mea
Mannheim Research Institute for the Economics of Aging

How comparable are different measures of self-rated health? Evidence from five European countries

Hendrik Jürges, Mauricio Avendano,
Johan Mackenbach

137-2007

L13, 17_D-68131 Mannheim_Phone +49 621 181-2773/1862_Fax +49 621 181-1863_www.mea.uni-mannheim.de

How comparable are different measures of self-rated health?

Evidence from five European countries

Hendrik Jürges [1], Mauricio Avendano [2], Johan Mackenbach [2]

1 Mannheim Research Institute for the Economics of Aging (MEA), University of Mannheim, Germany
2 Department of Public Health, Erasmus University Medical Center, Rotterdam, The Netherlands.

Address for correspondence:

Hendrik Jürges
MEA-Universität Mannheim
L13,17
68131 Mannheim
Germany
Fax: +49-621-181-1863
Email: juerges@mea.uni-mannheim.de

Abstract (300 words): Self-rated health (SRH) is a common health measurement in international research. Yet different versions of this item are often applied. This study compares the US (United States) version (from excellent to poor) and the EU (European) version (from very good to very bad) of SRH, and examines differences in their associations with demographic and objective health variables. Data were drawn from the Survey of Health, Ageing and Retirement in Europe (SHARE), comprising information from 11,622 respondents aged 50 years and over in five countries. Respondents were presented with both the EU and US versions. Information was collected on basic demographics and health variables including chronic diseases, symptoms, functional limitations and depression. Firstly, the distribution of each version of the SRH item was assessed, and both relative and literal concordance was examined. Subsequently, multivariate regression analysis was used to assess differences in the associations of both items with demographic and health indicators. The US version has a more symmetric distribution and smaller variance than the EU version. Although the EU version discriminates better at the negative end, the US version shows better discrimination at the negative end of the scale. 69% 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. Furthermore, the two versions were strongly correlated (polychoric correlation = 0.88), had similar associations with demographics and health indicators, and showed a similar pattern of variation across countries. Health levels based on different versions of the self-rated health item 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 EU 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*

Acknowledgements: This paper uses data from the early release 1 of SHARE 2004. This release is preliminary and may contain errors that will be corrected in later releases. The SHARE data collection has been primarily funded by the European Commission through the 5th framework programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life). Additional funding came from the US National Institute on Ageing (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.

Introduction

Self-rated health (SRH) is one of the most widely used health measures, and is one of the health indicators recommended by the World Health Organization (WHO) and the European Commission for health monitoring (Bardage, Plujim, Pedersen, Deeg, Jylhä, Noale et al., 2005; De Bruin, Picavet, & Nossikov, 1996). SRH is a comprehensive measurement that incorporates multiple dimensions of health (Simon, De Boer, Joung, Bosma, & Mackenbach, 2005), and is a major independent predictor of mortality (Appels, Bosma, Grabauskas, Gostautas, & Sturmans, 1996; Burstrom & Fredlund, 2001; DeSalvo, Bloser, Reynolds, He, & Muntner, 2006; Frankenberg & Jones, 2004; Miilunpalo, Vuori, Oja, Pasanen, & Urponen, 1997; Murata, Kondo, Tamakoshi, Yatsuya, & Toyoshima, 2006). Thus, despite its very general and seemingly subjective character, SRH appears to be very useful as a public health indicator(Robine, Jagger, & Romieu, 2002; World Health Organization & Statistics Netherlands, 1996). Differences between countries in SRH levels have been reported and are partly attributable to underlying differences in 'true' health (Bardage et al., 2005; Carlson, 1998). However, these variations could also reflect different cultural perceptions and measurement techniques. In particular, different versions of the SRH item have been used in different surveys, varying in wording and type of scale (Eriksson, Unden, & Elofsson, 2001). The impact of these differences on the way individuals report on their on health has not been studied. Furthermore, it is not known whether different versions of the SRH item may entail different dimensions of health or severity of health problems.

Two five-point scale versions of SRH have been used in international surveys: The first one, henceforth called 'US version' comprises categories ranging from "excellent" to "poor" and has been used in several studies within the United States such as the National Health Interview Survey (NHIS) or the Health and Retirement Study (HRS). The second one, henceforth called 'EU version', ranges from very good to 'very poor' has been applied in several European surveys such as the European Community Household Panel (ECHP) or the Health Survey for England (HSE). Comparing these two versions in the European context is important for two reasons: Firstly, it is not known whether both measures are genuinely comparable, or whether they can be made comparable *ex post*. Existing surveys incorporate only one of the two formats, and observed cross-country differences in SRH may simply be an artefact caused by the use of different versions of this item. Secondly, it has been argued that the US version may be more appropriate for Anglo-Saxon populations such as the United States, where for cultural reasons individuals may be in general more likely to make positive judgements using

wordings such as ‘excellent’. In contrast, this terminology may be less appropriate in European populations, where individuals may be less inclined to make use of this of terminology when making judgements about their health. From this perspective, the EU version would be expected to be more suitable to provide an accurate picture of the ‘true health’ of the European population as compared to the US version.

