Testing the validity of the vignette approach for improving comparability of self-reported health:
the case of Germany and Sweden

Abstract

Self-assessments are often subject to heterogeneity in reporting behaviour. Anchoring vignettes have been introduced to purge self-assessments of reporting heterogeneity. The aim of this paper is to evaluate the validity of the vignette approach for improving comparability of self-reported health. This is done by observing whether correcting for reporting heterogeneity with the use of anchoring vignettes brings the self-assessments closer to the respondents' objective health status, with the main focus being the comparison of Germany and Sweden. In particular, this paper compares adjusted self-reported health in the domain of mobility and breathing to an objective health index for German and Swedish respondents. The main results indicate that there is some mild evidence that the vignette approach improves cross-country comparability between Germany and Sweden, since it brings the self-assessments closer to the objective situation. More positive results are found for within-country validity, which reveal that adjusted self-reported health is in line, in terms of direction and significance, with objective health measures. Although this research gives some evidence in favour of the validity of the vignette approach, future research should focus on formally testing the assumptions of the anchoring vignettes, response consistency and vignette equivalence, in order to test the validity of the vignette approach more accurately.

Lisa Voois – 388675
Erasmus University Rotterdam

1. Introduction

Self-assessed health (SAH) is a commonly used survey method to collect data on individual health. This popular method entails asking individuals directly to self-report their level of health on an ordered scale. The method seems largely valid, since SAH has proven to be a useful predictor of objective health measures, such as mortality (Idler & Benyamini, 1997). However, self-assessments are likely to be incomparable across individuals, because individuals might tend to interpret an otherwise identical question differently (Brady, 1985). More specifically, take the example of survey respondents, say students, having to answer the question: “Generally, how do you rate your performance in school?” and suppose they can choose from the ordered scale very poor, poor, average, good and very good. Further, suppose that two students, Kate and Tom, both have an average of a 7. Lastly, suppose that Kate rates her performance in school as average and Tom rates his performance in school as good. As such, due to different reference levels against which Kate and Tom judge their performance in school, the results from the survey become incomparable between them. This might also apply to SAH; in particular, individuals with the same objective health status may have different personal frames of reference against which they rate their health (Jürges, 2006). For example, Mathers & Douglas (1998) found that the Aboriginal population of Australia tends to rate their health higher than the general population of Australia, but according to objective health indicators, such as mortality and morbidity, the Aboriginal population is in much poorer health than the general population. This indicates that these sub-populations might have different frames of reference against which they evaluate their health. Then, SAH becomes incomparable across sub-populations, or even across individuals, due to different reporting styles. Several studies have indeed already questioned the comparability of self-assessments across groups of individuals (Lindeboom & Doorslaer, 2004; Jürges, 2006).

Incomparability of survey results can have important policy implications. Following Jürges (2006), there appears to be no clear relationship between health care expenditures and SAH of the elderly across European countries. However, when adjusted SAH is used the relationship becomes positive. This example thus shows that adjusting for cross-country differences in response styles can considerably alter potential policy decisions and outcomes with respect to health care expenditures. Specifically, the example shows that not adjusting for reporting heterogeneity might falsely lead to the belief that investments in health care will not contribute to improved health, while adjusting for reporting heterogeneity seems to indicate the exact opposite. As such, it is in the social interest to ensure that survey results are comparable across individuals.

Moreover, ensuring that survey results are comparable across individuals is also relevant for scientific research. Not only do the social sciences often use survey results as data for their research, there is also a vast amount of scientific literature concerned with testing the comparability of survey results (Bago d'Uva et al., 2008; Bonsang & Soest, 2012; Angelini, Cavapozzi & Paccagnella, 2012). Much of the scientific literature relating to the comparability of SAH tries to establish whether differences in SAH across countries or individuals are genuine or can be attributed to reporting heterogeneity. For example, Lindeboom & Van Doorslaer (2004) find that age and gender lead to different reporting styles. More specifically, they find that female and older respondents tend to assess a given level of health more positively than male and younger respondents. Another important research area is whether health disparities are over- or underestimated when ‘biased’ SAH measures are used. For example, Johnston, Propper & Shields (2009) find that using widely available self-reported chronic health measures might underestimate the true income-health gradient. Bago d’Uva et al. (2008) also find that using unadjusted SAH leads to an underestimation of the income-health gradient. This implies that, given the same level of health, low-income respondents tend to report higher levels of health than high-income respondents. Furthermore, Bago d’Uva et al. (2008) find that the female health disadvantage might be over- or underestimated, depending on the specific country, when unadjusted SAH is used. These examples thus already indicate possible reporting heterogeneity in health by age, gender and income.

One way to adjust SAH for reporting heterogeneity is by using anchoring vignettes. The use of anchoring vignettes to correct for reporting styles was introduced by King et al. (2004). Basically, respondents are not only asked for self-assessments, but they are also asked to rate several hypothetical individuals, the vignettes, on the same ordered scale. The following is an example of a vignette description from the health literature[1]:

Alice has pain in her knees, elbows, wrists and fingers, and the pain is present almost all the time. Although medication helps, she feels uncomfortable when moving around, holding and lifting things.

This vignette description thus provides a given level of health, or more specifically pain, which is fixed across respondents. As such, anchoring vignettes make it possible to measure each individual's unique differential item functioning (DIF), i.e. their unique reporting style. The vignette descriptions can be seen as providing an 'anchor', which connects the respondent's subjective vignette rating to the fixed level of health represented by the vignette. As such, the vignette ratings can be used to adjust the self-assessments by removing reporting heterogeneity, which creates interpersonally comparable survey results. However, the validity of the vignette approach is dependent on two assumptions: response consistency and vignette equivalence (King et al., 2004). Response consistency requires that respondents should have the same frame of reference for evaluating their own health on a particular domain, e.g. mobility, and for evaluating the health of the accompanying hypothetical individual(s), i.e. the vignette(s). Otherwise, the reporting style for self-assessments will not be accurately measured. The second assumption, vignette equivalence, requires that, apart from random measurement error, all respondents understand the level of a particular variable portrayed in a certain vignette, e.g. mobility impairments, in the same way and on the same one-dimensional scale. Thus, any differences in how respondents perceive the level of a particular variable represented in a vignette must be random and independent of respondent characteristics.

King et al. (2004) originally applied the method of anchoring vignettes to correct for reporting heterogeneity in the domain of political efficacy. Since then, anchoring vignettes have been applied in other research areas as well. As mentioned above, vignettes have been widely applied in the domain of health (Bago d'Uva et al., 2008; Sirven, Santos-Eggimann & Spagnoli, 2012), but also in research related to life and job satisfaction among the elderly (Angelini et al., 2012; Bonsang & Soest, 2012) and recently vignettes have also been applied in research related to education (Voňková et al., 2015).

To conclude, Jürges & Soest (2012) report that although the possible advantages of using straightforward self-reported measures for comparative scientific research are evident, respondents from different countries or socioeconomic groups tend to have different frames of reference when evaluating themselves. As such, self-assessments should either be corrected with the help of anchoring vignettes or should be replaced by objective measures. Further research should indicate whether the two methods lead to similar adjustments in SAH. This is particularly interesting, since it enables an assessment of the validity of the vignette approach. More specifically, if the two methods yield similar results, the vignette approach will not only appear to be valid, but this also implies that the use of adjusted self-assessments in scientific research seems justified, since adjusted self-assessments seem to accurately represent objective health.

[1] Taken from the first wave (2004) of the Survey of Health, Ageing and Retirement in Europe (SHARE).

Therefore, the aim of this paper is to evaluate the validity of the vignette approach for adjusting SAH. Specifically, this paper tries to establish whether SAH adjusted by anchoring vignettes and replacing SAH by a health index based upon objective measures leads to similar results in respondents' health. In order to do so, this paper will refer to results from Jürges (2006).

