ERAMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis

Comparability of Self-Assessed Health Across Countries

03-12-2015

Author: Martijn Oeij

Student number: 346271

Supervisor: Dr. T.M. Bago d'Uva

Co-evaluator: Kim van Wilgenburg

ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

Contents

Abstract

This study aims to establish whether people from different countries respond to self-assessed health (SAH) questions differently, regardless of their health state. Reporting heterogeneity might cause results from SAH studies to be biased. Using anchoring vignettes and data from the first two waves of the Survey of Health, Ageing and Retirement in Europe (SHARE), this thesis investigates reporting heterogeneity in SAH across different EU countries. Also, there is a comparison made of the results between two waves of data, to look into possible differences in reporting heterogeneity over time. In this thesis SAH is divided among six health domains: pain, sleeping problems, mobility, memory, breathing problems and depression. Estimating an ordered probit model with the vignettes as the dependent variable shows that there indeed exists a difference in response scales when comparing countries across the six health domains. The answers to the vignettes are then used to correct the six SAH measures for reporting heterogeneity, using a HOPIT model. When comparing the results for corrected SAH across two waves of SHARE data, it can be seen that conclusions for SAH differences between countries are not constant over time. These results imply that reporting heterogeneity in SAH across countries exists and that results for these country differences are prone to changes over time when comparing two waves of panel data.

1. Introduction

Self-assessed health (SAH) is shown to be a good predictor for mortality (Idler & Benyamini, 1997). It is a measure widely used in research for within and between country comparisons, but there are questions about its accuracy. A typical SAH question would be phrased as “Would you say your health is...”, which respondents can answer with *excellent*, *very good*, *good*, *fair* or *poor* (SHARE, 2006). Logically, ill health, mental health and health behaviors/lifestyle have a great influence on SAH (Manderbacka, 1998). However, response variation in SAH is also found to be associated with the kind of survey, verbal or written, and the sequence of preceding questions (Crossley & Kennedy, 2002). Also, SAH is prone to interviewer effects (Blom & Korbmacher, 2013) and responses vary with different kinds of survey designs, e.g. different answer categories to the general SAH question (Lumsdaine & Exterkate, 2013). Hernandez-Quevedo et al. (2004) analysed British Household Panel Survey data in which they found that a change of the SAH question caused respondents to report differently, regardless of their objective health. Furthermore, there exist doubts about the reliability of self-reported measures in specific diseases, e.g. cancer and cardiovascular diseases (Newell et al., 1999). If biases exist in certain disease areas, this leads to biased calculations of the prevalence of sickness.

One important way in which SAH might be biased is due to social desirability. This means that respondents tend to respond in a way that is socially acceptable, e.g. leading to an underreporting of depression (Logan et al., 2008). Another example of a bias that could occur in SAH is reference bias, which means that respondents rate their own health according to some reference health state. These subjective reference health states are, however, not fixed over time, since respondents can change their reference. McPhail et al. (2010) show that asking respondents to consider descriptions of either an extremely good or an extremely bad health state, before SAH questions, changes the SAH responses.

When these biases occur randomly, they are random errors and there is no problem with drawing conclusions from such an analysis. However, when certain groups deal differently with SAH questions and surveys, there exists bias in estimated results and direct conclusions from these biased results are unreliable. Socioeconomic status seems an important factor for the way in which respondents react to a SAH question. This is due to the fact that people that come from different socioeconomic groups have different reference levels and therefore they frame their health accordingly. This causes variations in reference levels by e.g. income or age (Krause & Jay, 1994). Evidence for different reporting styles by socioeconomic status is present for older Europeans (Bago d'Uva et al., 2008a), Germany (Jürges, 2008) and France (Etilé & Milcent, 2006). Often people of higher socioeconomic groups have a higher reference health state and therefore respond more negative to a certain level of objective health, compared to the response of someone from a lower socioeconomic group. More specifically, different income groups show heterogeneous reporting in China, Indonesia and India (Bago d'Uva et al., 2008b). Reporting heterogeneity is also present in SAH responses for different genders and age groups (Lindeboom & Doorslaer, 2004; Peracchi & Rossetti, 2012). Chandola and Jenkinson (2000) find no reporting heterogeneity for ethnic groups when the outcomes are compared to a set of objective health measures. Overall, socioeconomic status seems to bias SAH results that make comparisons across socioeconomic status. This leads to a change in predictive power of SAH for mortality, e.g. by educational groups (Huisman et al., 2007). Jürges et al. (2008) argue that different measures of SAH are not directly comparable. These differences can occur in the phrasing of the SAH question or in the answer possibilities, e.g. the difference between the SAH answers in the WHO survey (very good, good, fair, bad, or very bad) and the US version (excellent, very good, good, fair, or poor)

Besides within country comparisons, there also is research done on between country differences. One way in which the SAH question differs across countries, is the phrasing. Sometimes exact translations are hard to find. However, the difference in the questions themselves is not the main point of this thesis. Assuming that the question is phrased exactly the same, there might still be a difference in the way that people in certain countries respond to SAH questions. Reference levels for different cultures might differ and reporting heterogeneity is then an important factor to take into account when comparing SAH for different countries. This thesis focuses on the subject of comparability of SAH across countries.

There exists literature that examines this comparability of SAH across countries. These papers often use one wave of panel data (e.g. Jürges, 2007), or combine multiple waves into one analysis (e.g. Kok et al., 2012). However, these papers do not include comparisons of results across multiple waves. Two waves of data might hold different results as to comparability of SAH. The central question for this research is the following:

Is self-assessed health comparable across countries and are the results the same across different waves of panel data?

To get insight into this comparability of SAH, the first thing that will be looked at is existing research. This will shine light upon different ways to answer the first part of the research question. After that, there will be an analysis of two waves of the Survey of Health, Ageing and Retirement in Europe (SHARE), which shows a comparison of results for both waves. Anchoring vignettes are used to measure possible reporting heterogeneity. These vignettes consists of one or multiple hypothetical health states that respondents have to rate in the same way they rate their own health. Section 3 elaborates on the SHARE dataset and how the vignettes exactly work. Furthermore, the HOPIT model uses these vignettes to analyse SAH and correct for reporting heterogeneity. The following sections of this thesis elaborate on these research steps.

2. Literature

2.1 Reporting heterogeneity between countries

Differences in response scales, or reporting heterogeneity, between countries in general surveys is present. De Jong et al. (2012) find that Russian people are more prone to report the extreme ends of the response categories (strongly agree/strongly disagree) and that Asian responses, except for Japan, show acquiescence bias, which means that they simply agree with a lot of stated questions. This gives insight into the way in which different cultures approach survey questions. One example of a survey question is self-reported emotional problems, which shows reporting heterogeneity when Britain and Norway are compared (Heiervang et al., 2008). Respondents in Norway are underreporting their emotional problems. It follows that one should make direct cultural comparisons of SAH with caution, because SAH can be sensitive to cultural environment (Jylhä et al., 1998). Hardy et al. (2014) show that respondents use both personal and cultural references when answering SAH questions.

Kok et al. (2012) find reporting heterogeneity across 10 European countries in self-assessed depression, but this reporting heterogeneity does not explain the differences for self-assessed depression across these countries. They use anchoring vignettes from SHARE data to identify differences in reporting scales, controlling for age, gender and education. Furthermore, a HOPIT model was estimated to correct self-assessed depression measures for reporting heterogeneity. This thesis will also focus on the vignette approach, combined with the HOPIT model.

Angelini et al. (2014) also find reporting heterogeneity in life satisfaction questions in European countries from SHARE data. They also use vignettes along with the HOPIT model to analyse differences in reporting scales in life satisfaction. Next to age, gender and education, their model consists of control variables for income, employment, objective health measures (number of chronic diseases, arthritis, symptoms, limitations with mobility, limitations with activities of daily living obesity and having been diagnosed with depression by a doctor) and social relationships (marital status, family bonds and extra-familiar activities). Angelini et al. (2012) find that self-reported working disability is also prone to reporting heterogeneity when comparing European countries from SHARE data. They also use a HOPIT model with vignettes and control for a large number of variables, including socioeconomic measures and health indicators. Kapteyn et al. (2007) conducted their own random experiment with vignettes, using internet panels from the Netherlands and the US. They find that when comparing the Netherlands to the US there also exist reporting heterogeneity in self-reported working disability. Salomon et al. (2004) show that reporting style differences in self-reported expectations for mobility are present for 6 Asian countries, using data from the World Health Survey, which also includes vignette data. They, however, do not follow any statistical approach to make a case for their results, but simply show distributions of vignette answers. Jürges (2007) also finds reporting style differences in self-reported general health in Europe using SHARE data and concludes that cross-country differences in SAH are overestimated when reporting styles are not taken into account. Using an ordered probit model, as well as an extensive list of chronic conditions and physical health measures, he estimates a health index for each respondent to summarize their health in one number, with near-death (0) and perfect health (1) as the end-points of the index. He compares this health index with self-reported general health and corrects for reporting style differences using a generalised ordered probit model.

If reporting heterogeneity is not accounted for in an analysis of SAH, then SAH results for different groups of respondents are not directly comparable. Two groups of individuals might have a significant different SAH, but this difference in SAH might exist because these two groups have different response scales. This means that reporting heterogeneity might cause differences in SAH between groups, that do not reflect true differences in objective health states.

2.2 Comparing multiple waves of panel data

When analysing reporting heterogeneity between countries, authors often choose to use one analysis for one or more waves of panel data. Therefore these papers do not look into possible differences between these waves of data. One could argue that respondent's references for SAH questions are not constant, but are prone to changes over time. It is possible that references for health in certain cultures or countries change over time. This could for instance be the case if the stigma for a particular disease becomes less, e.g. anti-stigma campaigns for mental illnesses (Byrne, 2000). This could have an effect on the mechanisms of reporting heterogeneity. Furthermore, with the rise of communication channels, such as the internet, people have the chance to become more educated about their own health, which over time could change people's references for SAH. These factors could all contribute and change reporting heterogeneity over a certain period of time, possibly also when comparing multiple waves of data.