Following this reasoning, the WHO regional office for Europe and the European community health-monitoring programme have recommended the use of the EU version in the European context(Murray, Salomon, Mathers, & Lopez, 2002; Robine et al., 2002; World Health Organization & Statistics Netherlands, 1996). The justification for this choice lies in the argument that this scale comprises a balanced set of 5 categories, two of which are positive (*very good, good*), one neutral (*fair*), and two negative (*bad, very bad*). Furthermore, this terminology is considered more general and universal as it is presumably translatable in a comparable way to all European languages(World Health Organization & Statistics Netherlands, 1996). Despite this alleged consensus, no studies have empirically examined these advantages in the EU version, and the scientific evidence for the choice of this version of SRH is scarce. In practice, it is uncertain whether this version is indeed preferable than those used in other surveys. As a consequence, despite WHO recommendations, European surveys have continued to apply different wordings in the SRH scale, which makes it impossible to compare results from different surveys(World Health Organization & Statistics Netherlands, 1996).

The present study aims at contributing to this debate by comparing the US and EU versions of SRH across five different European countries. We used data from the SHARE (Survey of Health, ageing and retirement in Europe) study, comprising information from over 11,000 respondents aged 50 and over from five European countries. A unique feature of this study is that participants were presented with both the US and EU version of SRH. This allows comparing both versions among individuals with exactly the same underlying health status. We assess differences in the distribution of health self-ratings across these two measures, and examine the association of the two items with demographic and health variables. This is the first study to assess the comparability of the two most commonly used versions of the SRH item using data from several European countries.

Methods

Study population and data collection

The survey of health, ageing and retirement in Europe (SHARE)

Details on the SHARE study in Europe have been described elsewhere (A. Börsch-Supan, Brugiavini, Jürges, Mackenbach, Siegrist, & Weber, 2005; A. Börsch-Supan & Jürges, 2005). Briefly, in 2004, a survey on health, ageing and retirement was conducted in representative samples of 10 European countries aged 50 or older, comprising a total of 22,777 men and women. Trained interviewers in each country conducted interviews, using a computer assisted personal interviewing (CAPI) program supplemented by a self-completion paper questionnaire. The set-up allowed each country to use exactly the same underlying structure and questionnaire. Translation of the CAPI questionnaire was conducted by expert agencies, and a pre-test and a pilot were conducted. Sample designs differed slightly by population. In most countries, samples were drawn from national or regional population registries. In some countries, a multi-stage sampling procedure was followed in which regions were first selected, and subsequently individuals within these regions were invited to participate. The only exceptions were Austria, Greece and Switzerland, where telephone directories were used as sampling frames. In these three countries and in Denmark, households were the unit of selection. In all other countries, individuals constituted the sampling frame. The average household response rate was 55.4%, ranging from 37% in Switzerland to 69% in France.

Self-rated health (SRH)

Self-rated health was measured in all individuals using both the US and EU versions. The English language version reads "Would you say your health is...". Individuals were subsequently presented with the following options depending on the item version: (1) US version: 'Excellent, very good, good, fair, or poor'; (2) EU version: very good, good, fair, bad, or very bad'. Half of the sample was given the US version at the beginning and the EU version at the end of the physical health part of the interview, whereas the other half was given the EU version first and the US version at the end of the physical health part of the interview. Table 1 shows that in four of the languages in participating countries (German, Greek, Spanish, and Dutch), categories mutual to both formats have exactly the same wording. In the other languages (Italian, French, Danish, and Swedish), translations differ slightly

between formats, so that answer categories are not verbally identical. Therefore, the present study is based on data for the first group of languages, including the following countries: Austria, Germany, Greece, the Netherlands, and Spain.

Table 1: Answer categories for SRH (self-rated health) using the European (EU) and the United States (US) versions in five European countries: The SHARE study

Language	Countries	SRH-EU	SRH-US
German	Austria, Germany	1 Sehr gut 2 Gut 3 Mittelmäßig 4 Schlecht 5 Sehr schlecht	1 Ausgezeichnet 2 Sehr gut 3 Gut 4 Mittelmäßig 5 Schlecht
Spanish	Spain	1 Muy Buena 2 Buena 3 Pasable 4 Mala 5 Muy mala	1 Excelente 2 Muy buena 3 Buena 4 Pasable 5 Mala
Greek	Greece	1.Πολύ καλή 2.Καλή 3.Μέτρια 4.Κακή 5.Πολύ κακή	1.Αριστη 2.Πολύ καλή 3.Καλή 4.Μέτρια 5.Κακή
Dutch	Netherlands, Belgium	1 Heel goed 2 Goed 3 Redelijk 4 Slecht 5 Heel slecht	1 Uitstekend 2 Heel goed 3 Goed 4 Redelijk 5 Slecht