Jürges (2006) decomposed cross-country differences in general SAH into parts explained by differences in 'true' health and parts explained by cross-country differences in reporting styles. He used data from the first wave (2004) of the Survey of Health, Ageing and Retirement in Europe (SHARE), which contains data from 10 European countries and primarily includes respondents aged 50 and over (Börsch-Supan, 2013). True health was estimated by computing a comparable health index, which was based upon a variety of objective health measures, such as diagnosed chronic diseases, mental illnesses and measurements like BMI. The index ranges from 0 to 1, where a value of 1 indicates the absence of any impairment. The existence of a health condition decreases the health index by a given proportion, or more specifically, by the disability weight for that specific health condition.

According to the main results, the comparison of Germany and Sweden based on general SAH is the exact opposite of the comparison based on objective health measures. This discrepancy provides motivation for testing the validity of the vignette approach for improving cross-country comparability of SAH. More specifically, the Germans are healthier than the Swedes when sorted by median health using the health index described above. Thus, according to the proxy for true health, German respondents are healthier than Swedish respondents. As such, it would also be expected that the self-assessments of general health follow this pattern. Yet, the Swedes actually report considerably higher levels of general SAH then their German counterparts. In fact, Germans rate their general SAH among the lowest in Europe, whilst their true health scores among the highest in Europe. This large discrepancy indeed turns out to be the result of differences in response styles across countries. Specifically, Swedish respondents tend to largely overestimate their health (compared to the SHARE average), whereas German respondents tend to systematically underestimate their health. If anchoring vignettes can solve this discrepancy, the vignette approach will appear valid for improving cross-country comparability of SAH.

Therefore, this paper will try to analyse whether the correction made for German and Swedish respondents by the anchoring vignettes is in line with a comparable health index based on the index constructed by Jürges. As such, it will try to test the validity of the vignette approach by observing whether correcting for reporting heterogeneity with the use of anchoring vignettes brings the self-assessments closer to the respondents' objective health status. Although the main focus of this paper, as motivated by Jürges' results, will lie on testing the comparison between Germany and Sweden, i.e. testing the validity of the vignette approach for improving cross-country comparability, some attention will also be paid to evaluating within-country validity with the help of socioeconomic variables.

The rest of this paper will be ordered as follows. Section two describes the data and methodology used. Section three describes the results of the analyses. In section four, the results are discussed, conclusions are drawn and additionally limitations of the research and suggestions for further research are discussed.

2. Data & Methodology

Data

Data from the first wave (2004) of the Survey of Health, Ageing and Retirement in Europe (SHARE) is used (Börsch-Supan, 2013). The first wave includes self-assessments and vignette ratings for several health domains, namely for pain, sleep, mobility, memory, breathing, affect and work disability. For the first six health domains, each self-assessment is accompanied by three vignette questions. For the last domain, work disability, nine vignette questions are included. Furthermore, wave 1 includes two versions of the vignettes, called type A and type B. The types differ with respect to the order of the questions and gender of the individuals described in the vignettes. The two types of vignette questionnaires were appointed randomly to the respondents. After each vignette description, respondents are asked to rate the health of the hypothetical individuals on the same ordered scale on which they evaluated themselves. Moreover, prior to rating the vignettes, respondents are asked to assume that the hypothetical individuals have the same age and background as themselves to ensure that respondents will evaluate the health of the hypothetical individuals in the same way as they judge their own health, i.e. with the same frame of reference.

In his analysis, Jürges makes use of self-reported general health provided in the first wave of the SHARE data. However, the vignette questionnaires do not include assessments for general health. They only include assessments for the seven health domains mentioned above. Since the health index constructed by Jürges mainly features physical impairments in the domain of mobility and breathing, such as asthma, arthritis, chronic lung disease, Parkinson disease et cetera, the domains mobility and breathing are chosen to be able to evaluate the validity of the vignette approach. In order to do this, a new health index will be constructed which includes only diseases and conditions which could contribute to mobility impairments or breathing problems.

Besides the self-assessments, the vignettes and the health index, a country dummy, which takes the value 1 for Germany and 0 for Sweden, will be included to test for reporting heterogeneity between the two countries. Furthermore, the socioeconomic variables age, gender and education will be added as control variables. Age is a continuous variable and represents the age of the respondent at the time of the survey, which ranges from 40 until 96. Gender is represented by the dummy variable female, which takes the value 1 when the respondent is female and 0 if the respondent is male. Education is measured according to the International Standard Classification of Education (ISCED 1997). This classification consists of 7 educational levels, ranging from level 0 to 6. For the purpose of this paper, some educational levels were merged to result in only 3 educational levels: 1) completed at most lower secondary or the second stage of basic education (ISCED 0-2), 2) completed upper secondary education (ISCED 3-4), 3) completed the first or second stage of tertiary education, such as an university degree (ISCED 5-6) (UNESCO, 1997). The first educational level is the reference category, while the latter two are dummy variables. Note that age, gender and education are added as control variables, because this research wants to investigate if cultural differences between Germany and Sweden result in different threshold values, not if differences in the composition of the respondents (age, educational level) result in different threshold values. Moreover, the control variables are also used to look at within-country validity of the vignette approach.

Mobility

Self-assessments and vignette ratings in the domain of mobility are obtained from the question: “Overall in the last 30 days, how much of a problem did you have with moving around?” and the five response categories are none, mild, moderate, severe and extreme.
The three vignette descriptions for mobility are, respectively:

Vignette 1: [Tom/Sue] has a lot of swelling in his/her legs due to his/her health condition. He/she has to make an effort to walk around his/her home, as his/her legs feel heavy.

Vignette 2: [Kevin/Lisa] does not exercise. He/she cannot climb stairs or do other physical activities because he/she is obese. He/she is able to carry the groceries and do some light household work.

Vignette 3: [Rob/Eve] is able to walk distances of up to 200 metres without any problems, but feels tired after walking one kilometre or climbing more than one flight of stairs. He/she has no problems with day-to-day activities, such as carrying food from the market.

Table 1 gives a summary overview of the descriptive statistics for the control variables in the domain of mobility.

Table 1 – Descriptive statistics for the control variables in the domain of mobility

	Germany		Sweden	
Variables	Mean	Standard deviation	Mean	Standard deviation
Female	0.566	0.496	0.522	0.500
Age	63.064	9.169	63.887	9.540
Educational level				
1) ISCED 0-2	0.171	0.377	0.534	0.499
2) ISCED 3-4	0.594	0.492	0.234	0.424
3) ISCED 5-6	0.235	0.424	0.232	0.422
Respondents (N)	502		406	

As can be seen in table 1, the composition of German respondents is relatively the same as the composition of Swedish respondents in terms of gender and age. However, German respondents seem to be higher educated than Swedish respondents. The educational levels show that the most striking difference can be found in the proportion of respondents having enjoyed at most lower secondary or the second stage of basic education (ISCED 0-2). Only 17% of the German respondents report level 1 education, in comparison to at least 53% of the Swedish respondents. As a result, the largest proportion of Swedish respondents has only finished level 1 education, while the largest proportion of German respondents has finished level 2 education. Lastly, it can also be inferred from the table that the sample size (N) is larger for Germany than for Sweden.

Besides looking at the composition of German and Swedish respondents, it will also be informative to look at the self-assessments and vignette ratings in more detail. Table 2 gives a summary overview of the self-assessments and vignette ratings in the domain mobility.