3. Data

3.1 SHARE

Data used in this thesis will be the first and second wave of the Survey of Health, Ageing and Retirement in Europe (SHARE) data. SHARE is a cross-country panel survey, which follows a representative population which is aged 50+ along with their spouses. Among other variables, SHARE contains a general questionnaire, which includes health valuations of the respondents. SHARE also provides extensive information about socioeconomic variables, such as income and education. This thesis will make use of a specific part of the SHARE dataset, which is a drop-off questionnaire that includes SAH questions. The first wave of data contains data for Germany, Sweden, The Netherlands, France, Italy, Spain, Greece and Belgium. The second wave adds three countries to the sample, which are Denmark, Poland and Czech Republic. Data for the first wave was collected between 2004-2005 and between 2006-2007 for wave 2. The total sample which is used in this thesis contains 11656 observations (4286 for wave 1, 7370 for wave 2).

Table 1 Descriptives of variables. Bold indicates wave 2

Country			*Female*		*Age*				*Years of education*			
	N		*%*		*Mean*	*SD*	***Mean***	***SD***	*Mean*	*SD*	***Mean***	***SD***
Germany	488	**1127**	0.57	**0.54**	63.32	9.09	**64.65**	**9.26**	13.14	3.03	**12.51**	**3.33**
Sweden	376	**460**	0.52	**0.54**	63.51	9.31	**65.68**	**9.99**	10.57	3.23	**11.28**	**3.77**
Netherlands	506	**513**	0.52	**0.53**	62.32	9.43	**61.69**	**9.87**	11.57	3.43	**11.38**	**3.93**
Spain	456	**486**	0.58	**0.55**	64.61	10.51	**63.71**	**10.13**	7.02	4.28	**7.85**	**5.00**
Italy	418	**668**	0.56	**0.54**	63.42	9.36	**64.93**	**8.85**	7.22	4.22	**8.27**	**4.43**
France	797	**356**	0.56	**0.56**	64.36	10.27	**63.65**	**9.91**	8.43	5.49	**11.95**	**4.16**
Greece	708	**531**	0.54	**0.55**	61.76	10.57	**64.03**	**10.79**	9.70	4.46	**8.85**	**4.21**
Belgium	537	**843**	0.56	**0.54**	63.47	9.65	**64.96**	**9.99**	10.29	3.80	**11.71**	**3.55**
Denmark	-	**963**	-	**0.56**	-	-	**63.96**	**9.78**	-	-	**13.16**	**3.43**
Czechia	-	**894**	-	**0.60**	-	-	**64.14**	**10.01**	-	-	**11.45**	**3.12**
Poland	-	**529**	-	**0.57**	-	-	**62.78**	**9.76**	-	-	**9.27**	**3.03**
Total	*4286*	***7370***	*0.55*	***0.55***	*63.32*	*9.91*	***64.13***	***9.84***	*9.70*	*4.61*	***11.00***	***4.12***

3.2 Self-assessed health

The variable of interest is self-assessed health. This thesis will look into reporting heterogeneity of this SAH variable. In this dataset, six health domains of SAH are used, which are *pain, sleeping problems, mobility, memory, breathing problems* and *depression*. In this thesis these six domains are assumed to represent SAH, which is somewhat different to one general SAH question, e.g. the one mentioned in the introduction. Respondents are asked to what degree they have problems within each of the domains, e.g. for the domain of pain the question asked would be "Overall in the last 30 days, how much of bodily aches or pains did you have?" (SHARE, 2006). Respondents would answer these questions with the following answer categories: (1) none, (2) mild, (3) moderate, (4) severe or (5) extreme (see appendix for all SAH questions). Figure 1 shows the distribution of SAH for the domain of pain across the countries (see appendix for all domains). Some interesting differences between countries can be seen in figure 1. Poland has a relatively large proportion of respondents with moderate or more problems with pain, while the Netherlands and Sweden have a relatively high proportion of respondents with no or mild problems with pain.

Figure 1 Distribution of SAH in the domain of pain for both waves, by country

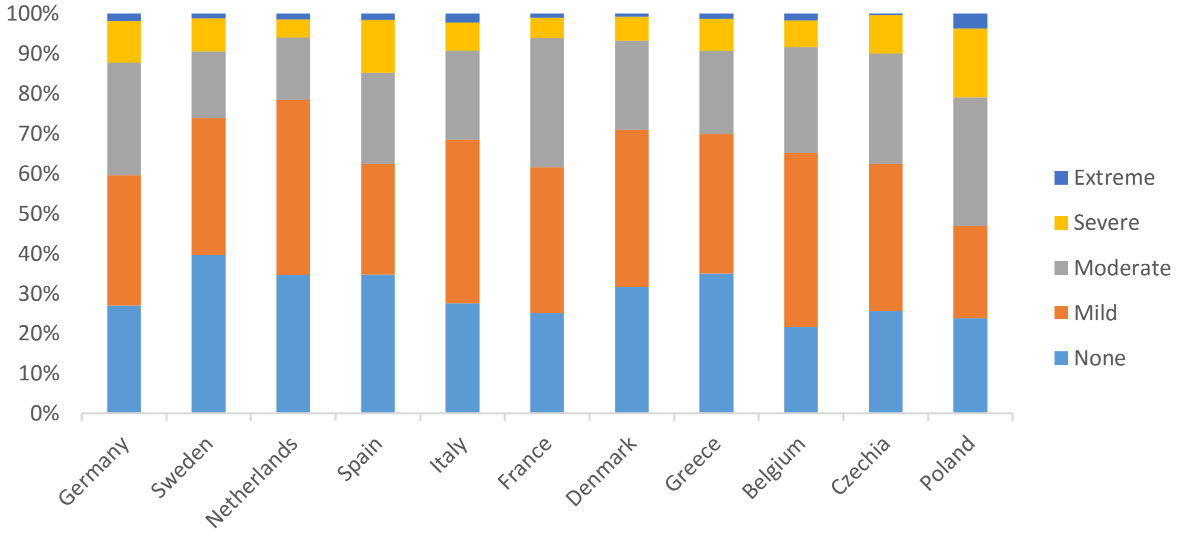
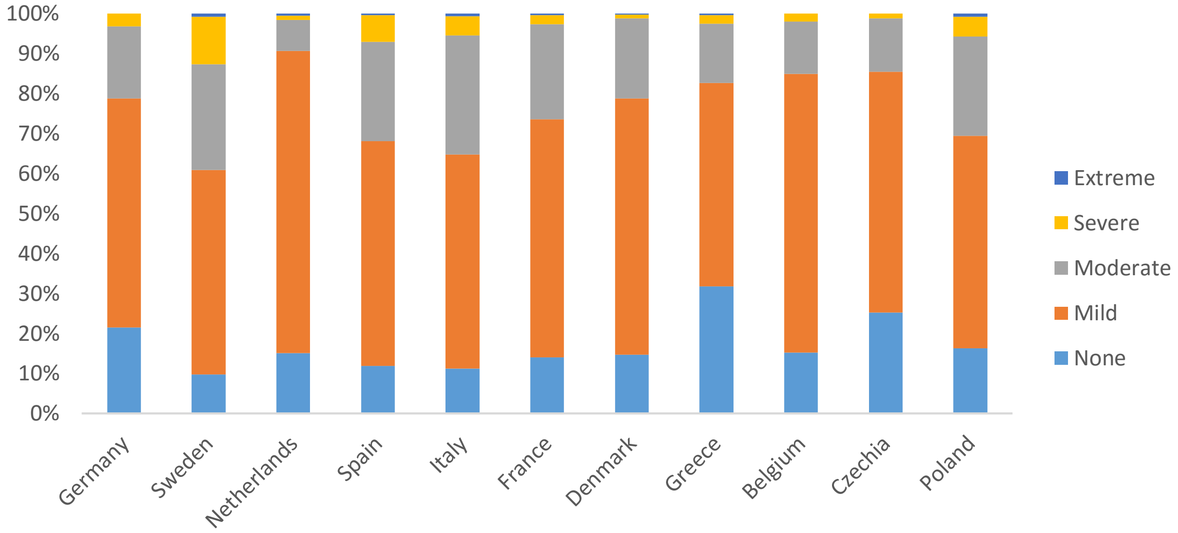


Figure 2 Distribution of vignette responses in the domain of pain for both waves, by country



3.2 Vignettes

One way of looking into reporting heterogeneity is using the vignette approach. After respondents answered the SAH questions, they were given vignette questions. These are present, besides the respondent's health questions, in the first two waves of SHARE data. A vignette consists of a hypothetical health state that respondents have to rate. Respondents are asked to evaluate these health states and assume the hypothetical person has the same age and background as the respondent. Vignettes are present for the same six health domains, as is the case for the SAH measures. The vignette used for the domain of pain is "Paul has a headache once a month that is relieved after taking a pill. During the headache he can carry on with his day-to-day affairs.", for which the respondent is asked "In your opinion, how much of bodily aches or pains does Paul have?". Respondents are then given the same five answer categories as in the SAH questions (see appendix for all vignettes). Names of the hypothetical persons are different for some countries, to match local common names. Since every respondent has to categorise the same health state, differences in responses to these vignettes could help identify reporting heterogeneity. Certain groups of people might rate the same health state differently.