Demographic and health covariates

SHARE contains a broad range of demographic and health measures, both on of physical and mental health. In the present study, the following factors found to be associated with SRH were included: (1) *Age* and *sex* of the respondent. (2*) Educational level* was based on self-reported of the highest level of education, and reclassified using the UNESCO International classification of education (ISCED-97)(Organization for Economic Cooperation and Development, 1999). The ISCED-97 classification scheme has 7 different levels (0 to 6), ranging from pre-primary level of education (e.g. kindergarten) to the second stage of tertiary education (Ph.D.). The original ISCED were recoded into three broader education levels: "low" (pre-primary to lower secondary education; ISCED 0 to 2), "medium" (upper secondary and post-secondary, non-tertiary education; ISCED 3 and 4), and "high" (first and second stage of tertiary education; ISCED 5 and 6). (3) *Diagnosed chronic diseases* were measured by asking individuals to indicate whether they were suffering from any of the diseases or symptoms presented in a showcard. The conditions listed were: heart attack or congestive heart failure, hypertension, high cholesterol, stroke or cerebral vascular disease, diabetes or high blood sugar, chronic lung disease,

asthma, arthritis, osteoporosis, cancer, stomach or duodenal ulcer, Parkinson disease, cataracts, and hip fracture. Individuals' answers were summarised in three categories: no condition, one or two conditions, and three or more conditions. (4) *Symptoms* were measured by asking individuals to indicate whether they were suffering from any of the symptoms presented in a showcard: 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, three or more symptoms. (5) ADL (activities of daily living) comprised information on limitations individuals have with daily activities, i.e., dressing, getting in/out of bed, bathing/showering, using the toilet and eating. (6) IADL (Instrumental activities of daily living) comprised data on limitations with activities such as preparing a meal, shopping, taking medication, making telephone calls, doing housework, or managing money; limitations with ADLs and IADLs were summarised in three categories: no limitation, one or two limitations, three or more limitations. (7) Depression was measured by self-report of diagnosis with depression.

Methods of analysis

Firstly, we assessed the distribution of answers to both versions of the SRH item in the full sample and by country. Cross-tabulations were used to show individual differences in health ratings. Subsequently, we assessed the degree of agreement or concordance between the two measurements. Three conceptually different types of concordance are used.

Literal concordance

Two answers are 'literally' concordant if they are consistent in terms of the verbal representation of the general health state, e.g., if an individual reports being in 'good' health independently of the SRH item presented. Literal concordance also includes cases whereby an individual reports to be in 'excellent' health in the US version and in 'very good' health in the EU version. This accounted for the fact that the EU scale does not allow for better than 'very good' health. Similarly, because the US version does not allow for worse than poor health, cases were coded as literally concordant if they reported to be in 'poor' health in the US version and in 'very poor' health in the EU version.

Relative concordance

Concordance between items was also assessed in terms of the relative position of the answer categories ('relative' concordance). Relative concordance can be understood as a tendency to use scale midpoints as an anchor, whereby individuals conceive the midpoint as the population average or median health (Schwarz, 1999). Thus, if individuals believe their health is above average, they will choose a health category above the midpoint, independently of the verbal representation. An equivalent reasoning applies to responses below the midpoint. Similarly, cases were coded as relatively concordant if they reported to be both in 'good' health when given the US version and 'fair' health when given the EU version, because both categories represent the midpoint of the respective item.

Polychoric correlations

Whereas the measures of concordance describe above implicitly treat general health as a categorical variable, polychoric correlations are based on the assumption that general health is a normally distributed continuous latent trait divided into ordered levels (Olsson, 1979). Pearson product-moment correlations are clearly inappropriate in this context because self-rated health is measured on an ordinal scale and any numbers attached to the categories are arbitrary. Polychoric correlations provide the appropriate alternative in this case, as they assess the relationship between continuous latent variables assumed to underlie the ordered categorical measures. Consider the *measurement model* $Y = bH + e$, where H is latent continuous 'true' general health, Y is an individual's perception of his or her health, e is a measurement error, and b is a regression coefficient. Y is converted into ordered categorical ratings as the individual applies thresholds associated with the rating categories (e.g., very good, good). One interpretation of the polychoric correlation is that if individuals' general health had been measured on the *same* continuous scale, the correlation between both measures would have been equal to this value. In the absence of measurement error and if both scales really measured the same concept, the correlation should be 1. This approach is analogous to the assessment of inter-rater agreement or reliability, with the exception that we compare two ratings on the same unit by the same rater, rather than a rating on the same unit by two raters. The reliability interpretation of polychoric correlations between the two SRH-items applies to the underlying continuous construct, and not to the categorizations. Polychoric correlations were estimated by maximum likelihood (Olsson, 1979).

Ordered probit regressions

To assess whether associations of SRH with demographic and health variables differed for the US and EU versions, ordered probit regressions were estimated (Aitchison & Silvey, 1957; McKelvey & Zavoina, 1975). As polychoric correlations, the ordered probit model is based on the assumption that there is a latent continuous variable underlying the observed ordinal responses. Thresholds divide the real line into segments corresponding to the various ordinal categories. The latent continuous variable 'general health' *H* is modelled as a linear function of covariates X (the demographics and health indicators) plus a disturbance term *u* that has a standard normal distribution. In the language of structural equation modelling, this would be our *structural model*. This approach is comparable to the estimation of ordered logit models, but the latter assumes a disturbance term that follows a logistic cumulative distribution function.

We estimated separate ordered probit models for both SRH-items, allowing for cross-equation correlations of the error terms in order to account for the fact that measurements were made on the same individuals ('seemingly unrelated estimation'). Identification of the ordered probit model is achieved by setting the standard deviation of the error term to 1, thus allowing meaningful comparisons of parameters across equations. Ordered probit regression coefficients summarise the effect of a one-unit increase in the explanatory variables on the continuous (latent) outcome variable.