Table 2 – Self-reported health and vignette ratings in the domain of mobility

	Germany		Sweden	
Ratings	Frequency	Percentage	Frequency	Percentage
Self-rating				
1) Extreme	2	0.40	3	0.74
2) Severe	36	7.17	19	4.68
3) Moderate	97	19.32	74	18.23
4) Mild	135	26.89	157	38.67
5) None	232	46.22	153	37.68
Total	502	100.00	406	100.00
Vignette 1				
1) Extreme	43	8.60	97	24.13
2) Severe	283	56.60	235	58.46
3) Moderate	132	26.40	59	14.68
4) Mild	36	7.20	11	2.74
5) None	6	1.20	0	0.00
Total	500	100.00	402	100.00
Vignette 2				
1) Extreme	26	5.21	7	1.75
2) Severe	216	43.29	142	35.41
3) Moderate	182	36.47	183	45.64
4) Mild	58	11.62	64	15.96
5) None	17	3.41	5	1.25
Total	499	100.00	401	100.00
Vignette 3				
1) Extreme	4	0.80	2	0.51
2) Severe	84	16.87	38	9.60
3) Moderate	247	49.60	136	34.34
4) Mild	134	26.91	163	41.16
5) None	29	5.82	57	14.39
Total	498	100.00	396	100.00

According to table 2, there seems to be little difference between the self-reported mobility of German and Swedish respondents. Most German respondents report no mobility impairment, while most Swedish respondents report a mild impairment. German respondents, however, report relatively slightly more moderate and severe mobility impairments compared to Swedish respondents. Overall, there seems to be no clear indicator that the Swedes assess their mobility higher than the Germans, in contrast to Jürges' result for general SAH. Before comparing the vignette ratings of Germany and Sweden, the frequencies in table 2 already show considerable variation within countries. Thus, although the vignettes represent fixed descriptions of mobility impairments, respondents rate the severity of this fixed impairment quite differently. This variation might therefore be an indicator of reporting heterogeneity within countries with respect to socioeconomic characteristics, such as age and educational level. However, this paper is specifically interested in differences in response styles across countries. According to the first vignette, Germans tend to rate the health of the hypothetical individual described in the vignette higher than the Swedes. Specifically, a substantially larger proportion of the Swedish respondents tends to rate the impairment as extreme compared to the proportion of German respondents. For the second vignette, Swedish respondents seem to rate the impairment slightly more mildly than German respondents. Particularly, a larger proportion of Swedish respondents reports a mild or moderate impairment in comparison with German respondents, with the largest proportion of Swedes reporting a moderate impairment. In turn, most German respondents report a severe impairment. For the third vignette, Swedish respondents seem to clearly rate the impairment more mildly than their German counterparts. Specifically, Swedish respondents are more likely to report no or a mild impairment compared to German respondents, whereas German respondents are more likely to rate the impairment as moderate or severe. Overall, there seems to be some evidence that the Swedes overestimate the fixed health levels, however differences in reporting styles between the two countries seem to be rather small. Lastly, note that the number of respondents differs slightly per vignette, since it was decided to keep respondents who rated at least one vignette, and thus not necessarily rated all vignettes.

Breathing

Self-assessments and vignette ratings in the domain of breathing are obtained from the question: “In the last 30 days, how much of a problem did you have because of shortness of breath?” and the five response categories are none, mild, moderate, severe and extreme.
The three vignette descriptions for breathing are, respectively:

Vignette 1: [Mark/Karen] has no problems with walking slowly. He/she gets out of breath easily when climbing uphill for 20 meters or a flight of stairs.

Vignette 2: [Paul/Karen] suffers from respiratory infections about once every year. He/she is short of breath 3 or 4 times a week and had to be admitted in hospital twice in the past month with a bad cough that required treatment with antibiotics.

Vignette 3: [Henri/Maria] has been a heavy smoker for 30 years and wakes up with a cough every morning. He/she gets short of breath even while resting and does not leave the house anymore. He/she often needs to be put on oxygen.

Table 3 gives a summary overview of the descriptive statistics for the control variables in the domain of breathing.

Table 3 – Descriptive statistics for the control variables in the domain of breathing

	Germany		Sweden	
Variables	Mean	Standard deviation	Mean	Standard deviation
Female	0.561	0.497	0.526	0.500
Age	63.010	9.223	63.961	9.554
Educational level				
1) ISCED 0-2	0.172	0.378	0.535	0.499
2) ISCED 3-4	0.593	0.492	0.235	0.424
3) ISCED 5-6	0.234	0.424	0.230	0.421
Respondents (N)	499		409	

According to table 3, the age and gender of respondents are relatively similar for Germany and Sweden. However, German respondents are again higher educated than Swedish

respondents, with the biggest difference being the relative number of respondents who have enjoyed at most lower secondary or the second stage of basic education (ISCED 0-2).
These descriptive statistics are largely similar to the descriptive statistics in the domain of mobility, given that mainly the same respondents rated the vignettes in the domain of mobility and breathing. However, in the domain of breathing the number of German respondents is slightly smaller and the number of Swedish respondents is slightly larger than in the domain of mobility (table 1).

As before, it will be informative to look at the self-assessments and vignette ratings in more detail. Table 4 gives a summary overview of the self-assessments and vignette ratings in the domain of breathing.

Table 4 – Self-reported health and vignette ratings in the domain of breathing

	Germany		Sweden	
Ratings	Frequency	Percentage	Frequency	Percentage
Self-rating				
1) Extreme	4	0.80	9	2.20
2) Severe	17	3.41	34	8.31
3) Moderate	51	10.22	90	22.00
4) Mild	100	20.04	119	29.10
5) None	327	65.53	157	38.39
Total	499	100.00	409	100.00
Vignette 1				
1) Extreme	8	1.61	14	3.47
2) Severe	159	31.99	146	36.14
3) Moderate	245	49.30	179	44.31
4) Mild	73	14.69	62	15.35
5) None	12	2.41	3	0.74
Total	497	100.00	404	100.00
Vignette 2				
1) Extreme	46	9.26	112	27.93
2) Severe	274	55.13	199	49.63
3) Moderate	120	24.14	59	14.71
4) Mild	38	7.65	28	6.98
5) None	19	3.82	3	0.75
Total	497	100.00	401	100.00
Vignette 3				
1) Extreme	195	39.08	156	38.52
2) Severe	226	45.29	202	49.88
3) Moderate	41	8.22	29	7.16
4) Mild	18	3.61	15	3.70
5) None	19	3.81	3	0.74
Total	499	100.00	405	100.00

According to table 4, there seems to be clear evidence that the Germans rate their health in the domain of breathing higher than the Swedes. Specifically, more than 65% of German respondents reports no breathing problems, in comparison to only 38% of the Swedes. Moreover, Swedish respondents report relatively more severe and extreme breathing problems. As such, according to the self-assessments, the Germans seem to be healthier than the Swedes in the domain of breathing. The frequencies for the vignette ratings in table 4 again show substantial variation within countries, which might be evidence for reporting heterogeneity by socioeconomic characteristics within countries. In contrast, there seems to be little difference between reporting styles across countries, according to the first vignette. The Germans seem to rate the breathing problems slightly more mildly than the Swedes, since German respondents are more likely to report no breathing problems and less likely to report extreme breathing problems compared to the Swedes. However, overall differences seem small. Germans seem to clearly rate the breathing problems described in the second vignette more mildly than the Swedes. Specifically, more than 27% of the Swedes rates the problems as extreme, in comparison to only 9% of the Germans. As a result, the Germans report relatively more often either no, mild or moderate breathing problems than the Swedes. For the third vignette, there seems to be no evidence of reporting heterogeneity by country. Germans seem to rate the breathing problems described in the vignette slightly more mildly, since they are more likely to report no breathing problems. However, overall differences are small or non-existent. Overall, there seems to be evidence that the Germans might overestimate their health, given that they also seem to overestimate the fixed health levels portrayed in the vignettes.

Health index

As mentioned before, a health index is constructed, based on Jürges' health index, but only including health conditions and measures in the domain of mobility and breathing. Specifically, the following conditions and measures are included: heart attack or heart failure, chronic lung disease, asthma, arthritis or rheumatism, osteoporosis, cancer or malignant tumour, Parkinson disease, hip or femoral fracture, low grip strength, which has proven to be a useful indicator of mobility impairments (Sallinen et al., 2010), low walking speed, obesity, ever treated for depression, since depression and anxiety can significantly contribute to impaired breathing, or more formally, dyspnea (Neuman et al., 2006) and other conditions, since these primarily include back, hip or other joint problems. Obesity is defined as having a BMI equal to or larger than 30, low walking speed is defined as walking 0.4 metres per second or less and grip strength is defined as low for men when their maximum grip strength is below 37 kg and for women when their maximum grip strength is below 21 kg.