With a vignette it is also possible to anchor the respondent's responses for that particular health domain and compute their SAH that is adjusted for reporting heterogeneity. Section 4 elaborates on how this is exactly done. There is, however, a difference between the availability of the vignettes for the two waves of SHARE data. Wave 1 has three vignettes for each respondent per health domain, while wave 2 only has one vignette. The one vignette that is used in wave 2 is one of the vignettes from wave 1. In this thesis only this one vignette per health domain is used. Respondents with missing values for either the SAH or the vignettes were left out of the analysis. Figure 2 shows the distribution of vignette responses for the domain of pain (see appendix for all domains). From figure 2 it is already noticeable that there might exist differences in responses to the vignette. The Netherlands for example has a large proportion of its respondents claiming the hypothetical person from the vignette has no or moderate problems with pain, compared to countries such as Sweden, Italy and Spain. Assuming that respondents use the same scale for these vignettes as for themselves, these differences in responses to the vignettes indicate reporting heterogeneity in SAH between countries, i.e. differences in figure 1 might not be differences in only true health.

3.3 Control Variables

Years of education is entered in the analysis as a continuous variable. This is done simply because this leads to a simple interpretation of the results. To check for any misspecifications in the models, link tests were done and showed no evidence for misspecification for the chosen variables. These link tests regress the dependent variable of a model on the prediction of the model and this prediction squared. When the latter has no explanatory power, there is no evidence for misspecification in the model. Years of education are directly asked to the respondents in wave 2, but not in wave 1. In the case of wave 1, years of education are derived from 1997 International Standard Classification of Education (ISCED-97) codes. A cross table of years of education in both waves for respondents that were in both waves, showed that the derivation for years of education in wave 1 was fairly accurate. However, there was, for some years of education, a slight underestimation in the derivations from the ISCED-97 codes compared to the years of education that were directly asked to the respondent.

As another proxy for socioeconomic status, household income was left out. The reason for this is that adding household income to the analysis led to a rather large number of missing values. To check the impact of leaving out household income as a variable, two separate analysis were done: one with household income added, and one without household income. This showed that leaving out household income had little effect on the significance for the rest of the coefficients. Also, the coefficient for household income was insignificant.

Age was divided into four age categories: <56, 56-65, 66-75 and >75. These age groups each have a share ranging from 15%-38%.[1] The effect of age on SAH or on reporting heterogeneity is most likely not linear, which often causes researchers to add age squared or age dummies to their models. The reason for choosing age as dummies in this thesis, is to avoid difficult interpretations compared with coefficients of age squared.

[1] Since this is not an equal distribution, the group with the largest relative share (age 56-65) was split into two groups and the analysis was repeated to see if this has an impact on the conclusions. There were no major changes in the results so the original distribution of age dummies is maintained.

4. Methods

To analyse the data and answer the research question, the methods in this thesis will consist of two parts. The first part is to look at reporting heterogeneity across countries. The second corrects SAH for reporting heterogeneity.

4.1 Reporting heterogeneity between countries

Looking into reporting heterogeneity across countries, this thesis uses an ordered probit model. The vignettes are entered as dependent variables, so there will be an estimation of six models. Since the vignettes are ordinal variables, an ordered probit model is an appropriate statistical approach. The ordered probit model assumes an underlying latent variable for the dependent variable, measured on a continuous scale. Then this model computes cut-points on the latent scale, to define which value of the latent variable corresponds with a certain value of the ordinal dependent variable. These cut-points are the same for all respondents, so this model assumes no reporting heterogeneity in the dependent variable. However, the dependent variables are the vignettes, which means that adding independent variables to the model already gives an insight into possible reporting heterogeneity for SAH, since all respondents answer the same vignette.

To show this model, let y_{ij}^{v} be the response to vignette in health domain *j* for respondent *i* and assume that the latent variable Y_{ij}^{v*} is the underlying factor for y_{ij}^{v}. More specifically

$$Y_{ij}^{v*} = X_i\beta + \varepsilon_i, \qquad \varepsilon_i|X_i \sim N(0,1) \tag{1}$$

where X_i is a vector of covariates. The variance of the error term, ε_i, is set to 1, since this allows for the ordered probit model to be estimated. Using the latent variable to identify the observed categorical response gives

$$y_{ij}^{v} = k \Leftrightarrow \tau^{k-1} \leq Y_{ij}^{v*} < \tau^{k} \tag{2}$$

with $k = 0, \dots, K, \tau^0 < \tau^1 < \cdots < \tau^{K-1} < \tau^K$, and $\tau^0 = -\infty, \tau^K = \infty$. The cut-points τ^k are constant for all respondents.

The ordered probit models for this part will consist of the six vignettes as dependent variables: one for each health domain. This model will include socioeconomic variables as control variables. These socioeconomic variables are gender, age and education. By adding country dummies, next to control variables, to the model, an estimation can be made to what degree respondents from different countries respond differently to the same vignette. This analysis is done for both waves. Because of the fact that wave 2 only has one vignette in each health domain, the models for wave 1 are also only done for the same vignette.

To facilitate the comparison of results between both waves of data, an additional analysis is done where one ordered probit model is made for each wave, with a general vignette as the dependent variable. To do this the dataset is transformed into a long format with six observations, one for each vignette domain, for each respondents. One general vignette is made that consists of the responses to the vignettes. The vignette domains are then added to the model as covariates. This gives a general overview of the various effects of the variables on vignette responses as a whole.

4.2 Correcting SAH for reporting heterogeneity

The ordered probit model in the previous section is helpful in getting an insight into reporting heterogeneity in SAH. The next step is to adjust the SAH for reporting heterogeneity. This is done with the Hierarchical Ordered Probit (HOPIT) model. This model is based on the model by Bago d'Uva

(2012). The dependent variables, in this case, are the SAH questions. So, again, there are six model estimated. The HOPIT model adjusts the cut-points for the latent variable with the help of the answers to the vignettes for the health domain that corresponds to the same domain of SAH. This means that the cut-points are now specific to a single respondent, but the values of the latent variable are the same for all respondents. Then, much like the ordered probit model, the cut-points are used to translate the latent variable into the answer categories, in this case the SAH categories. This method makes it possible to compare SAH that is corrected for reporting heterogeneity, across different groups of people, e.g. countries.

The HOPIT model first estimates reporting behaviour with the vignettes. Using the same notation as equation 1, the vignette is now specified as

$$Y_{ij}^{v*} = \alpha_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0,1) \tag{3}$$

where the only variation in the perception of the vignette is assumed random. The relationship between the observed vignette response and the latent health level of the vignette is defined as

$$y_{ij}^{v} = k \Leftrightarrow \tau_{ij}^{k-1} \leq Y_{ij}^{v*} < \tau_{ij}^{k} \tag{4}$$

with $k = 0, \dots, K, \tau^0 < \tau^1 < \dots < \tau^{K-1} < \tau^K$, and $\tau^0 = -\infty, \tau^K = \infty$. Assuming that all respondents perceive the vignette the same way leads to the conclusion that the only factor determining differences in vignette responses are the cut-points. These are defined as

$$\tau_{ij}^{k} = X_i \gamma^k \tag{5}$$

where γ^k is a vector of coefficients that need to be estimated. The HOPIT model allows the cut-points to shift non-parallel with respect to one another, whereas the ordered probit model only allows for parallel cut-point shifts. This means that the HOPIT model allows for variation in the relative distance between the cut-points on the latent health scale, which makes the model capable of correcting for stronger or weaker reporting heterogeneity in some levels of latent health. The next part of the HOPIT estimation is using these individual cut-points in the estimation of SAH. The latent level of a respondent's own health is defined as

$$Y_{ij}^{s*} = Z_i \beta + \varepsilon_i, \qquad \varepsilon_i | Z_i \sim N(0, \sigma^2) \tag{6}$$

where Z_i is a vector of covariates now including a constant. The observed SAH responses y_{ij}^{s} is determined by

$$y_{ij}^{s} = k \Leftrightarrow \tau_{ij}^{k-1} \leq Y_{ij}^{s*} < \tau_{ij}^{k} \tag{7}$$

with $k = 0, \dots, K, \tau^0 < \tau^1 < \dots < \tau^{K-1} < \tau^K$, and $\tau^0 = -\infty, \tau^K = \infty$ and where τ_{ij}^{k} is defined in equation 5.

The coefficients of the HOPIT model cannot be interpreted directly, only the sign has direct information. Therefore, an estimation of partial effects is necessary. Partial effects show changes in probabilities for a certain outcome that a certain variables initiates, which are expressed as changes in percentage points (p.p.). This requires a reference individual, which in this case is a male aged under 56, has 10 years of education and is from Germany. Since there are six health domains that are being investigated, there will be six models for each wave of SHARE data. To see what influence the correction for reporting heterogeneity has on the results, a comparison is also made between the HOPIT partial effects, and the partial effects of ordered probit models with the SAH questions as dependent variables. These SAH ordered probit models follow the same principle as discussed in equation 1 and 2, with the only difference being that Y_{ij}^{v*} and y_{ij}^{v} are replaced by Y_{ij}^{s*} and y_{ij}^{s} respectively. Furthermore, it is also possible to look at the coefficients for certain cut-points. If coefficients of covariates are significant in equation 5, this would mean there is reporting heterogeneity.

With these SAH HOPIT models that adjust for reporting heterogeneity, it is now also possible to compare results for both waves of data. To make comparisons more easy, data for Denmark, Poland and Czech Republic is omitted from this part of the analysis, since data for these countries only exist in wave 2. Also, for each domain the same single vignette is used in wave 1, as the one vignette that is present in wave 2. The models for both waves consist of the same variables, which makes comparing the results easy.

5. Results

This section of this thesis describes the results that were obtained using the models explained in section 4. These results will help in formulating answers to the research question. The first part goes in debt about reporting heterogeneity in SAH. The second part consists of the HOPIT models in which SAH is corrected for reporting heterogeneity. The final part of the results comprises a comparison between the results for both waves of SHARE data.

