Lahelma E. The shape of the relationship between income and self-assessed health: an international study. *Int J Epidemiol*. 2005;34(2):286–93. doi:[10.1093/ije/dyh338](https://doi.org/10.1093/ije/dyh338).
28. van Doorslaer E, Wagstaff A, Bleichrodt H, Calonge S, Gerdtham UG, Gerfin M, et al. Income-related inequalities in health: some international comparisons. *J Health Econ*. 1997;16(1):93–112. doi:[10.1016/S0167-6296\(96\)00532-2](https://doi.org/10.1016/S0167-6296(96)00532-2).
29. van Doorslaer E, Koolman X. Explaining the differences in income-related health inequalities across European countries. *Health Econ*. 2004;13(7):609–28. doi:[10.1002/hec.918](https://doi.org/10.1002/hec.918).
30. Su YP, Ferraro KF. Social relations and health assessments among older people: do the effects of integration and social contributions vary cross-culturally? *J Gerontol B Psychol Sci Soc Sci*. 1997;52(1):S27–36.
31. Jürges H. True health vs response styles: exploring cross-country differences in self-reported health. *Health Econ*. 2007;16(2):168–78. doi:[10.1002/hec.1134](https://doi.org/10.1002/hec.1134).
32. Jylha M, Guralnik JM, Ferrucci L, Jokela J, Heikkinen E. Is self-rated health comparable across cultures and genders? *J Gerontol B Psychol Sci Soc Sci*. 1998;53(3):S144–52.
33. Rubin J. Multiple imputation in sample surveys and censuses. New York: Wiley; 1987.