Results

Descriptive comparisons between the two scales

Table 2 shows the distribution of answers to both versions of the SRH item, by country and for all five countries as a whole. The EU version had a more skewed distribution than the US version. When presented with the EU version, only 1.7% of the individuals rated their health as “very poor” (the bottom category), whereas more than 15% selected the top category "very good". In contrast, when presented with the US version, about the same proportion of individuals selected the top and bottom categories, and the distribution of answers around the midpoint was fairly symmetrical. This pattern was clear in the pooled dataset of all countries, and was slightly more marked in Greece and Austria.

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

	Austria		Germany		Netherlands		Spain		Greece		Total	
	EU	US	EU	US	EU	US	EU	US	EU	US	EU	US
Excellent	N.A.	9.4	N.A.	4.7	N.A.	12.7	N.A.	3.5	N.A.	7.1	N.A.	7.5
Very good	17.8	24.7	11.3	17.1	18.2	18.0	9.9	15.1	21.6	26.9	15.4	19.8
Good	44.0	37.2	44.7	41.1	51.4	43.2	40.7	38.9	41.8	36.4	44.9	39.8
Fair	29.1	22.5	32.1	29.2	24.8	22.1	34.0	31.6	29.6	24.1	29.9	26.1
Poor	7.3	6.2	10.2	8.0	4.8	4.0	12.3	10.9	6.0	5.5	8.2	6.9
Very poor	1.8	N.A.	1.8	N.A.	0.7	N.A.	3.1	N.A.	1.0	N.A.	1.7	N.A.
N	1,896		2,885		2,747		2,262		1,918		11,622	

N.A. indicates not applicable

Individuals appear to be in better health when confronted with the US version of the SRH item (Table 2). When confronted with the US version, 7.5% of participants reported to be in 'excellent' health and 19.8% in 'very good' health. In contrast, when contrasted with the EU version, the proportion of individuals in 'very good' health (the ‘best’ health possible in this version) was only about half the proportion of those in 'very good' and 'excellent' health when presented the US version (15.5% vs. 27.3%). Similarly, whereas about 7% of respondents reported that their health was poor when presented with the US version, about 9.9% reported their health was poor or very poor when presented with the EU version. This pattern was also observed for each country individually (Table 2). These findings suggest that similar verbal presentations in both the EU and US versions elicited different assessments. Individuals provided different answers to both SRH item versions, suggesting that responses to these items are not directly comparable.

Differences at the individual level are shown in Table 3, which shows cross-tabulations of answers given by participants to both item versions. Among those who reported that they were in ‘good' health when confronted with the EU version, 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. About 10% of these participants reported that they were in ‘excellent’, ‘fair’ or ‘poor’ health, which were discordant ratings. A similar pattern was observed for the other health categories.

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

	SRH-US					
SRH-EU	Excellent	Very Good	Good	Fair	Poor	Total (col. %)
Very good	***37.9***	**51.3**	10.5	0.3	0.0	15.4
Good	3.6	*24.9*	**65.7**	5.7	0.1	44.9
Fair	0.2	2.3	*27.8*	**66.6**	3.2	29.9
Poor	0.0	0.0	4.5	*41.2*	**54.3**	8.2
Very poor	0.5	0.5	0.0	10.8	***88.1***	1.7
Total (row %)	7.5	19.8	39.8	26.1	6.9	100.0

Note: Numbers in italics indicate relative concordance, numbers in boldface indicate literal concordance

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, i.e., those who said very good (SRH-EU) and excellent (SRH-US), or very poor (SRH-EU) and poor (SRH-US). In the total sample, 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. Austrian respondents had both the lowest percentage of literally concordant and the highest percentage of relatively concordant answers. Spain had the lowest percentage of relatively concordant answers and the highest percentage of discordant answers.

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

				Polychoric correlation	
Country	% Literally concordant	% Relatively concordant	% Discordant	PC	RMSEA
Austria	64.7	36.9	7.4	0.871	0.061
Germany	70.4	28.1	6.7	0.896	0.049
Netherlands	71.6	29.3	9.1	0.892	0.048
Spain	67.3	27.1	10.5	0.856	0.063
Greece	69.6	31.2	6.8	0.895	0.045
Total	69.0	30.1	8.1	0.857	0.060

The final column of Table 4 shows coefficients of polychoric correlation between the two versions of SRH. Model fit as judged by the root mean square error of approximation (RMSEA) was in the good (<0.05) to acceptable (<0.08) range. The polychoric correlation in the full sample was 0.857, which suggests that the continuous latent variables measured by the two versions were strongly correlated. Correlation coefficients were highest in Germany, the Netherlands and Greece. The lowest correlation was observed in Spain, which is consistent with the large number of discordant answers in this population.