Table 2 Ordered probit models with the six health vignettes as dependent variables

	Pain Vignette		Sleep Vignette		Mobility Vignette	
	Wave 1	Wave 2	Wave 1	Wave 2	Wave 1	Wave 2
Female	-0.0433	-0.0311	-0.0425	-0.0564*	-0.0379	-0.116***
Age 56-65	-0.0613	0.0578	-0.0695	0.0622	-0.0222	-0.0369
Age 66-75	-0.00906	0.113**	-0.144**	0.0518	-0.0251	-0.0211
Age 76	-0.00674	0.139**	-0.111	0.0894*	-0.0327	-0.0359
Years of Education	-0.00129	-0.0130***	-0.00770*	0.00823*	-0.00169	0.00828*
Sweden	1.052***	0.162*	0.383***	0.332***	-0.519***	0.0891
Netherlands	-0.0707	-0.150**	-0.145*	-0.0401	-0.202**	-0.237***
Spain	0.409***	0.264***	-0.256***	-0.0257	0.167*	0.139*
Italy	0.497***	0.317***	-0.441***	-0.0603	-0.640***	-0.377***
France	0.195**	0.134*	-0.473***	-0.11	-0.278***	-0.210**
Greece	-0.375***	-0.181**	0.495***	0.147*	-0.268***	-0.605***
Belgium	0.0772	-0.078	-0.311***	-0.0121	-0.505***	-0.523***
Denmark	-	0.135**	-	-0.0832	-	-0.253***
Czechia	-	-0.189***	-	-0.309***	-	-0.307***
Poland	-	0.267***	-	0.181**	-	0.119

	Memory Vignette		Breath Vignette		Depression Vignette	
	Wave 1	Wave 2	Wave 1	Wave 2	Wave 1	Wave 2
Female	-0.0314	-0.122***	-0.108**	-0.181***	-0.107**	-0.134***
Age 56-65	0.000767	0.0823*	-0.00805	0.0239	-0.071	0.0305
Age 66-75	0.136**	0.159***	-0.0227	0.0677	-0.0394	0.07
Age 76	0.095	0.226***	-0.0281	0.0453	0.0316	0.0481
Years of Education	-0.0157***	-0.0194***	-0.00553	0.00813*	-0.0035	-0.00326
Sweden	0.992***	0.00801	0.105	0.346***	0.0994	0.397***
Netherlands	-0.296***	-0.306***	-0.158*	-0.018	0.0938	0.011
Spain	0.318***	0.370***	0.459***	0.752***	0.269***	0.650***
Italy	-0.0722	-0.0388	-0.319***	-0.136*	0.161*	0.0605
France	0.0689	0.0954	-1.431***	0.0153	0.168**	0.053
Greece	-0.471***	-0.0153	-0.127*	-0.147*	0.677***	0.200**
Belgium	-0.127	-0.125*	-0.880***	-0.264***	0.0784	-0.197***
Denmark	-	-0.192***	-	0.0365	-	0.366***
Czechia	-	-0.172***	-	0.0184	-	-0.187***
Poland	-	0.607***	-	0.561***	-	0.211***

** p<0.10, **p<0.05, *** p<0.01*

5.1 A first look into reporting heterogeneity

Table 2 shows the coefficients from the ordered probit models with the vignettes as dependent variables for both waves. The signs and p-values can be interpreted directly from this table. The interpretation is as follows: a significant positive coefficient means that there is a higher probability when that variable increases to report more health problems for the same objective health state, since the dependent vignette variables are increasing in bad health (1=no problems, 5=extreme problems). What follows from this is that all significant coefficients from table 2 point to the fact that reporting heterogeneity is present, because every respondent got to evaluate the same hypothetical health state. The magnitude of the coefficients is not directly interpretable. This is because the coefficients of the model are based on a latent variable.

What can be concluded from table 2, is that, especially for the domains of pain and mobility, most country coefficients are significant. Also, in the domain of sleeping problems all the country dummies have significant coefficients. This means that respondents from those countries answer the vignette significantly different than respondents from Germany, which is the reference country. More precisely, respondents from the Netherlands, Belgium and Czech republic all tend to report less health problems than Germans in the health domains for which the coefficients are significant. Sweden, Spain and Poland have positive significant coefficients in at least four vignettes, meaning respondents from those countries report significantly more health problems than Germans, with the exception of the coefficient for Sweden in the wave 1 mobility domain and the coefficient for Spain in the wave 1 sleep domain. Italy has significant coefficients in four health domains: pain, sleep (only in wave 1) mobility and breathing problems. Respondents from Italy tend to report significantly more problems with pain than Germans. However, Italians also tend to significantly report less problems with sleeping (in wave 1), mobility and breathing than Germans do. It is interesting to see that the depression vignette has a relatively low number of significant country coefficients.

Age does not seem to be a good predictor of vignette responses in three health domains: mobility, breathing problems and depression. The only significant coefficients for age groups are in the health domains pain (in wave 2), sleep and memory. What is interesting about these age dummies is that older groups of respondents, report significantly more health problems in these three health domains, compared to the group aged <56. The significant coefficients in the models for gender show that females report less health problems. Years of education seems to have different influences on vignette answers in the different health domains. More years of education are linked to reporting less health problems in the domains of pain and memory, and to reporting more health problems in the domains of mobility and breathing problems.[2]

Although not the intention of this part of the analysis, it is already easy to see some big differences between the results of both waves. Especially the contrast between the significant coefficients for gender and years of education in wave 2 and the insignificant coefficients of the same variables in wave 1 is quite contradicting. To facilitate the comparing of both waves, there also was a model estimated (not shown here) of wave 2 with only data from countries that were present in wave 1. This model held very little changes with respect to the signs and significance of the coefficients.

[2] To look if the effect of years of education was the same for all countries, separate models were estimated that only looked at the effect of gender, age and education on vignette answers within each of the countries. This did not hold any major differences compared to the model shown in table 2.

Table 3 Ordered probit models with one general vignette

	Wave 1	*Wave 2*
Female	**-0.0601*****	***-0.117******
Age 56-65	**-0.0356***	*0.0205*
Age 66-75	-0.0165	***0.0406****
Age 76	-0.00945	***0.0980******
Years of Education	**-0.00563*****	*-3.9E-05*
Sweden	**0.325*****	***0.213******
Netherlands	**-0.124*****	***-0.118******
Spain	**0.215*****	***0.354******
Italy	**-0.145*****	*-0.0422*
France	**-0.305*****	*-0.00743*
Greece	0.000213	***-0.102******
Belgium	**-0.285*****	***-0.202******

** p<0.10, **p<0.05, *** p<0.01*

Table 3 shows the coefficients from the general vignette model. This table gives a general overview of the various influences that the covariates have on general vignette responses. The coefficients for the vignette dummies are not shown, since they give no useful information. A few variables show consistent signs and significance across both waves of data. Females tend to report less health problems than men. People from Sweden and Spain report more health problems than Germans, whereas respondents from the Netherlands and Belgium report less health problems than Germans. Results across both waves of data show some changes in significance, but no major sign changes occur.

5.2 Correcting SAH for reporting heterogeneity

Table 4 shows the estimated coefficients for the second cut-point taken from the SAH HOPIT models. This means that it is the cut-point between the answers no or mild problems, and moderate, severe or extreme problems. When coefficients in this table are significant, it means that the second cut-point for this variable differs from that of a reference, which implies reporting heterogeneity. Significant and positive coefficients for example show that this variable is associated with less inclination to reporting health problems. As an example it can be seen that the coefficient for the second cut-point is positive for women, for the domains in which it is significant. This means that women tend to report no or mild health problems for higher values of latent health than men. Since the latent health scale is increasing in bad health, this means that women are less likely to report health problems. The same reasoning can be applied to the other covariates. The conclusions that can be taken from these coefficients of the second cut-points, are not much different than the conclusions derived from the vignette models in table 2. Most of the significant coefficients have the same sign.

Whereas table 4 shows the first part of the HOPIT estimation, table 5 and 6 show the second step, which is the adjustment of SAH for reporting heterogeneity. This is expressed as the partial effects for the reference individual[3] for outcome=1,2 (no or mild problems) derived from the SAH ordered

[3] The reason for showing the partial effect of two outcomes, instead of just one outcome, is that choosing just one outcome resulted in errors, possibly due to a small number of observations in that outcome with respect to the chosen reference individual. The current outcome (=1,2) and reference individual (male, age<56, years of education=10, German) for the partial effects were chosen based on trial and error with changes in both the outcomes and the reference individual.