Note that most of these conditions and measures are actually quasi-objective, since the SHARE questionnaire asks the respondents to indicate conditions which were ever diagnosed by a doctor. BMI is also self-reported, since respondents self-report their height and weight. Measures such as walking speed and grip strength are, however, completely objective. Only one health index is constructed, since many conditions contribute both to mobility impairments and to breathing problems, among others obesity, chronic lung disease, rheumatoid arthritis, which is often associated with interstitial lung disease (Michaud & Wolfe, 2007) and osteoporosis, which is frequently linked to chronic obstructive pulmonary disease (COPD) (Jørgensena et al., 2007). Specifically, for the latter two, the associated lung diseases might not have been recogized by a physician (and thus not selected), when they are actually present. The disability weights are used which Jürges generated from within his sample, which are largely similar to the ones found in earlier studies. Table 5 summarizes the mentioned health conditions and measures by country, and additionally shows the sample size and the implied disability weights generated by Jürges.

Table 5 – Prevalence of health conditions and measures by country

		Germany			Sweden		
Condition or measure	Disability weight	Mean	Standard deviation	N	Mean	Standard deviation	N
Heart attack or heart failure	0.098	0.107	0.309	506	0.125	0.331	409
Chronic lung disease	0.097	0.047	0.213	506	0.034	0.182	409
Asthma	0.054	0.036	0.185	506	0.071	0.257	409
Arthritis or rheumatism	0.093	0.130	0.340	506	0.093	0.291	409
Osteoporosis	0.075	0.063	0.244	506	0.017	0.130	409
Cancer or malignant tumour	0.089	0.069	0.254	506	0.088	0.284	409
Parkinson disease	0.145	0.004	0.063	506	0.000	0.000	409
Hip or femoral fracture	0.055	0.014	0.117	506	0.015	0.120	409
Other joint problems	0.093	0.166	0.372	506	0.298	0.458	409
Ever treated for depression	0.047	0.449	0.500	107	0.477	0.502	107
Low grip strength	0.048	0.133	0.340	490	0.159	0.367	389
Low walking speed	0.118	0.083	0.280	36	0.103	0.307	39
Obesity	0.052	0.165	0.371	504	0.150	0.357	407

First of all, table 5 shows that not all respondents participated in the grip strength or walking speed test. Specifically, for the measurement of walking speed only respondents aged 75 or over were eligible to take the test, which explains the low number of observations. Furthermore, not all respondents answered the question whether they had ever been treated for depression (by a doctor or psychiatrist), given that this question only had to be answered if respondents had ever suffered from symptoms of depression that lasted at least two weeks. This also explains the rather high mean values found for this variable. Lastly, some respondents did not report their height and weight, and as such some observations for obesity are missing. Since the health index will range from 0 to 1, where a value of 1 indicates the absence of any impairment and the existence of a health condition decreases the health index by the disability weight for that specific condition, missing observations are set to indicate no impairment. That is, missing observations for grip strength are coded as normal grip strength, missing observations for walking speed are coded as normal walking speed, respondents who did not answer the question whether they had ever been treated for depression are coded as not having been treated for depression and respondents who did not report their height and weight are coded as having a normal BMI.

Based on Jürges’ results, it would be expected that the Germans have overall less prevalence of the health conditions and measures, and are thus healthier than the Swedes in the domain of mobility and breathing. However, overall differences between the two countries are rather small. As such, it does not seem that German respondents are healthier than Swedish respondents. Lastly, note that the sample size includes respondents who rated at least one vignette in the domain of mobility or one vignette in the domain of breathing.

Methodology

For each health domain, three models are estimated to analyse differences in reporting styles. The models are, respectively, the standard ordered probit model for self-reported health, a generalized ordered probit model for the vignette ratings and an interval regression for self-reported health. The latter two regressions make up the hierarchical ordered probit

(HOPIT) model, which is used to adjust self-reported health. Lastly, a fractional probit model is generated to model the health index.

The first model, the ordered probit model, is the standard model for the estimation of probabilities when responses are ordinal. The model assumes an unobserved true health variable Y_i^*, on which observed self-reported health Y_i, which is reported on an ordered scale, is based. Y_i^* is specified as:

$$Y_i^* = \ x_i\beta + \varepsilon_i, \qquad \varepsilon_i \sim N(0,1)$$

where x_i is a vector of observed respondent characteristics and ε_i is a random error term independent of x_i. The variance of the error term and the constant term are not identified, since true health is unobserved and self-reported health is an ordinal variable, and are therefore normalized to 1 and 0, respectively. Y_i is related to Y_i^* as follows:

$$Y_i = k \ \leftrightarrow \ \mu^{k-1} \ \leq \ Y_i^* < \ \mu^k, \qquad k = 1, \ldots, 5$$

where k represents the response categories, which range from 1 to 5, and μ^k are the cut-points between the response categories. Moreover, $\mu^0 < \ u^1 < \cdots < \mu^5$ and $\mu^0 = -\infty$, $\mu^5 = \infty$. The response categories k, and self-reported health Y_i, are increasing in good health. Thus, the response category extreme equals 1 and the response category none equals 5.

True health Y_i^* is located somewhere along the ordered scale. As such, true health should lie in between the first μ^{k-1} and last μ^k cut-point. Cut-points, or thresholds, between the different response categories are fixed among respondents in the ordered probit model, which corresponds to the assumption of homogenous reporting. As discussed previously, this assumption might not hold since individuals tend to have different frames of reference against which they evaluate their health. Reporting heterogeneity will make self-reported health incomparable across individuals, since the ordered probit model will produce biased estimates of the coefficients β. Specifically, the coefficients β will mirror both health and reporting effects.

The generalized ordered probit model, proposed by Terza (1985), is used to model the vignette ratings. It extends the standard ordered probit model by allowing the cut-points to be functions of x_i. More specifically, each vignette j $(= 1,2,3)$ represents a fixed true health level Y_{ij}^{v*}, which should rule out any association between the true health level Y_{ij}^{v*} and respondent characteristics x_i. As such, any variation in vignette ratings can be ascribed to reporting heterogeneity. Therefore, the true health level Y_{ij}^{v*} represented by any vignette j as perceived by individual i can be written as follows:

$$Y_{ij}^{v*} = \ \alpha_j + \ \varepsilon_{ij}^v, \qquad \varepsilon_{ij}^v \sim N(0,1)$$

where α_j is an intercept normalized to 0 and ε_{ij}^v is again a random error term, which is independent of measurement error in each vignette ε_{ij}^v and independent of x_i.

The observed vignette rating Y_{ij}^v is related to Y_{ij}^{v*} in the following way:

$$Y_{ij}^v = k \ \leftrightarrow \ \mu_i^{k-1} \leq \ Y_{ij}^{v*} < \ \mu_i^k, \qquad k = 1, \ldots, 5$$

where $\mu_i^0 \ < \mu_i^1 \ldots < \mu_i^5$ and $\mu_i^0 \ = \ -\infty$, $\mu_i^5 \ = \ \infty$ for all i.

Thus, since the true health level Y_{ij}^{v*} is independent of respondent characteristics x_i, all variation in vignette ratings can be attributed to reporting heterogeneity. As such, the cut-points can be written as functions of x_i:

$$\mu_i^k = \gamma_0^k + x_i\gamma^k, \qquad k = 1, \ldots, 4$$

where γ_0^k represents the intercepts in the cut-points.

This generalised ordered probit model allows for non-parallel cut-points shifts, since reporting behaviour across cut-points γ^k is allowed to vary. Non-parallel cut-point shifts occur when reporting heterogeneity is more evident at some levels of health, e.g. the threshold from mild to moderate, than at others, e.g. the threshold from severe to extreme. On the other hand, if reporting heterogeneity is constant across cut-points ($y^1 = \cdots = y^4$), then there are parallel cut-point shifts. Allowing for non-parallel cut-point shifts is important, since significant cut-point shifts can remain undetected if the model does not allow for non-parallel cut-point shifts (Jones et al., 2013). Note that the generalized ordered probit model represents the vignette component of the HOPIT model.