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	Nether-lands	Spain	Greece	Total
Age 50-59	30.9	34.3	41.6	31.7	38.5	35.7
Age 60-69	38.9	38.6	31.6	30.5	29.1	33.9
Age 70-79	21.3	20.4	19.2	26.5	22.7	21.8
Age 80+	8.9	6.7	7.6	11.3	9.7	8.6
Male	42.2	46.8	46.7	42.2	45.6	45.0
Female	57.8	53.2	53.3	57.8	54.4	55.0
Low Education	31.5	17.8	57.5	85.3	63.7	49.9
Medium Education	48.8	56.7	23.0	7.6	22.2	32.4
High Education	19.7	25.5	19.5	7.1	14.1	17.7
No Diagnosed Condition	30.9	27.2	32.1	20.5	27.1	27.7
One Or Two Conditions	54.1	52.8	52.6	51.2	54.5	52.9
Three Or More Conditions	15.0	20.0	15.3	28.3	18.4	19.4
No Symptom	32.3	29.4	38.6	27.3	35.9	32.7
One Or Two Symptoms	50.7	50.5	48.1	42.7	46.2	47.8
Three Or More Symptoms	17.0	20.1	13.3	30.0	17.9	19.5
No (I)ADL Limitation	79.2	84.4	83.5	73.8	80.5	80.6
One Or Two (I)ADL Limitations	14.0	11.0	12.3	17.4	15.0	13.7
Three Or More (I)ADL Limitations	6.8	4.6	4.2	8.8	4.5	5.7
Never Diagnosed With Depression	91.8	89.5	83.3	81.9	96.5	88.1
Ever Diagnosed With Depression	8.2	10.5	16.7	18.1	3.5	11.9

Differences in associations with covariates

Table 5 shows the distribution of covariates in the entire sample and separately by country. About two thirds of participants were between 50 and 69 years of age and one third was 70 or older, and 55% of them were women. About half of participants had a low educational level. However, this percentage varied greatly across countries. In particular, Mediterranean countries such as Spain and Greece had relatively low levels of education as compared to countries such as Germany and Austria. More than two thirds of participants reported one or more diagnosed conditions or symptoms, whereas 20% reported at least one limitation with ADL. About 12% of participants had ever been diagnosed with depression. Health levels varied across countries. The prevalence of chronic diseases, symptoms and

limitations was highest in Spain, while the prevalence of these health problems was lowest in the Netherlands.

Table 6: Ordered probit regressions (fully adjusted models) of self-rated health (SRH) for the EU 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.

	SRH-EU		SRH-US		Cross-equation test
Covariate	Coeff.	SE	Coeff.	SE	p-Value
Age 60-69	0.068	0.026	0.073	0.025	
Age 70-79	0.239	0.029	0.212	0.029	0.008
Age 80+	0.263	0.044	0.159	0.044	
Female	-0.098	0.022	-0.092	0.021	0.754
Medium Education	-0.201	0.028	-0.243	0.027	0.092
High Education	-0.397	0.032	-0.442	0.031	
One or two chronic conditions	0.761	0.027	0.694	0.025	0.004
Three or more chronic conditions	1.173	0.037	1.146	0.037	
One or two symptoms	0.485	0.025	0.471	0.024	0.685
Three or more symptoms	0.955	0.037	0.931	0.037	
One or two (I)ADL problems	0.479	0.033	0.482	0.033	0.345
Three or more (I)ADL problems	1.065	0.054	1.125	0.060	
Ever diagnosed with depression	0.202	0.034	0.204	0.033	0.958
Austria	0.030[ns]	0.024	-0.075	0.024	
Germany	0.277	0.021	0.299	0.020	
Netherlands	-0.145	0.020	-0.128	0.020	<0.001
Spain	0.025[ns]	0.023	0.033[ns]	0.022	
Greece	-0.187	0.023	-0.129	0.022	
Threshold 1	-0.263	0.033	-0.848	0.034	
Threshold 2	1.456	0.035	0.201	0.032	
Threshold 3	2.897	0.042	1.620	0.034	
Threshold 4	4.030	0.054	3.075	0.042	

Note: larger values = worse health. Source. SHARE 2004, release 1: Austria, Germany, Greece, Netherlands, Spain; [ns] coefficient not significant at 99% level

Table 6 shows the ordered probit regression models, which summarize the effect of a one-unit increase in the explanatory variables on continuous (latent) general health. With the exception of three country effects, all coefficients were significantly different from zero, suggesting that all factors were significantly associated with both versions of SRH. Standard errors were marginally smaller when the US version was used, which reflects the more even distribution of this item compared to the EU version. Results indicate that SRH ratings decrease with age even after adjusting for health indicators, which may reflect unmeasured health differences by age. Women reported better health than men even after adjusting for health and demographic indicators, and higher education was associated with better SRH. As expected, worst SRH ratings were strongly associated with a larger number of chronic

conditions, symptoms, ADL and IADL limitations, and depression diagnosis (Table 6). Country differences were also significant. After adjustment for demographic and health variables, SRH ratings were best in Greece and the Netherlands, and worst in Germany.

The associations of SRH with most demographic and health variables were very similar for both the US and EU versions (Table 6). There were only two exceptions to this rule: Firstly, the effect of being older than 80 years old and over on SRH was significantly larger for the EU than for the US version. This probably reflected the better discrimination of the EU version at the negative end of the scale. Secondly, the association of chronic diseases with SRH was larger for the EU and than for the US version. This difference was nevertheless small, and statistical significance was most likely due to the large sample size. For all other variables, there was no evidence of different associations with the EU and US versions of SRH.

Table 6 shows that after adjusting for all demographic and health covariates, individuals rated their health better than average in Greece and the Netherlands, whereas individuals in Germany showed the worst self-ratings of health. This pattern was consistent for both the US and EU versions. However, when measuring health with the EU version, Spanish participants appeared to have slightly better health than Austrians, whereas the opposite pattern was observed when measuring health with the US version. This reflected the discrepancy of estimates for Austria in both version of SRH. Excluding this country from the analysis resulted in identical country rankings of SRH for the US and EU versions.