Table 4 Estimated coefficients of the second cut-point from the SAH HOPIT models for the six health domains

	Pain		Sleep		Mobility	
	Wave 1	*Wave 2*	Wave 1	*Wave 2*	Wave 1	*Wave 2*
Constant	**0.4622*** **	**0.4194*** **	**-0.8568*** **	-0.0719	**-0.5565*** **	**-0.3644*** **
Female	**0.0777* **	**0.0964** **	0.0616	**0.0632* **	**0.0681* **	**0.1301*** **
Age 56-65	0.0285	0.0147	0.0334	-0.0408	0.0549	0.0271
Age 66-75	-0.0434	0.0414	0.0622	**-0.0923* **	0.026	0.0793
Age >75	-0.0141	-0.1053	0.0288	**-0.1599*** **	0.0416	0.0418
Years of education	**0.0138*** **	**0.0232*** **	-0.0024	**-0.0109** **	0.0027	-0.003
Sweden	**-0.9066*** **	-0.0104	**-0.1718* **	**-0.2327*** **	**0.6608*** **	0.0218
Netherlands	**0.5198*** **	**0.4759*** **	0.1359	0.0113	**0.3684*** **	**0.3083*** **
Spain	**-0.3324*** **	**-0.1349* **	**0.2158** **	0.0488	**-0.2039** **	**-0.1435** **
Italy	**-0.3597*** **	**-0.1887*** **	**0.4877*** **	**0.1024* **	**0.7009*** **	**0.4406*** **
France	-0.0979	-0.027	**0.2673*** **	0.0005	**0.2525*** **	**0.1323* **
Greece	**0.3507*** **	**0.1813** **	**-0.3739*** **	-0.1014	**0.3275*** **	**0.5475*** **
Belgium	**0.2117** **	**0.3186*** **	**0.2247*** **	-0.0281	**0.6090*** **	**0.5802*** **

	Memory		Breathing		Depression	
	Wave 1	*Wave 2*	Wave 1	*Wave 2*	Wave 1	*Wave 2*
Constant	**0.3198*** **	**0.5369*** **	**-1.0882*** **	**-0.3034*** **	0.0014	0.0176
Female	0.0569	**0.1652*** **	**0.1174*** **	**0.1975*** **	**0.1329*** **	**0.1846*** **
Age 56-65	0.0134	-0.0148	0.009	**-0.0938* **	0.0685	-0.0071
Age 66-75	**-0.1228** **	**-0.0972* **	0.0049	**-0.1306** **	0.0324	-0.0583
Age >75	**-0.1147* **	**-0.2578*** **	0.0445	**-0.2291*** **	-0.0542	-0.0513
Years of education	**0.0215*** **	**0.0233*** **	0.0074	**-0.0123** **	0.0054	0.0044
Sweden	**-0.9999*** **	-0.0631	-0.0219	**-0.1342* **	0.0145	**-0.3130*** **
Netherlands	**0.7666*** **	**0.7169*** **	**0.3837*** **	-0.0937	-0.0248	0.0257
Spain	**-0.3582*** **	**-0.3973*** **	**-0.3909*** **	**-0.7179*** **	**-0.2700*** **	**-0.6868*** **
Italy	0.1001	-0.0048	**0.5479*** **	**0.1296** **	-0.0783	0.0105
France	-0.0123	-0.0755	**1.4836*** **	**-0.1769** **	**-0.2279*** **	-0.0789
Greece	**0.3918*** **	**-0.1667** **	**0.3065*** **	**0.1607** **	**-0.5016*** **	**-0.1810*** **
Belgium	**0.3536*** **	**0.3383*** **	**1.0574*** **	**0.2388*** **	-0.0401	**0.2800*** **

** p<0.10, **p<0.05, *** p<0.01*

probit models and the SAH HOPIT models. A partial effect means that it is the probability change of being in the best two health categories when some aspect of the reference individual changes. For instance looking at the partial effects from the SAH HOPIT models, it can be derived from table 5 and 6 that being a woman, compared to being a male, decreases the probability of being in the best two categories in all health domains[4] with breathing problems in wave 2 as an exception. Differences between the partial effects from the ordered probit model and the HOPIT model show that SAH is adjusted, using the estimated individual cut-points, i.e. it shows the impact of correcting for reporting heterogeneity. Conclusions drawn from table 4 and the other cut-points are used by the model to transform the SAH responses. Looking at the gender covariate, it can be seen in table 4 that women report less health problems, so SAH needs to be adjusted upwards, i.e. more health problems than reported. The same conclusion can be taken from table 5 and 6, since all partial effects in the HOPIT models are more negative than those of the ordered probit model. This means that adjusting for reporting heterogeneity leads to women being even less likely to be in the best two categories in all health domains, except breathing problems in wave 2.

[4] All partial effects are ceteris paribus.

Table 5 Wave 1 partial effects for outcome=1,2 of the SAH ordered probit and HOPIT models in the six health domains. The reference individual for these partial effects is a male, aged <56, German and has 10 years of education.

	Pain		Sleep		Mobility	
	Ordered Probit	HOPIT	Ordered Probit	HOPIT	Ordered Probit	HOPIT
Female	**-0.1090*** **	**-0.1243*** **	**-0.1240*** **	**-0.1263*** **	**-0.0449*** **	**-0.0533*** **
Age 56-65	-0.0126	**-0.0305* **	-0.0159	-0.027	**-0.0239* **	-0.0247
Age 66-75	**-0.0587*** **	**-0.0689*** **	**-0.0284* **	**-0.0564*** **	**-0.1322*** **	**-0.1401*** **
Age >75	**-0.1548*** **	**-0.1683*** **	**-0.0587*** **	**-0.0842*** **	**-0.3008*** **	**-0.3151*** **
Years of education	**0.0099*** **	**0.0096*** **	**0.0059*** **	**0.0066*** **	**0.0090*** **	**0.0101*** **
Sweden	**0.1860*** **	**0.2615*** **	**0.1388*** **	**0.1679*** **	0.0023	**-0.0875*** **
Netherlands	**0.1234*** **	**0.1201*** **	0.0256	0.0332	**0.0711*** **	**0.0648*** **
Spain	**0.0930*** **	**0.1525*** **	**0.0499** **	0.022	**0.0711*** **	**0.0933*** **
Italy	**0.0664*** **	**0.1506*** **	-0.0124	**-0.0893*** **	**0.0861*** **	-0.0008
France	**0.0618*** **	**0.0886*** **	-0.0201	**-0.0949*** **	**0.1286*** **	**0.1128*** **
Greece	**0.0954*** **	0.0298	**0.1063*** **	**0.1533*** **	**0.1387*** **	**0.1275*** **
Belgium	0.0372	**0.0457* **	**-0.0568*** **	**-0.0949*** **	**0.0718*** **	0.0119

	Memory		Breathing		Depression	
	Ordered Probit	HOPIT	Ordered Probit	HOPIT	Ordered Probit	HOPIT
Female	**-0.0338*** **	**-0.0401*** **	-0.0057	**-0.0249** **	**-0.1009*** **	**-0.1121*** **
Age 56-65	-0.0036	0.001	-0.0138	-0.0037	**0.0355*** **	**0.0323** **
Age 66-75	**-0.0750*** **	**-0.0442** **	**-0.0642*** **	**-0.0563*** **	**0.0293** **	**0.0316** **
Age >75	**-0.1609*** **	**-0.1357*** **	**-0.0875*** **	**-0.0805*** **	-0.0218	-0.0056
Years of education	**0.0082*** **	**0.0057*** **	**0.0031*** **	**0.0035*** **	**0.0067*** **	**0.0076*** **
Sweden	**0.0384** **	**0.1292*** **	**-0.1538*** **	**-0.1391*** **	**-0.1142*** **	**-0.0946*** **
Netherlands	**0.0282* **	-0.0139	**0.0373*** **	**0.0488** **	**0.0708*** **	**0.1185*** **
Spain	0.0077	**0.0512** **	**0.0547*** **	**0.0924*** **	0.0237	**0.0567** **
Italy	0.0125	-0.0008	**0.0484*** **	0.008	-0.0213	0.0048
France	0.0117	0.0267	-0.0002	**-0.3204*** **	**0.0416** **	**0.0649*** **
Greece	**0.0607*** **	-0.0182	**0.0299** **	**0.0475** **	**-0.0346* **	**0.0520** **
Belgium	-0.0119	-0.0353	0.0189	**-0.1486*** **	0.0136	**0.0399* **

** p<0.10, **p<0.05, *** p<0.01*

When comparing the partial effects from the ordered probit model and the HOPIT model, no sign changes can be seen. However, some differences in significance occur between partial effects from the HOPIT model and the ordered probit model. Especially the country dummies hold some corrections for reporting heterogeneity that result in significant or insignificant differences in probabilities compared to Germany. For example, Belgium has no significant difference in probabilities of being in the best two health categories for breathing problems in the ordered probit model. However, correcting for reporting heterogeneity shows that respondents from Belgium now have a lower probability to be in this group compared to respondents from Germany, as can be seen in the partial effects from the HOPIT model. The opposite happens for Belgium in the domain of mobility. A significant partial effect in the probit model, is an insignificant partial effect in the HOPIT model, which implies that correcting for reporting heterogeneity results in no significant difference between Belgium and Germany in the domain of mobility.

5.3 Comparing HOPIT results between waves

Looking at the partial effects from the SAH HOPIT models from table 5 and 6 and comparing wave 1 and wave 2, shows that the socioeconomic variables age and years of education have little surprises.

Table 6 Wave 2 partial effects for outcome=1,2 of the SAH ordered probit and HOPIT models in the six health domains with the same reference individual.