Lastly, an interval regression for self-reported health is estimated. This interval regression represents the second component of the HOPIT model. Just as the ordered probit model, the interval regression assumes an unobserved true health variable Y_i^*, on which observed self-reported health Y_i is based, and relates Y_i^* to Y_i. However, in the interval regression the cut-points are no longer fixed, but can differ across individuals based on the vignette component, i.e. based on the generalized ordered probit model. Y_i^* is defined as:

$$Y_i^* = \beta_0 + x_i\beta + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

where the error term ε_i is again independent of respondent characteristics x_i and independent of the error terms for the vignette ratings ε_{ij}^v. Y_i is related to Y_i^* as follows:

$$Y_i = k \leftrightarrow \mu_i^{k-1} \leq Y_i^* < \mu_i^k, \qquad k = 1, \ldots, 5$$

where $\mu_i^0 < \mu_i^1 < \cdots < \mu_i^5$ and $\mu_i^0 = -\infty$, $\mu_i^5 = \infty$. Note that this corresponds to the ordered probit model, only now the cut-points μ_i^k are allowed to differ across individuals (i). Following the response consistency assumption, the cut-points for self-reported health are set equal to the cut-points for the vignettes, which are written as a function of x_i.

Lastly, a fractional probit model is used to model the health index. The health index is obtained using various health conditions and measures in the domain of mobility and breathing (see data section) and functions as a proxy for true health Y_i^*. The fractional probit model is specified in largely the same way as the ordered probit model, namely as:

$$Y_i^* = x_i\beta + \varepsilon_i, \qquad \varepsilon_i \sim N(0,1)$$

However, Y_i^* is now defined as a fractional response, that is:

$$0 \leq Y_i^* \leq 1$$

The fractional probit model is chosen, since the dependent variable, i.e. the health index, can theoretically take values between, and inclusive, 0 and 1.

3. Results

Mobility

Based on Jürges' results, it could be expected that the Swedes self-report higher levels of health in the domain of mobility than the Germans. The self-assessments are analysed using the standard ordered probit model. Table 6 shows the results of the regression.

Table 6 – Ordered probit for self-reported health in the domain of mobility

Variables	Coefficients	P-value
Germany	0.0487903	0.549
Female	-0.0202792	0.785
Age	-0.0186139	0.000
ISCED 3-4	0.0195348	0.837
ISCED 5-6	0.2077847	0.049
Sample size (N)	908	

First of all, notice that the coefficients of the covariates represented in table 6 should be interpreted qualitatively. That is, a positive coefficient should be interpreted as having a positive effect on reporting a higher category of self-reported health. As was expected from inspecting the frequencies of the response categories in the data section, there seems to be no difference between the self-reported health of German respondents and Swedish respondents in the domain of mobility. German respondents tend to rate their health in the domain of mobility higher according to the positive coefficient, however this positive effect on reporting a higher category is insignificant. Therefore, the self-assessments in the domain of mobility don't seem to follow Jürges' results: in the domain of mobility there is no significant difference in the self-reported health of Germans and Swedes. Furthermore, the results in table 6 indicate two significant relationships at a 5% significance level, namely a significant negative relationship between age and health and a significant positive relationship between health and higher education (ISCED 5-6) in comparison to the reference category (ISCED 0-2).

Based on Jürges' results, the Swedes are likely to overestimate the vignettes in comparison to the Germans. This is analysed using the generalized ordered probit model. Table 7 shows the results of this regression.

Table 7 – Generalized ordered probit for vignette ratings in the domain of mobility

	Cut-point 1		Cut-point 2		Cut-point 3		Cut-point 4	
Variables	Coefficients	P-value	Coefficients	P-value	Coefficients	P-value	Coefficients	P-value
Germany	-0.2558849	0.004	0.0562592	0.336	0.1907635	0.003	0.0272014	0.784
Female	0.2544268	0.003	-0.0243697	0.646	-0.1270183	0.032	-0.2141251	0.025
Age	-0.0116523	0.016	-0.008093	0.006	-0.0027211	0.403	0.0053725	0.301
ISCED 3-4	-0.0071902	0.946	-0.0697662	0.309	0.0088748	0.908	0.1161712	0.319
ISCED 5-6	-0.07192	0.536	0.0530557	0.478	-0.0297421	0.722	0.1285257	0.326
Sample size (N)	2696[i)]							

[i)] Note: this corresponds with the previous 908 observations, only the data was converted to a 'long form' to accommodate the generalized ordered probit model.

Regarding table 7, notice that the response categories are increasing in good health. As such, cut-point 1 represents the shift from extreme to severe, cut-point 2 represents the shift from severe to moderate, cut-point 3 represents the shift from moderate to mild and cut-point

4 represents the shift from mild to none. A significant positive shift in the cut-points represents higher health standards. Before turning to the variable of interest, Germany, it is worth inspecting the other variables. Females tend to have higher health standards regarding the distinction between extreme and severe mobility impairments, however as health increases females tend to have lower health standards. Older individuals, on the other hand, tend to have somewhat lower health standards across all health levels, except for the distinction between reporting a mild or no mobility impairment. However, the magnitude of the cut-point shifts for age seems rather small in comparison to the other variables, especially in comparison to the shifts by gender and country, which might indicate effectively no reporting heterogeneity by age. At best, older individuals might slightly overestimate their health at lower health levels in comparison to their younger counterparts. Moreover, there seems to be no reporting heterogeneity concerning education levels, since for both education levels none of the coefficients are significant.

The results show two significant cut-points shifts for Germany, namely cut-point 1 and cut-point 3. However, the effects regarding reporting behaviour vary across the two cut-points. Specifically, Swedes have a significantly higher threshold value for the shift from extreme to severe (cut-point 1). This implies that Swedish respondents are less likely to report that a given vignette description coincides with a severe, rather than an extreme, mobility impairment compared to the Germans. This effect turns around for the third cut-point. Here, the Germans have higher health standards. In other words, the Swedes are more likely to classify the mobility impairment as mild, rather than as moderate, in comparison to German respondents. Thus, different reporting styles are indeed observed for Germany and Sweden. A test of significance indeed confirms the observation that there is evidence of reporting heterogeneity by country, since the null hypothesis of homogeneous reporting is strongly rejected (P-value = 0.0007). However, it is less clear whether Swedish respondents overestimate or underestimate fixed health levels relative to German respondents. At a low health level, cut-point 1, the Swedes seem to underestimate their health, whereas for a higher health level, cut-point 3, they seem to overestimate their health. Thus, at different levels of mobility impairments, the Swedes (and Germans) have different reporting styles.

Given the observation that reporting heterogeneity seems present by country, gender and age, but not by education, it seems likely that the simple ordered probit model suffers from reporting heterogeneity. A test of joint significance indicates that there is indeed strong evidence of reporting heterogeneity for all variables across all cut-points, since the null hypothesis of reporting homogeneity is strongly rejected (P-value = 0.0000). Lastly, it is worth mentioning that the coefficients differ quite substantially across cut-points, which indicates non-parallel cut-point shifts. As expected from the results, there is strong evidence of non-parallel cut-point shift by all variables considered (P-value = 0.0000) and also for Germany separately (P-value = 0.0003). As such, the choice of the non-parallel model over the parallel model seems justified.

Lastly, adjusted self-reported health in the domain of mobility is investigated using an interval regression. Table 8 shows the results of this regression.

Table 8 – Interval regression for self-reported health in the domain of mobility

Variables	Coefficients	P-value

Germany	0.1364818	0.199
Female	-0.185826	0.056
Age	-0.0236412	0.000
ISCED 3-4	0.0781392	0.528
ISCED 5-6	0.3395711	0.014
Sample size (N)	908	

First of all, note that the magnitude of the coefficients in table 8 cannot be directly compared to the coefficients in table 6, given that in the ordered probit model the scale was normalized to 1, while this is not the case for the interval regression. However, the significance of the coefficients can be compared. In comparing tables 6 and 8, it is noticeable that the coefficient for female turns significant, while the sign of the coefficient remains the same. That is, women tend to be in worse health than men, which seems to mimic the female health disadvantage. Specifically, in table 8 the coefficient for female is significant at a 10% significance level, while in table 6 the coefficient for female was highly insignificant (P-value = 0.785). This most likely implies that females, in comparison to males, tend to overestimate their health, which seems to correspond to the overall lower health standards found for females in table 7. The result found in table 6 was therefore most likely a mixture of a true negative health effect and a tendency for women to report higher levels of health. As such, ignoring reporting heterogeneity could hide a true negative health effect of being female.