Table 7 shows the results (p-values) of cross-equation tests of parameter differences for the two version of SRH separately by country. Overall, there was a general pattern of similar associations of SRH measures with health and demographic variables. There were however some differences in Greece, where the association of SRH with age and education was different for the US and EU item version. Furthermore, the magnitude of associations between SRH and the number of symptoms in Austria and the number of conditions in Spain was different for the EU and US versions in these populations. Overall, however, results for most countries showed a similar pattern as for the pooled analysis, namely that associations between covariates and SRH are not statistically different for both the US and EU versions.

Table 7: p-values of cross-equations tests for single-country models among men and women aged 50 years and over in five European countries: The SHARE study

	Austria	Germany	Netherlands	Spain	Greece
Age	0.665	0.157	0.197	0.217	0.000
Sex	0.912	0.656	0.979	0.191	0.735
Education	0.637	0.562	0.660	0.394	0.001
N of conditions	0.277	0.647	0.196	0.006	0.654
N of symptoms	0.034	0.435	0.240	0.670	0.385
N adl/iadl limitations	0.938	0.299	0.933	0.777	0.274
Ever depressed	0.676	0.615	0.890	0.724	0.208
Joint test	0.455	0.564	0.622	0.051	0.000

Discussion

WHO has recommended the EU version as the standard measurement of self-rated health in the European context (Robine et al., 2002; World Health Organization & Statistics Netherlands, 1996). Results from our study suggest that this version does not have clear advantages with respect to the US version of SRH. In fact, the US version has a more symmetrical distribution than the EU version. Furthermore, although the EU version discriminates better at the negative end, the US version shows better discrimination at the positive end of the scale. Comparisons of individual answers to both items further show that the two versions are not fully comparable, neither in a literal or relative sense. Whereas approximately two thirds of respondents provided concordant answers in a literal sense, only one third gave concordant answers in a relative sense. Despite these discrepancies, less than 10% of respondents were discordant in either sense. Furthermore, the US and EU versions were highly correlated, had similar associations with demographic and more objective health indicators, and showed a similar pattern of variation across countries. Overall, this study shows that the two most commonly used versions of SRH are not directly comparable, but are in fact different categorizations of the same latent continuous health variable.

Limitations of the study

This is the first study measuring self-rated health using two different items in a large sample of men and women in several countries. However, some limitations should be considered. The present study is based on data 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. Thus, in populations including individuals at younger ages, the differences between the US and EU item versions may be more marked. In particular, the US version

might be more appropriate in younger cohorts, since it discriminates better at the positive end of the scale. Nevertheless, it is unlikely that associations of SRH with demographic and health variables would be different for both items. Thus, the main conclusions of our study are likely to apply to both younger and older populations in Europe.

Respondents were presented with both versions of the SRH item along with other question on their own health. The order of presentation (i.e., at the beginning or the end) may have had an impact on the response individuals provided, which may bias our comparisons of both versions (Crossley & Kennedy, 2002). In order to tackle this problem, we randomised the order of presentation of both versions at the beginning or at the end of the survey. As a result, half of participants are presented with the US version at the beginning and the EU version at the end, whereas the other half was presented with both versions in the opposite order. Analysis indicated that the order of presentation had little impact on individual's levels of SRH. Thus, it is unlikely that the order of presentation influenced the main conclusions of our study.

Finally, objective health outcomes such as chronic disease and symptoms were measured by self-report. Individuals may underreport certain symptoms and chronic diseases. Furthermore, the concordance between SRH levels and other health measures may have been artificially increased since both assessments stemmed from self-reports. However, previous studies have shown that SRH is a strong predictor of 'more objective' measures including mortality (Appels et al., 1996; Burstrom & Fredlund, 2001; DeSalvo et al., 2006). Furthermore, since all individuals were presented with both versions of SRH, any underreporting would not have been systematic with respect to SRH measure. Thus, comparisons of associations between these objective health outcomes and measures of SRH are unlikely to be biased by underreporting.

Comparison with previous studies

The predictive power of subjective global health assessments has been shown in numerous studies (Appels et al., 1996; Burstrom & Fredlund, 2001; DeSalvo et al., 2006; Kaplan, Goldberg, Everson, Cohen, Salonen, Tuomilehto et al., 1996). 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 more objective health outcomes and demographic variables.

Consistent with findings from single populations (Eriksson et al., 2001), we found that different measures of SRH are strongly correlated, and differences in their distribution are relatively small. Nevertheless, we found that the EU version of SRH has a slightly more skewed distribution than the US version. Despite this, our results confirm findings from previous research suggesting that different measures of SRH seem to represent parallel assessments of subjective health (Eriksson et al., 2001), and further suggest that this pattern is similar across several European countries.