	Pain		Sleep		Mobility	
	Ordered Probit	HOPIT	Ordered Probit	HOPIT	Ordered Probit	HOPIT
Female	**-0.0835*****	**-0.0995*****	**-0.0996*****	**-0.1092*****	**-0.0442*****	**-0.0660*****
Age 56-65	**-0.0460*****	**-0.0476****	**-0.0238***	-0.0135	**-0.0305****	**-0.0306***
Age 66-75	**-0.0940*****	**-0.0903*****	**-0.0562*****	**-0.0453*****	**-0.1024*****	**-0.1013*****
Age >75	**-0.1705*****	**-0.1485*****	**-0.0570*****	**-0.0398***	**-0.2330*****	**-0.2239*****
Years of education	**0.0129*****	**0.0112*****	**0.0056*****	**0.0081*****	**0.0130*****	**0.0145*****
Sweden	**0.0636*****	**0.1041*****	**0.0394****	**0.0750*****	**0.0758*****	**0.0766*****
Netherlands	**0.1013*****	**0.0716*****	**0.0622*****	**0.0811*****	**0.0556*****	**0.0344***
Spain	**0.0866*****	**0.1440*****	**0.0494*****	**0.0445****	**0.0915*****	**0.0943*****
Italy	**0.1058*****	**0.1683*****	**0.0640*****	**0.0651*****	**0.1111*****	**0.0642*****
France	0.0247	0.0388	-0.0257	-0.039	**0.1172*****	**0.0830*****
Greece	**0.1159*****	**0.0936*****	0.0161	0.0213	**0.1281*****	**0.0485****
Belgium	0.0178	-0.0152	**-0.0542*****	**-0.0340***	**0.0610*****	-0.0166

	Memory		Breathing		Depression	
	Ordered Probit	HOPIT	Ordered Probit	HOPIT	Ordered Probit	HOPIT
Female	-0.0046	**-0.0310*****	0.0046	-0.0133	**-0.0653*****	**-0.0877*****
Age 56-65	-0.005	0.0036	-0.0074	-0.0021	**0.0200****	**0.0222***
Age 66-75	**-0.0877*****	**-0.0578*****	**-0.0638*****	**-0.0583*****	0.001	0.0103
Age >75	**-0.1731*****	**-0.0989*****	**-0.1238*****	**-0.1035*****	**-0.0451*****	**-0.0377****
Years of education	**0.0074*****	**0.0051*****	**0.0067*****	**0.0080*****	**0.0045*****	**0.0054*****
Sweden	-0.0006	-0.0014	0.0035	**0.0225***	0.0189	**0.0544*****
Netherlands	0.0042	**-0.0553*****	**0.0228****	**0.0514*****	**0.0318*****	**0.0476*****
Spain	-0.0072	**0.0498*****	**0.0431*****	**0.0766*****	**-0.0676*****	0.0109
Italy	-0.0112	-0.0144	**0.0555*****	**0.0461*****	**-0.0405*****	**-0.0317***
France	**-0.0387****	-0.0211	**-0.0384****	-0.0175	**-0.0284***	-0.0193
Greece	**-0.0259***	-0.0266	**-0.0453*****	**-0.0551*****	**-0.0236***	-0.0071
Belgium	**-0.0443*****	**-0.0786*****	-0.0084	**-0.0294****	**-0.0358*****	**-0.0594*****

** p<0.10, **p<0.05, *** p<0.01*

Education has significant and positive coefficients in all models, meaning that an extra year of education raises the probability of being in the best two categories of SAH. Furthermore, the negative relation between age and health seems to also apply in this analysis. The oldest age group has in all cases to highest negative significant coefficient, which means the oldest group has the lowest probability of being in the best two health categories compared to the younger age group. However, an interesting difference in the age categories is present in the depression domain. In this domain age groups 56-65 (both waves) and 66-75 (wave 1) have a significantly higher probability of being in the best two categories compared to the youngest group aged <56.

Comparing the waves in the HOPIT models shows not much of a difference in the socioeconomic variables. Some coefficients are significant in one wave, but not in the other. However, when this happens, the significant variable often has a relatively small partial effect, e.g. the coefficient for female in the breathing domain wave 1 model has a partial negative effect of 2.49 percentage points, while the partial effect of female in the pain wave 1 model is 12.43 percentage points.

Some interesting differences can be seen, when comparing partial effects of the country dummies for both waves in a certain domain of SAH. Interestingly, in the domain of pain all countries have a higher probability to be in the best two SAH categories, compared to Germany. France (8.86 p.p.), Greece (9.39 p.p) and Belgium (4.57 p.p.) have only one significant partial effect in both waves in the pain domain. The domain of sleeping problems is subject to some changes between waves. Sweden and Belgium show consistency across waves with significant partial effects that have the same sign. However, other country dummies have one insignificant partial effect. The biggest difference here is the partial effect of Greece, which goes from 15.33 p.p. in wave 1 to insignificant in wave 2. Italy is the most interesting case in the domain of sleep, because Italy has two significant partial effects, but their signs differ. In wave 1 Italians have 8.93 p.p. lower probability to be in the best two SAH categories for sleeping problems compared to Germans, which changes to a 6.51 p.p. higher probability in wave 2. This difference could be partly due to the answers to the vignettes. Looking at the cut-points from table 4, it can be seen that Italy has a significant positive ($p<0.01$) coefficient in wave 1 (Italians report less health problems compared to Germans in the same vignette) for sleeping problems, and a smaller less significant coefficient ($p<0.1$) for the second cut-point in wave 2. Considering that Italians report less health problems in wave 1, the sign difference for the partial effects in the HOPIT models is somewhat logical. This is because reporting less health problems, means that correcting for this causes the SAH to become worse, which is exactly what happens in wave 1 for Italy in the domain of sleeping problems, i.e. lower probability to be in the best SAH categories.

Sweden is a country that has three sign changes in the HOPIT models. This happens for the domains of mobility, breathing problems and depression. Also, in the domain of memory, the partial effect in wave 1 is significant and relatively large (12.92 p.p.) and insignificant in wave 2. Another sign change occurs for Belgium in the domain of depression. These sign changes for Sweden and Belgium follow the same reasoning as explained for the sign change of Italy in the domain of sleeping problems, which is that these changes in SAH partial effects can be partly derived from the changes in coefficients from the cut-points, shown for the second cut-point in table 4. The only sign change that does not follow this reasoning, is the change in signs for the partial effects of Greece in the domain of breathing problems. In this case the sign or significance of the coefficient for the cut-points in table 4 does not change across waves. Wave 1 holds a slightly lower second threshold in this case, which implies that Greece became less optimistic in wave 2. The expected correction for this is that SAH for wave 1 would be corrected more than for wave 2, i.e. wave 1 should have a more negative partial effect than in wave 2. However, the partial effects for Greece in breathing problems show the exact opposite. One could argue that since vignette responses for Greece are constant (same sign for the cut-points), the change in partial effects is related to the fact that Greek people really have more problems with breathing in wave 2 compared to Germans.

6. Discussion

6.1 Conclusion

Using anchoring vignettes, this thesis shows that SAH is not directly comparable across countries. Certain European countries have different reporting styles when it comes to SAH. The vignette models from table 2 and 4 show that differences exist in the way that people evaluate a certain (hypothetical) health state. This reporting heterogeneity between countries is present in all six health domains. Overall, people from the Netherlands, Belgium and the Czech Republic are less likely to report health problems than Germans, whereas the Spanish, Swedish and Polish tend to report more health problems. The domain of depression has the least significant coefficients for the country dummies. However, Kok et al. (2012) also use the vignette method to show that there are large differences in reporting thresholds in the domain of depression. They found these differences using Sweden as the reference country. Evidence from this thesis and other literature suggests that one should be cautious in drawing conclusions from SAH on its own. Policy makers should always bear in mind that subjective measures of health are prone to inter-personal and inter-cultural differences in perception of SAH questions and health states.

The vignette models also show, in wave 2, that females report significantly less health problems than men. This result is opposite of what Peracchi and Rossetti (2012). They found women reporting more health problems, also using vignettes. However, they employ different methods[5] compared to the methodology of this thesis, which is one of the possible reasons for the difference in results. Years of education show mixed results in the vignette models and leads to either reporting more health problems (mobility/breathing problems) or less health problems (memory). The association between education and reporting less problems with memory is in line with what Kok et al. (2012) found. Results from this thesis also show that there exists reporting heterogeneity between certain age groups.

Vignettes can also be used to correct for reporting heterogeneity and compute a corrected SAH model. This is done with the HOPIT model. Results from this method show that still differences exist in SAH when corrected for reporting heterogeneity. Women, people of old age and lower educated people have a lower probability to be in the best SAH categories. Furthermore, the HOPIT models also show that there are large differences between countries in corrected SAH. However, some countries are prone to changes of signs for probabilities to be in a certain SAH category when comparing two waves of data. This change in signs could have two reasons. One reason would be that real health has changed in comparison with the reference country (Germany). Another reason is that answers to the vignette have changed. The latter seems to be plausible, since the countries that show sign changes in the HOPIT models also show sign or significance changes in the vignette models. These results show that definitive conclusions from models using vignettes concerning differences between countries in SAH should be made with caution, if only one wave of panel data is considered. Perceptions and references of respondents concerning health might be subject to change over time. Therefore, future research is needed in the subject of changes in vignette perceptions.

[5] One of those differences is that Peracchi and Rossetti collapse the vignette answers to 3 categories: no problems, mild problems and serious problems. To see if this causes the difference in results, there was a separate analysis done with the same vignette categories. However, results from this analysis were the same as the models shown in this thesis.

Overall, in analysing SAH and reporting heterogeneity the vignette approach proves to be a useful methodology. However, vignettes can also be used outside the subject of SAH. Any application, in which research is being conducted on or with individuals having to categorize their own situations or that of others, is most likely subject to some sort of reporting bias. Vignettes offer an option at grasping how respondents differ in their response scales and this can be applied in many fields of research where perception is an important factor, e.g. general well-being, politics (King et al., 2004), economics or welfare. SAH corrected for reporting heterogeneity with anchoring vignettes offer a way of quantifying health, next to objective health measures. For instance, looking at table 5 and 6, it can be seen that Germany scores worse than a relatively large number of EU countries in the domain of pain and mobility, and to a lesser extent in other domains. It would therefore be wise for German policy makers to look at and learn from these other countries that do seem to handle health problems better in these health domains.

6.2 Limitations

Some choices were made with regards to transforming certain variables. One variable in particular has a transformation that is debatable, which is education. In this thesis the choice to use years of education was made to make interpretation easy. However, the first problem with this variable is that measurements were not the same in both waves of data and this could cause the results to be biased. The second problem is that years of education are not directly comparable across countries. One year in of education might have a different impact on real knowledge, depending on the country in which this education is given. Data was available to divide education into ISCED-97 categories. However, doing this resulted in extreme differences between countries, e.g. a category of education would exist largely of respondents from one country. Therefore, years of education was chosen as a control variable. This does not necessarily have a major impact on the results obtained for the differences in reporting heterogeneity between countries, which is the main subject of this thesis.