As can be seen in table 8, the variable of interest, Germany, does not become significant at a 10% level. However, table 7 did indicate reporting heterogeneity by country. To estimate the overall direction of this reporting heterogeneity, partial effects of Germany on the probability of reporting no mobility impairment are estimated for a reference individual. The reference individual is a 65-year-old Swedish male who has enjoyed at most lower secondary or the second stage of basic education (ISCED 0-2). Table 9 shows the partial effects.

Table 9 – Partial effects of Germany on the probability of reporting no mobility impairment for a reference individual

Model	Coefficient	P-value
Ordered probit	0.0187543	0.550
Interval regression	0.0404126	0.202

According to table 9, both effects are insignificant, as was to be expected from table 8. However, the magnitude of the coefficients changes slightly. Specifically, disregarding significance, if the reference individual would be German, compared to Swedish, his probability of reporting no mobility impairment would increase by approximately 2 percentage points according to unadjusted health. This effect increases for adjusted health. Specifically, again disregarding significance, if the reference individual would be German, instead of Swedish, this would increase his probability of having no mobility impairment by 4 percentage points. This seems to indicate that the Swedes indeed somewhat overestimate their health in comparison to the Germans, since the partial effect purged of reporting heterogeneity increases the probability of reporting no mobility impairment. However, although the partial effect in table 9 increases in magnitude once it is purged of reporting heterogeneity, it does not reach statistical significance. Therefore, since both partial effects are not statistically significant, the apparent result that the Swedes somewhat overestimate their health in comparison to the Germans, should be interpreted with some caution. As such, these results do not give substantial evidence that the Swedes overestimate their health in the domain of mobility in comparison to the Germans.

Breathing

The self-assessments in the domain of breathing are also analysed using the standard ordered probit model. Table 10 shows the results of this regression.

Table 10 – Ordered probit for self-reported health in the domain of breathing

Variables	Coefficients	P-value
Germany	0.6268216	0.000
Female	-0.1184411	0.125
Age	-0.0102604	0.016
ISCED 3-4	0.0245993	0.803
ISCED 5-6	0.1650384	0.130
Sample size (N)	908	

In accordance with the data section, table 10 shows that the Germans self-report higher levels of health in the domain of breathing than the Swedes. This positive relationship between being German and self-reported health is significant at a 1% level. Note that this is contrary to what would have been expected based on Jürges analysis of general health, where the Germans rated their health far lower than the Swedes. Furthermore, the results in table 10 indicate one more significant relationship, namely a significant negative relationship at a 5% significance level between age and health.

The vignette ratings are again analysed using a generalized ordered probit model. Table 11 shows the results of this regression.

Table 11 – Generalized ordered probit for vignette ratings in the domain of breathing

	Cut-point 1		Cut-point 2		Cut-point 3		Cut-point 4	
Variables	Coefficients	P-value	Coefficients	P-value	Coefficients	P-value	Coefficients	P-value
Germany	-0.3322403	0.000	-0.2289152	0.000	-0.1426439	0.064	-0.732317	0.000
Female	0.0028957	0.962	-0.1493389	0.005	-0.1427914	0.037	-0.0643151	0.560
Age	-0.0190269	0.000	-0.0158213	0.000	-0.0116903	0.001	-0.0095288	0.094
ISCED 3-4	0.1081164	0.166	-0.0125446	0.854	-0.0849558	0.338	0.0769524	0.581
ISCED 5-6	0.2402485	0.004	0.1887562	0.014	0.1671879	0.102	0.6313932	0.002
Sample size (N)	2703[ii]							

[ii] Note: this corresponds with the previous 908 observations, only the data was converted to a 'long form' to accommodate the generalized ordered probit model.

Regarding table 11, females seem to have lower health standards given that cut-points 2 and 3 are significant, at a 1% and 5% level respectively. Specifically, women are more inclined to rate a given breathing problem as moderate, rather than severe (cut-point 2), and as mild, rather than moderate (cut-point 3) in comparison to men. Older individuals seem to have lower health standards across all health levels. However, the magnitude of the cut-point shifts by age seems again rather small in comparison to the other variables, which might indicate effectively no reporting heterogeneity by age, or at best indicate that older respondents slightly overestimate their health in comparison to younger respondents. There also seems to be reporting heterogeneity for the higher educated (ISCED 5-6) in comparison to the reference category (ISCED 0-2). Higher educated respondents seem to have higher health standards across all health levels, except maybe for the distinction between moderate and mild breathing problems (cut-point 3). This most likely indicates that higher educated respondents, in comparison to respondents who have finished at most lower secondary or the second stage of basic education (ISCED 0-2), tend to underestimate their health, due to higher health standards. Note that this result was not found in the domain of mobility. Moreover, there seems to be no reporting heterogeneity for level 2 education (ISCED 3-4), since none of the coefficients are significant.

According to table 11, German respondents tend to have lower health standards across all health levels, given that all cut-points are significant and have the same sign. Note that all cut-points are significant at a 1% level, except for the distinction between moderate and mild breathing problems (cut-point 3), which is significant at a 10% level. This indicates that the Germans and the Swedes tend to have different reporting styles. Specifically, the results indicate that German respondents are likely to overestimate their health, since they have lower health standards, relative to Swedish respondents. Note that this result is contrary to Jürges' result. A test of significance indeed confirms the observation that there is evidence of reporting heterogeneity by country, since the null hypothesis of homogenous reporting is strongly rejected (P-value = 0.0000).

Since the results in table 11 indicate that reporting heterogeneity is present by country, gender, age and higher education, it seems likely that the standard ordered probit model suffers from reporting heterogeneity. A test of joint significance indicates that there is indeed strong evidence of reporting heterogeneity for all variables across all cut-points, since the null hypothesis of reporting homogeneity is strongly rejected (P-value = 0.0000). Lastly, it is worth mentioning that the coefficients vary considerably across cut-points, mostly in terms of magnitude, which tends to indicate non-parallel cut-point shifts. As expected, there is indeed strong evidence of non-parallel cut-point shift by all variables considered (P-value = 0.0033) and also for Germany individually (P-value = 0.0002). As such, the choice of the non-parallel model over the parallel model seems again justified.

Lastly, adjusted self-reported health in the domain of breathing is generated using an interval regression. Table 12 shows the results of the interval regression.

Table 12 – Interval regression for self-reported health in the domain of breathing

Variables	Coefficients	P-value
Germany	0.3725431	0.002
Female	-0.2903704	0.007
Age	-0.024227	0.000
ISCED 3-4	0.0483609	0.725
ISCED 5-6	0.6671479	0.000
Sample size (N)	908	

In comparing tables 10 and 12, it is noticeable that the coefficients for female and higher education (ISCED 5-6) turn significant, while the sign of the coefficients remains the same. That is, women tend to be in worse health than men and higher educated (ISCED 5-6) respondents tend to be in better health than lower educated respondents (ISCED 0-2), which corresponds to the results in the domain of mobility. Specifically, in table 12 the coefficients for female and higher education are significant at a 1% level, while in table 10 the coefficients for female and higher education were just slightly above 10% significance. This most likely indicates that women, in comparison to men, tend to overestimate their health, which corresponds to the overall lower health standards found for women in table 11. Similarly, higher educated respondents, in comparison to lower educated respondents, tend to underestimate their health, which corresponds to the overall higher health standards found for higher educated respondents in table 11. The result found for women in table 10 is therefore most likely a combination of a true negative health effect and an inclination for women to report better health. Again, similarly, the result found in table 10 for higher educated respondents, seems to be a mixture of a true positive health effect and an inclination for higher educated respondents to report worse health. Note that the same result for women was found in the domain of mobility.