Differences between countries in the level of self-rated health and the association of this variable with socioeconomic and health factors have been reported (Bardage et al., 2005; Carlson, 1998; Jürges, In press; Kunst, Bos, Lahelma, Bartley, Lissau, Regidor et al., 2005; Mackenbach, Martikainen, Looman, Dalstra, Kunst, & Lahelma, 2005; Su & Ferraro, 1997; van Doorslaer & Koolman, 2004; van Doorslaer, Wagstaff, Bleichrodt, Calonge, Gerdtham, Gerfin et al., 1997). 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 minor differences in the wording of response categories. Thus, cross-country comparisons of population health based on different versions of the SRH item may lead to spurious health variations across populations. On the other hand, the associations of SRH with demographic factors such as socioeconomic status were similar for the two SRH item versions. Thus, comparisons of how demographic and other factors relate to self-rated health across surveys using a different 5-point SRH scale (Appels et al., 1996; Bardage et al., 2005; Jylha, Guralnik, Ferrucci, Jokela, & Heikkinen, 1998; Su & Ferraro, 1997; van Doorslaer & Koolman, 2004; van Doorslaer et al., 1997) are unlikely to be biased.

Interpretation and implications

Most health and social surveys contain only one version of the SRH 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 EU version, thus 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 SRH by appropriately rescaling items. Imagine survey A uses the EU-version and survey B applies the US-version of SRH,

while all other variables are measured in the same way. In this situation, the fact that survey A misses data on the US version (and vice versa) can be interpreted as a missing data problem. Since data are missing at random, they can be imputed using conditional probabilities known from surveys such as SHARE (Table 3). In order to 'convert' the EU into the US version, a random number from a uniform distribution on the unit interval, say X, should be drawn. A respondent who has answered 'very good' to the EU version can be coded as being in 'excellent' health if $X < 0.379$ (Table 3), as being in 'very good' health if $0.379 \leq X < 0.892$ (i.e., 0.379 + 0.513), and so on. In principle, this procedure can be repeated several times, yielding multiple imputations (Rubin, 1987).

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 midpoint as the population average health when judging their own health status, independently of the verbal representation (Schwarz, 1999). 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. However, respondents may still use the relative midpoint as average for judging their health when presented with the first item. In either case, the implication of these findings is that presenting all respondents with a 5-point scale is not enough to ensure comparability, as respondents largely use 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 EU 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 SRH 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 SRH as different categorisations of an underlying (latent) continuous health variable. This includes the ordered probit and logit models, and simple probit and logit models if the SRH item is dichotomised, e.g., into good vs. less than good health.

WHO recommends the use of the EU 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 EU 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*)(Robine et al., 2002; World Health Organization & Statistics Netherlands, 1996). In our study, however, this balanced set of categories resulted in a skewed distribution of SRH. In terms of statistical efficiency, the US version has in fact some advantages. In particular, responses to the US version are more evenly distributed across the 5-point scale, resulting in smaller standard errors of the estimated parameter. The fact that both versions are similarly associated with demographic and health determinants further weakens the case for recommending the EU version, as both versions seem to represent parallel assessments of the same latent health variable. Thus, in studies of older European populations, there does not seem to be a strong argument for preferring the so-called EU version. These results invite a reassessment of WHO recommendations. In fact, the choice of an SRH version should be based on several considerations, including aspects such as the age distribution of the population studied, e.g., in older populations, the EU version tends to show a skewed distribution. More crucial than the choice of one or the other item of SRH is the application of a common version in all surveys, or alternatively the rescaling of items for surveys that apply different versions of this item.

REFERENCES

Aitchison, J., & Silvey, S. D. (1957). The Generalization of Probit Analysis to the Case of Multiple Responses. *Biometrika,* 44(11/12), 131-140.

Appels, A., Bosma, H., Grabauskas, V., Gostautas, A., & Sturmans, F. (1996). Self-rated health and mortality in a Lithuanian and a Dutch population. *Soc Sci Med,* 42(5), 681-689.

Bardage, C., Plujim, S. M. F., Pedersen, N., Deeg, D. J. H., Jylhä, M., Noale, M., et al. (2005). Self-rated health among older adults: a cross-national comparison. *Eur J Ageing,* 2, 149-158.

Börsch-Supan, A., Brugiavini, A., Jürges, H., Mackenbach, J. P., Siegrist, J., & Weber, G. (2005). Health, Ageing and Retirement in Europe. Morlenbach: Strauss GmbH.

Börsch-Supan, A., & Jürges, H. (2005). The Survey of Health, Ageing and Retirement in Europe - M, Methodology. Mannheim: MEA.

Burstrom, B., & Fredlund, P. (2001). 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,* 55(11), 836-840.

Carlson, P. (1998). Self-perceived health in East and West Europe: another European health divide. *Soc Sci Med,* 46(10), 1355-1366.

Crossley, T. F., & Kennedy, S. (2002). The reliability of self-assessed health status. *J Health Econ,* 21(4), 643-658.

De Bruin, A., Picavet, H. S. J., & Nossikov, A. (1996). Health interview surveys. Towards international harmonization of methods and instruments. Geneve: WHO, Regional Publications European Ser no 58.

DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., & Muntner, P. (2006). Mortality prediction with a single general self-rated health question. A meta-analysis. *J Gen Intern Med,* 21(3), 267-275.

Eriksson, I., Unden, A. L., & Elofsson, S. (2001). Self-rated health. Comparisons between three different measures. Results from a population study. *Int J Epidemiol,* 30(2), 326-333.

Frankenberg, E., & Jones, N. R. (2004). Self-rated health and mortality: does the relationship extend to a low income setting? *J Health Soc Behav,* 45(4), 441-452.

Jürges, H. (In press). True Health vs Response Styles: Exploring Cross-country Differences in Self-Reported Health. *Health Econ*.