Furthermore, a limitation of this study would be that the data only allows for comparisons of people with age 50 or older. Differences between countries in reporting styles might differ for younger generations. Therefore further research is needed that investigates reporting heterogeneity in these younger generations. Also, there were a lot of missing values or observations in the total sample. SHARE offers a representative sample, but this cannot be said with certainty about the sub-sample that this thesis uses. For future research it is advisable to look into which respondents from what country have a lot of missing values. It could bias the results if only respondents from a certain background decide to not participate in the vignette questionnaire.

Finally, another limitation of this study, is that of the approach of the vignette. Like many methods, the vignette approach has its assumptions, which must hold in order for it to be a valid method. *Response consistency* assumes that respondents use the same scale when evaluating themselves and when evaluating the vignette person. Research suggests that for some health domains (mobility, concentration, breathing and affect) this assumption might not hold (Kapteyn et al., 2011). This would mean that corrections for reporting heterogeneity in these health domains could be biased and therefore these corrections could be underestimated or overestimated. *Vignette equivalence* assumes that different respondents interpret the same vignette in the same way. This assumption is also open to some discussion (Jürges & Winter, 2013), which would imply that reporting heterogeneity is not only present in SAH, but might also occur in the vignettes. This leads to the conclusion that correcting SAH for reporting heterogeneity with vignettes would not be justified, since this would not correct for reporting heterogeneity in an unbiased way. Further research is needed in the area of the assumptions concerning vignette questions. This would shine light upon different ways in which the vignettes can be improved, so that the probability increases that the assumptions hold.

References

Angelini, V., Cavapozzi, D., & Paccagnella, O. (2012). Cross-Country Differentials in Work Disability Reporting Among Older Europeans. *Social Indicators Research Vol. 105*(2), 211-226.

Angelini, V., Cavapozzi, D., Corazzini, L., & Paccagnella, O. (2014). Do Danes and Italians Rate Life Satisfaction in the Same Way? Using Vignettes to Correct for Individual-Specific Scale Biases. *Oxford Bulletin of Economics and Statistics Vol. 76*(5), 643–666.

Bago d'Uva, T. (2012). Chapter 4 Repairing Heterogeneity in Health. In A. M. Jones, N. Rice, T. Bago d'Uva, & S. Balia, *Applied Health Economics.* Londen: Routledge {ISBN: 9780415397728}.

Bago d'Uva, T., Doorslaer, E. v., Lindeboom, M., & O'Donnell, O. (2008b). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics Vol. 17*(3), 351-375.

Bago d'Uva, T., O'Donnell, O., & Van Doorslaer, E. (2008a). Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. *International Journal of Epidemiology Vol. 37*(6), 1375-1383.

Blom, A. G., & Korbmacher, J. M. (2013). Measuring Interviewer Characteristics Pertinent to Social Surveys: A Conceptual Framework. *Survey Methods: Insights from the Field*. Retrieved from http://surveyinsights.org/?p=817

Byrne, P. (2000). Stigma of mental illness and ways of diminishing it. *Advances in Psychiatric Treatment Vol. 6*(1), 66-72.

Chandola, T., & Jenkinson, C. (2000). Validating Self-rated Health in Different Ethnic Groups. *Ethnicity & Health Vol. 5*(2), 151-159.

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

De Jong, M. G., Lehmann, D. R., & Netzer, O. (2012). State-Dependence Effects in Surveys. *Marketing Science Vol. 31*(5), 838-854.

Etilé, F., & Milcent, C. (2006). Income-related reporting heterogeneity in self-assessed health: evidence from France. *Health Economics Vol. 15*(9), 965-981.

Hardy, M. A., Acciai, F., & Reyes, A. M. (2014). How Health Conditions Translate into Self-Ratings: A Comparative Study of Older Adults across Europe. *Journal of Health and Social Behavior Vol. 55*(3), 320-341.

Heiervang, E., Goodman, A., & Goodman, R. (2008). The Nordic advantage in child mental health: separating health differences from reporting style in a cross-cultural comparison of psychopathology. *Journal of Child Psychology and Psychiatry Vol. 49*(6), 678-685.

Hernandez-Quevedo, C., Jones, A. M., & Rice, N. (2004). Reporting bias and heterogeneity in self-assessed health. Evidence from the British Household Panel Survey. *Ecuity III Project Working Paper*, No.19.

Huisman, M., Lenthe, F. v., & Mackenbach, J. (2007). The predictive ability of self-assessed health for mortality in different educational groups. *International Journal of Epidemiology Vol. 36*(6), 1207-1213.

Idler, E. L., & Benyamini, Y. (1997). Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies. *Journal of Health and Social Behavior Vol. 38*(1), 21-37.

Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. *Health Economics Vol. 16*(2), 163-178.

Jürges, H. (2008). Self-assessed health, reference levels and mortality. *Applied Economics Vol. 40*(5), 569-582.

Jürges, H., & Winter, J. (2013). Are anchoring vignettes ratings sensitive to vignette age and sex? *Health Economics, Vol. 22*(1), 1-13.

Jürges, H., Avendano, M., & Mackenbach, J. P. (2008). Are different measures of self-rated health comparable? An assessment in five European countries. *European Journal of Epidemiology Vol. 23*(12), 773-781.

Jylhä, M., Guralnik, J. M., Ferrucci, L., Jokela, J., & Heikkinen, E. (1998). Is Self-Rated Health Comparable Across Cultures and Genders? *Journal of Gerontology: Social Sciences Vol. 53B*(3), S144-S152.

Kapteyn, A., Smith, J. P., & Van Soest, A. (2007). Vignettes and Self-Reports of Work Disability in the United States and the Netherlands. *American Economic Review Vol. 97*(1), 461-473.

Kapteyn, A., Smith, J. P., Soest, A. v., & Voňková, H. (2011). Anchoring Vignettes and Response Consistency. *Working Paper No. WR-840*, RAND, Santa Monica.

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 Vol. 98*(1), 191-207.

Kok, R., Avendano, M., Bago d'Uva, T., & Mackenbach, J. (2012). Can Reporting Heterogeneity Explain Differences in Depressive Symptoms Across Europe? *Social Indicators Research Vol. 105*(2), 191-210.

Krause, N. M., & Jay, G. M. (1994). What Do Global Self-Rated Health Items Measure? *Medical Care Vol. 32*(9), 930-942.

Lindeboom, M., & Doorslaer, E. v. (2004). Cut-point shift and index shift in self-reported health. *Journal of Health Economics Vol. 23*(6), 1083-1099.

Logan, D. E., Claar, R. L., & Scharff, L. (2008). Social desirability response bias and self-report of psychological distress in pediatric chronic pain patients. *Pain Vol. 136*(3), 366-372.

Lumsdaine, R. L., & Exterkate, A. (2013). How survey design affects self-assessed health responses in the Survey of Health, Ageing, and Retirement in Europe (SHARE). *European Economic Review Vol. 63*, 299-307.

Manderbacka, K. (1998). Examining what self-rated health question is understood to mean by respondents. *Scandinavian journal of social medicine Vol. 25*(2), 145-153.

McPhail, S., Beller, E., & Haines, T. (2010). Reference bias: presentation of extreme health states prior to eq-vas improves health-related quality of life scores. A randomised cross-over trial. *Health and Quality of Life Outcomes Vol. 8*(146).

Newell, S. A., Girgis, A., Sanson-Fisher, R. W., & Savolainen, N. J. (1999). The Accuracy of Self-Reported Health Behaviors and Risk Factors Relating to Cancer and Cardiovascular Disease in the General Population: A Critical Review. *American Journal of Preventive Medicine Vol. 17*(3), 211–229.

Peracchi, F., & Rossetti, C. (2012). Heterogeneity in health responses and anchoring vignettes. *Empirical Economics Vol. 42*(2), 513-538.

Salomon, J. A., Tandon, A., & Murray, C. J. (2004). Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes. *British Medical Journal Vol. 328*(7434), 258.

SHARE. (2006). *Questionnaire Wave 2.* Retrieved from http://www.share-project.org/data-access-documentation/questionnaires/questionnaire-wave-2.html

Appendix

8.1 SAH questions[6]

Pain

"Overall in the last 30 days, how much of bodily aches or pains did you have?"

Sleeping problems

"In the last 30 days, how much difficulty did you have with sleeping?"

Mobility

"Overall in the last 30 days, how much of a problem did you have with moving around?"

Memory

"Overall in the last 30 days how much difficulty did you have with concentrating or remembering things?"

Sleeping problems

"In the last 30 days, how much of a problem did you have because of shortness of breath?"

Depression

"Overall in the last 30 days, how much of a problem did you have with feeling sad, low, or depressed?"

8.2 Vignettes

Pain

"Paul has a headache once a month that is relieved after taking a pill. During the headache he can carry on with his day-to-day affairs. In your opinion, how much of bodily aches or pains does Paul have?"

Sleeping problems

"Alice falls asleep easily at night, but two nights a week she wakes up in the middle of the night and cannot go back to sleep for the rest of the night. In your opinion, how much difficulty does Alice have with sleeping?"

Mobility

"Rob 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 has no problems with day-to-day activities, such as carrying food from the market. In your opinion, how much of a problem does Rob have with moving around?"

[6] Source: SHARE, 2006

Memory

"Lisa can concentrate while watching TV, reading a magazine or playing a game of cards or chess. Once a week she forgets where her keys or glasses are, but finds them within five minutes. In your opinion, how much difficulty does Lisa have with concentrating or remembering things?"