As can be seen in table 12, the Germans are healthier than the Swedes at a 1% significance level. According to table 11, the Germans have lower health standards, which would indicate that the Germans overestimate their health in the domain of breathing in comparison to the Swedes. In order to test this, partial effects are estimated for a reference individual to compare the magnitude of the ordered probit model and the interval regression. The reference individual is the same as before: a 65-year-old Swedish male who has at most enjoyed level 1 education (ISCED 0-2). Table 13 shows the partial effects of Germany on the probability of reporting no breathing problems for the reference individual.

Table 13 – Partial effects of Germany on the probability of reporting no breathing problems for a reference individual

Model	Coefficient	P-value
Ordered probit	0.2456238	0.000
Interval regression	0.1057458	0.002

According to table 13, the coefficients of the two models vary quite a bit. Using unadjusted health, if the reference individual would be German, instead of Swedish, this would increase his probability of reporting no breathing problems by approximately 25 percentage points. This partial effect decreases when using adjusted health. Specifically, if the reference individual would be German, compared to Swedish, this would increase his probability of having no breathing problems by approximately 11 percentage points. This indicates, in accordance with the lower health standards found for German respondents in table 11, that German respondents overestimate their health in comparison to Swedish respondents, since the partial effect purged of reporting heterogeneity decreases the probability of having no breathing problems. As such, the health effect found in table 10 seems to be a combination of a positive health effect of being German and a tendency for Germans to report better health in the domain of breathing. Note that if reporting heterogeneity was not considered, the health gap in the domain of breathing between Germans and Swedes would be overestimated.

Lastly, the overall results seem to show the importance of adjusting for reporting heterogeneity. Not correcting for reporting styles would mask a true negative health effect of being female, a true positive health effect of being higher educated and would fail to recognize that Germans tend to overestimate their health in comparison to the Swedes in the domain of breathing.

Health index

According to Jürges' paper, it would be expected that the Germans are healthier than the Swedes according to objective health measures, which are represented by the health index. Table 14 shows the results of the fractional probit model.

Table 14 – Fractional probit model for the health index in the domain of mobility and breathing

Variables	Coefficients	P-value
Germany	0.0414693	0.315
Female	-0.0693487	0.059
Age	-0.0149054	0.000
ISCED 3-4	0.0550176	0.249
ISCED 5-6	0.137614	0.016
Sample size (N)	915	

As can be inferred from table 14, women and older respondents are significantly, at a 10% and 1% level respectively, less healthy than men and younger respondents, based on the proxy for true health. Higher educated individuals (ISCED 5-6) are significantly, at a 5% level, healthier in comparison to their lower educated counterparts (ISCED 0-2). These results seem to be in accordance with adjusted health in the domain of mobility and breathing. However, note that there is no significant difference between the health of Germans and Swedes. Specifically, although the coefficient is positive, it does not reach statistical significance. This seems to be in line with adjusted health in the domain of mobility, however not with adjusted health in the domain of breathing. Moreover, as was expected from investigating self-reported health, respondents who have enjoyed upper secondary education (ISCED 3-4) are not significantly healthier than lower educated respondents (ISCED 0-2).

4. Discussion

Conclusion

The aim of this paper is to try to test the validity of the vignette approach for improving comparability of self-reported health by observing whether correcting for reporting heterogeneity with the use of anchoring vignettes brings self-assessments closer to the respondents' objective health status. As mentioned in the introduction, Jürges' result that the comparison of Germany and Sweden is opposite when using self-assessments and objective health measures, provides motivation for testing the cross-country validity of the vignette approach. As such, the main focus of this paper is to test the cross-country comparability of self-reported health by comparing the results of the vignette approach to the results of the objective health index for Germany and Sweden in the domain of mobility and breathing.

According to the (adjusted) self-assessments, there appears to be no significant difference between the health of Germans and Swedes in the domain of mobility and the Germans appear to be healthier than the Swedes in the domain of breathing. Moreover, according to the objective health index there is no significant difference between the health of German and Swedish respondents in both domains, which is in accordance with (adjusted) self-reported health in the domain of mobility. Given that the analysis is not opposite when comparing self-reported health and adjusted self-reported health, it seems somewhat difficult to test the validity of the vignette approach based on the comparison of Germany and Sweden, especially for the domain of mobility, since objective health seems to be in correspondence with both adjusted and unadjusted self-reported health. However, adjusted self-reported health in the domain of breathing seems to be closer to objective health, given that, once the self-assessments have been purged of reporting heterogeneity, the probability of reporting no breathing problems when being German, instead of Swedish, decreases (table 13). Thus, the difference between objective health and self-reported health becomes smaller when self-reported health is adjusted for heterogeneity in reporting behaviour. Still, although the coefficient decreases in magnitude, the partial effect remains significant and is thus not in line with objective health. Nonetheless, the anchoring vignettes seem to correct the self-assessments in the right direction.

Although the main goal of the analysis is the comparison of Germany and Sweden for evaluating the cross-country validity of the vignette approach, some interesting results are observed regarding other variables. Specifically, the comparison of Germany and Sweden does not give much insight into the validity of the vignette approach, but the control variables for gender and education seem to indicate more clearly that self-reported health adjusted by the anchoring vignettes is indeed closer to the objective situation.

Specifically, the results for the variable female indicate that in both domains, adjusted self-reported health is closer to objective health. According to the self-assessments, there is no significant difference between the health of women and men. However, when the self-assessments are purged of reporting heterogeneity, women appear to be in worse health than men, which corresponds to objective health measures. The results for the variable higher education (ISCED 5-6) also show that adjusting for reporting heterogeneity brings self-reported health closer to objective health, but only in the domain of breathing. Specifically, according to the self-assessment there is no significant difference between the health of higher educated respondents (ISCED 5-6) and lower educated respondents (ISCED 0-2). However, when self-reported health is corrected for response styles, higher educated respondents turn out to be healthier than their lower educated counterparts, which is in accordance with objective measures.

Overall, the results seem to indicate that self-reported health adjusted by anchoring vignettes is indeed closer to the objective situation, represented by the health index. The vignette approach seems to give some mild evidence for improving the cross-country comparability of self-reported health. Specifically, although the anchoring vignettes seem to adjust in the right direction, they are not able, it seems, to purge the self-assessments completely of reporting heterogeneity, as is the case for Germany in the domain of breathing. The vignette approach seems to give stronger evidence for improving within-country comparability of self-assessments, since adjusting for reporting heterogeneity brings the self-assessments of female and higher educated respondents in line with objective health measures in terms of the direction and significance of the coefficients. As such, it seems justified to draw the mild conclusion that the vignette approach seems valid, given that it brings the self-assessments and objective measures closer together.

Limitations and suggestions for further research

However, this research is also subject to limitations. First of all, this research does not allow for a formal testing of the assumptions of the vignette approach, response consistency and response equivalence. Response consistency requires that respondents use the same reference level when evaluating themselves and the vignettes, such that reporting heterogeneity is accurately measured. Vignette equivalence entails that perceived differences elicited by a vignette description are random and independent of individual characteristics. This research only informally tests the validity of the vignette approach, by observing whether the direction and significance of self-reported health is closer to objective health once the self-assessments have been purged of reporting heterogeneity, but it does not allow for comparing the magnitude of the adjustments made by the two methods. As such, this research does not account for differences in the magnitude of the coefficients nor does it elucidate why the self-assessments corrected for reporting styles are not similar, in terms of significance, to objective measures, as is the case for Germany in the domain of breathing. It could be that response consistency and/or response equivalence do not hold. The formal testing of these assumptions should be able to shed more light on this. Specifically, formal tests of response consistency can also be used to compare whether the thresholds identified by the anchoring vignettes are similar to the thresholds identified by objective measures, and thus compare the magnitude of the adjustments made by the two methods (Bago d'Uva et al., 2011). Therefore, this seems a good direction for further research.