Jylha, M., Guralnik, J. M., Ferrucci, L., Jokela, J., & Heikkinen, E. (1998). Is self-rated health comparable across cultures and genders? *J Gerontol B Psychol Sci Soc Sci,* 53(3), S144-152.

Kaplan, G. A., Goldberg, D. E., Everson, S. A., Cohen, R. D., Salonen, R., Tuomilehto, J., et al. (1996). Perceived health status and morbidity and mortality: evidence from the Kuopio ischaemic heart disease risk factor study. *Int J Epidemiol,* 25(2), 259-265.

Kunst, A. E., Bos, V., Lahelma, E., Bartley, M., Lissau, I., Regidor, E., et al. (2005). Trends in socioeconomic inequalities in self-assessed health in 10 European countries. *Int J Epidemiol,* 34(2), 295-305.

Mackenbach, J. P., Martikainen, P., Looman, C. W., Dalstra, J. A., Kunst, A. E., & Lahelma, E. (2005). The shape of the relationship between income and self-assessed health: an international study. *Int J Epidemiol,* 34(2), 286-293.

McKelvey, R. J., & Zavoina, W. (1975). A Statistical Model for the Analysis of Ordinal Level, Dependent Variables. *Journal of Mathematical Sociology,* 4, 103-120.

Miilunpalo, S., Vuori, I., Oja, P., Pasanen, M., & Urponen, H. (1997). 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,* 50(5), 517-528.

Murata, C., Kondo, T., Tamakoshi, K., Yatsuya, H., & Toyoshima, H. (2006). Determinants of self-rated health: Could health status explain the association between self-rated health and mortality? *Arch Gerontol Geriatr*.

Murray, C., Salomon, J., Mathers, C., & Lopez, A. (2002). Summary measures of population health: concepts, ethics, measurement and applications. In W. H. Organization (Ed.). Geneve.

Olsson, U. (1979). Maximum Likelihood Estimation Of The Polychoric Correlation Coefficient. *Psychometrika,* 44(4), 443-459.

Organization for Economic Cooperation and Development (1999). Classifying Educational Programmes. Manual for ISCED-97 Implementation in OECD Countries. Paris: OECD.

Robine, J. M., Jagger, C., & Romieu, I. (2002). Selection of a Coherent Set of Health Indicators for the European Union. Phase II: Final report. Montpellier: Euro-REVES.

Rubin, J. (1987). Multiple imutation in sample surveys and censuses. New York: Wiley.

Schwarz, N. (1999). How questions share the answers. *American Psychologist,* 54(2), 93-105.

Simon, J. G., De Boer, J. B., Joung, I. M., Bosma, H., & Mackenbach, J. P. (2005). How is your health in general? A qualitative study on self-assessed health. *Eur J Public Health,* 15(2), 200-208.

Su, Y. P., & Ferraro, K. F. (1997). 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,* 52(1), S27-36.

van Doorslaer, E., & Koolman, X. (2004). Explaining the differences in income-related health inequalities across European countries. *Health Econ,* 13(7), 609-628.

van Doorslaer, E., Wagstaff, A., Bleichrodt, H., Calonge, S., Gerdtham, U. G., Gerfin, M., et al. (1997). Income-related inequalities in health: some international comparisons. *J Health Econ,* 16(1), 93-112.

World Health Organization, & Statistics Netherlands. (1996). Health interview surveys: Towards international harmonization of methods and instruments. Copenhagen: WHO Regional Office for Europe, WHO Regional Publications, European Series, no. 58.

Discussion Paper Series

Mannheim Research Institute for the Economics of Aging Universität Mannheim

To order copies, please direct your request to the author of the title in question.

Nr.	Autoren	Titel	Jahr
125-07	Matthias Sommer		07
126-07	Axel H. Börsch-Supan, Anette Reil-Held, Christina B. Wilke	How an Unfunded Pension System looks like Defined Benefits but works like Defined Contributions: The German Pension Reform	07
127-07	Karsten Hank Isabella Buber	Grandparents Caring for Their Grandchildren: Findings from the 2004 Survey of Health, Ageing and Retirement in Europe	07
128-07	Axel Börsch-Supan	European welfare state regimes and their generosity towards the elderly	07
129-07	Axel Börsch-Supan Alexander Ludwig Mathias Sommer	Aging and Asset Prices	07
130-07	Axel Börsch-Supan	Nachfrageseitiger Wettbewerb im Gesundheitswesen	07
131-07	Florian Heiss, Axel Börsch-Supan, Michael Hurd, David Wise	Pathways to Disability: Predicting Health Trajectories	07
132-07	Axel Börsch-Supan	Rational Pension Reform	07
133-07	Axel Börsch-Supan	Über selbststabilisierende Rentensysteme	07
134-07	Axel Börsch-Supan, Hendrik Jürges	Early Retirement, Social Security and Well-Being in Germany	07
135-07	Axel Börsch-Supan	Work Disability, Health, and Incentive Effects	07
136-07	Axel Börsch-Supan, Anette Reil-Held, Daniel Schunk	The savings behaviour of German households: First Experiences with state promoted private pensions	07
137-07	Hendrik Jürges, Mauricio Avendano, Johan Mackenbach	How comparable are different measures of self-rated health? Evidence from five European countries	07