The fact that the vignettes do not seem to be able to completely purge self-assessments of reporting heterogeneity, might also be because different vignettes might lead to different, or even contrary, measures of DIF. Although this research does not test whether certain vignettes lead to contradicting DIF, Voňková & Hullegie (2011) found that the vignette method is indeed sensitive to the choice of the vignette. As such, further research should be focused on how to 'correctly' formulate vignettes such that the DIF will be accurately measured. Moreover, the validity of the vignette approach might also be sensitive to the health domain being investigated (Voňková & Hullegie, 2011). As such, the limited scope of the research is another limitation to which this research is subject. Specifically, this research only evaluates the validity of the vignette approach in the domain of mobility and breathing and only for a subset of variables. The overall results indicate that the vignette approach is successful in bringing self-assesments and objective measures closer together, however this should not be taken as an indicator that the vignette approach is completely valid. Rather, more research should be done to investigate the validity of the vignette approach more broadly, specifically in other health domains.

Other limitations of this research concern the use of (quasi-) objective health measures and the construction of the health index. As mentioned in the data section, most of the objective health measures used to construct the health index are actually quasi-objective. Respondents are asked to indicate conditions which were diagnosed by a doctor. However, respondents might have misunderstood the doctor and indicated conditions which were never diagnosed, or respondents might suffer from conditions which were never diagnosed. Moreover, it seems unlikely that these conditions alone should capture all the variation in objective health to function as an accurate proxy for true health. Respondents might not suffer from any of the conditions, but still experience mobility impairments or breathing problems. Furthermore, in order to construct the health index missing observations were set to indicate no impairment. As such, it could be that respondents' health is not accurately measured. Lastly, for the purpose of this research, one health index is constructed, which captures conditions and measures both in the domain of mobility and breathing. Although mobility impairments and breathing problems can be closely related, it is still possible that certain respondents have difficulty moving around, without suffering from shortness of breath. As such, the health index might not accurately measure the true level of health in the domain of mobility and/or in the domain of breathing. However, it is beyond the scope of this research to create a specific index by domain, especially since accurate and truly objective measures for mobility impairments and breathing problems are lacking. Future research concerned with testing the validity of the vignette approach should therefore be concerned with finding more truly objective measures of health, such as elaborate tests performed by a physician, to ensure that true health is accurately measured (per domain).

Lastly, another avenue for further research entails differences in language use that can affect the relationship between true health and self-reported health. Specifically, this research does not follow Jürges' result that the Swedes vastly overestimate their health in comparison to the Germans. Of course, this could be (partially) attributed to the fact that this research tests health only on two domains, mobility and breathing, and not on general health. However, Jürges (2006) gives another possible explanation, namely differences in language use for the response categories. For example, Jürges uses the ordinal scale excellent, very good, good, fair and poor. Excellent is translated as 'ausgezeichnet' in German, which Germans often consider as a sarcastic overstatement and consequently would not often use or choose in a health context. Given that the word 'ausgezeichnet' does not appear in the German response categories used in the domain of mobility and breathing, this might provide a partial explanation why Swedish respondents do not vastly overestimate their health in comparison to German respondents. Considering this, future research could focus on investigating the effects of different translations of the response categories on reporting heterogeneity.

Acknowledgements

This paper uses data from SHARE Waves 1, 2, 3 (SHARELIFE), 4 and 5 (DOIs: 10.6103/SHARE.w1.500, 10.6103/SHARE.w2.500, 10.6103/SHARE.w3.500, 10.6103/SHARE.w4.500, 10.6103/SHARE.w5.500), see Börsch-Supan et al. (2013) for methodological details. The SHARE data collection has been primarily funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812) and FP7 (SHARE-PREP: N°211909, SHARE-LEAP: N°227822, SHARE M4: N°261982). Additional funding from the German Ministry of Education and Research, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

Bibliography

Angelini, V., Cavapozzi, D., & Paccagnella, O. (2012). Cross-Country Differentials in Work Disability Reporting Among Older Europeans . *Social Indicators Research* , 211-226.

Angelini, V., Cavapozzi, D., Corazzini, L., & Paccagnella, O. (2012). Age, Health and Life Satisfaction Among Older Europeans . *Social Indicators Research* , 293-308.

Bago d'Uva, T., Lindeboom, M., O'Donnell, O., & Van Doorslaer, E. (2011). Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity . *The Journal of Human Resources* , 875-906.

Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M., & O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics* , 351-375.

Bonsang, E., & Soest, A. (2012). Satisfaction with Job and Income Among Older Individuals Across European Countries . *Social Indicators Research* , 227-254.

Börsch-Supan, A. (2013). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 1. Release version: 5.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w1.500

Börsch-Supan, A., A. Brugiavini, H. Jürges, J. Mackenbach, J. Siegrist and G. Weber. (2005). Health, ageing and retirement in Europe – First results from the Survey of Health, Ageing and Retirement in Europe. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

Börsch-Supan, A. and H. Jürges (Eds.). (2005). The Survey of Health, Ageing and Retirement in Europe – Methodology. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., Zuber, S. (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). International Journal of Epidemiology DOI: 10.1093/ije/dyt088.

Brady, H. (1985). The Perils of Survey Research: Inter-Personally Incomparable Responses. *The Society for Political Methodology* , 269-291.

Idler, E., & Benyamini, Y. (1997). Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies . *Journal of Health and Social Behaviour* , 21-37.

Jørgensena, N., Schwarza, P., Holmeb, I., Henriksenc, B., Petersend, L., & Backerb, V. (2007). The prevalence of osteoporosis in patients with chronic obstructive pulmonary disease - A cross sectional study. *Respiratory Medicine* , 177-185.

Jürges, H. (2006). True Health vs. Response Styles: Exploring Cross-Country Differences in Self-Reported Health. *Health Economics* , 163-178.

Jürges, H., & Soest, A. (2012). Comparing the Well-Being of Older Europeans: Introduction . *Social Indicators Research* , 187-190.

Johnston, D., Propper, C., & Shields, M. (2009). Comparing subjective and objective measures of health: Evidence from hypertension for the income/health gradient . *Journal of Health Economics* , 540-552.

Jones, A., Rice, N., Bago d'Uva, T., & Balia, S. (2013). Reporting heterogeneity in health. In A. Jones, N. Rice, T. Bago d'Uva, & S. Balia, *Applied Health Economics* (pp. 70-105). New York: Routledge.

King, G., Murray, C., Salomon, J., & Tandon, A. (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research . *American Political Science Review* , 191-207.

Lindeboom, M., & Van Doorslaer, E. (2004). Cut-point shift and index shift in self-reported health . *Journal of Health Economics* , 1083-1099.

Mathers, C., & Douglas, R. (1998). Measuring progress in population health and wellbeing. In R. Eckersley, *Measuring progress: Is life getting better?* (pp. 125-155). Collingwood: CSIRO Publishing .

Michaud, K., & Wolfe, F. (2007). Comorbidities in rheumatoid arthritis. *Best Practice & Research Clinical Rheumatology* , 885-906.

Neuman, A., Gunnbjörnsdottir, M., Tunsäter, A., Nyström, L., Franklin, K., Norrman, E., et al. (2006). Dyspnea in relation to symptoms of anxiety and depression: A prospective population study. *Respiratory Medicine* , 1843-1849.

Sallinen, J., Stenholm, S., Rantanen, T., Heliövaara, M., Sainio, P., & Koskinen, S. (2010). Hand-Grip Strength Cut-Points to Screen Older Persons at Risk for Mobility Limitation. *Journal of the American Geriatrics Society* , 1721-1726.

Sirven, N., Santos-Eggimann, B., & Spagnoli, J. (2012). Comparability of Health Care Responsiveness in Europe . *Social Indicators Research* , 255-271.

Terza, J. (1985). Ordinal probit: a generalisation. *Communication in Statistics Theory and Methods* , 1–11.

UNESCO. (1997). *International Standard Classification of Education.* Paris: UNESCO.
Voňková, H., & Hullegie, P. (2011). Is the anchoring vignette method sensitive to the domain and choice of the vignette? *Journal of the Royal Statistical Society* , 597-620.

Voňková, H., Zamarro, Gema, Deberg, V., & Hitt, C. (2015). Comparisons of Student Perceptions of Teacher’s Performance in the Classroom: Using Parametric Anchoring Vignette Methods for Improving Comparability.