Breathing problems

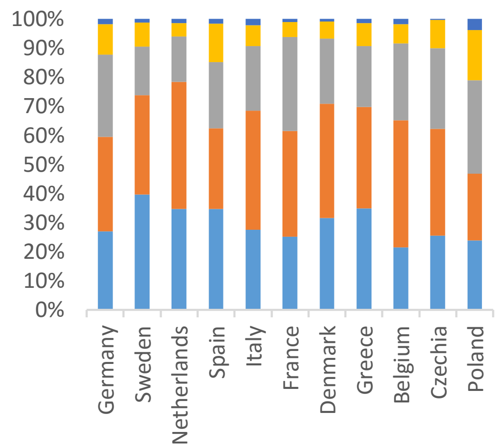
"Mark has no problems with walking slowly. He gets out of breath easily when climbing 20 meters uphill or a flight of stairs. How much of a problem does Mark have because of shortness of breath?"

Depression

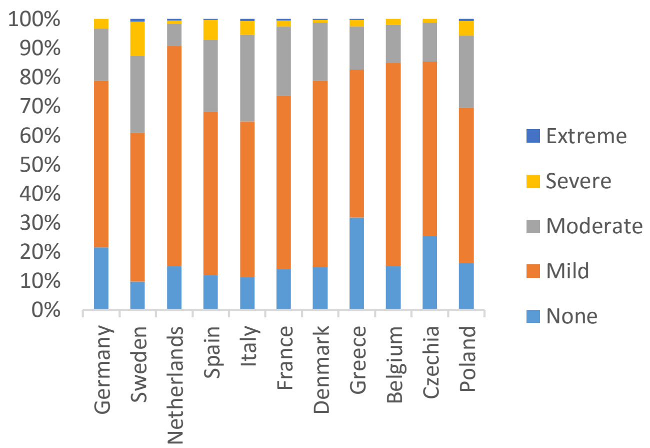
"Karen enjoys her work and social activities and is generally satisfied with her life. She gets depressed every 3 weeks for a day or two and loses interest in what she usually enjoys but is able to carry on with her day-to-day activities. How much of a problem does Karen have with feeling sad, low, or depressed?"

8.3 Distribution of SAH and vignette responses, by country

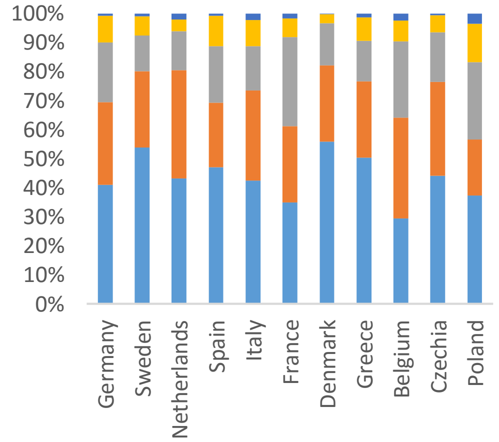
Pain Self Rating



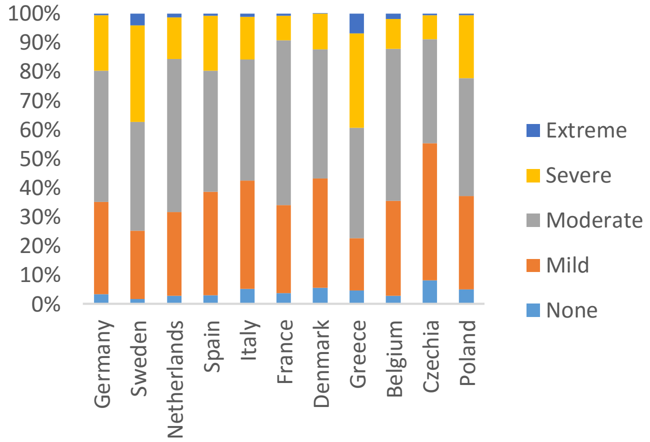
Pain Vignette



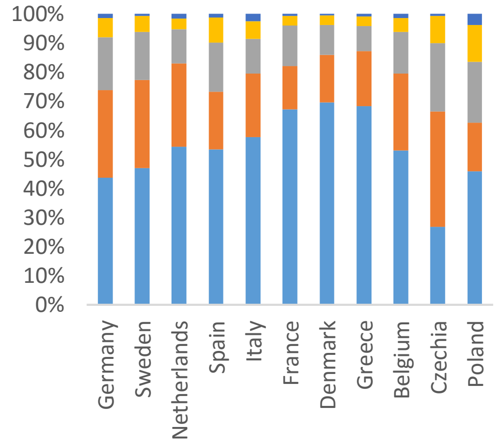
Sleep Self Rating



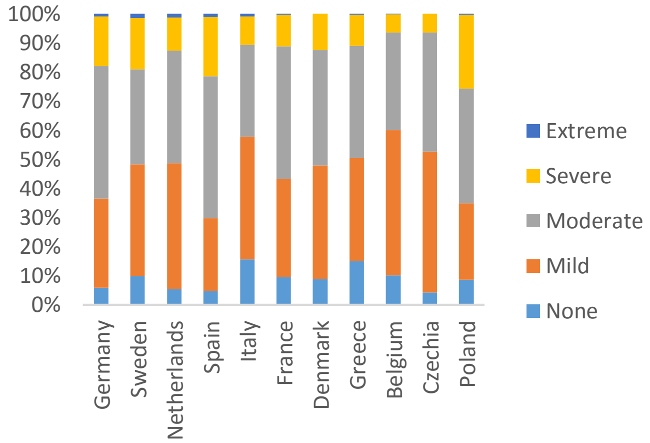
Sleep Vignette



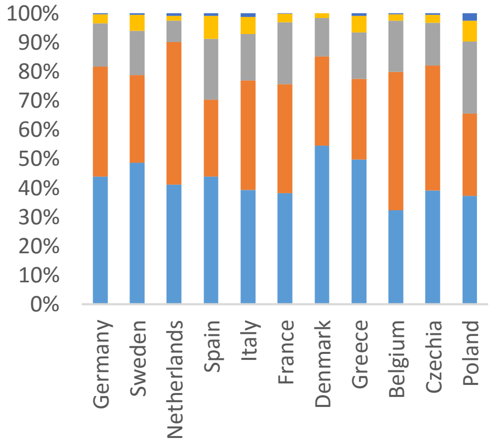
Mobility Self Rating



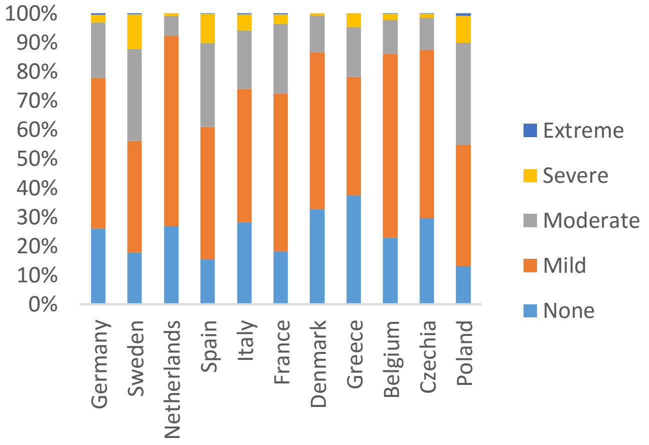
Mobility Vignette



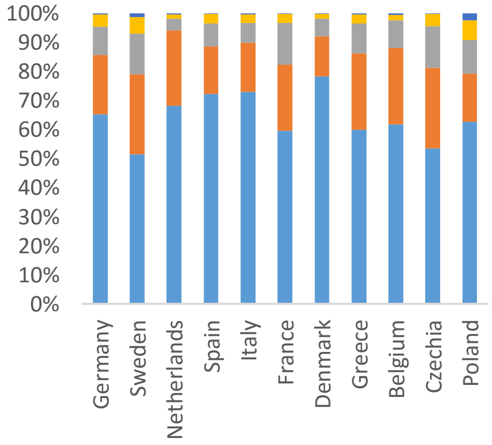
Memory Self Rating



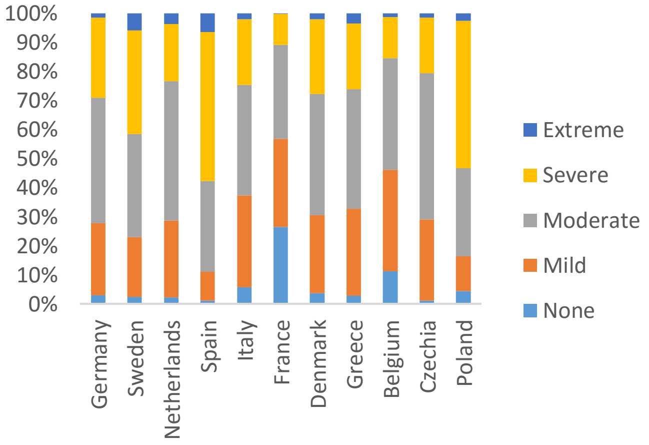
Memory Vignette



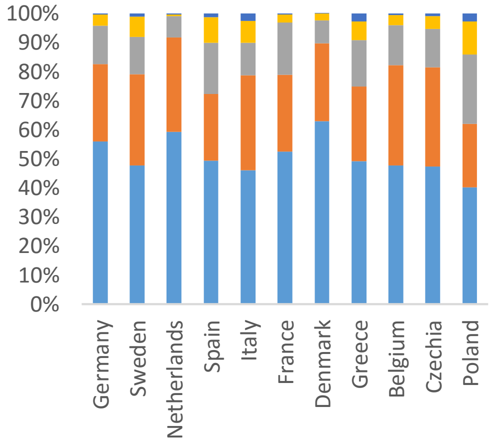
Breathing Self Rating



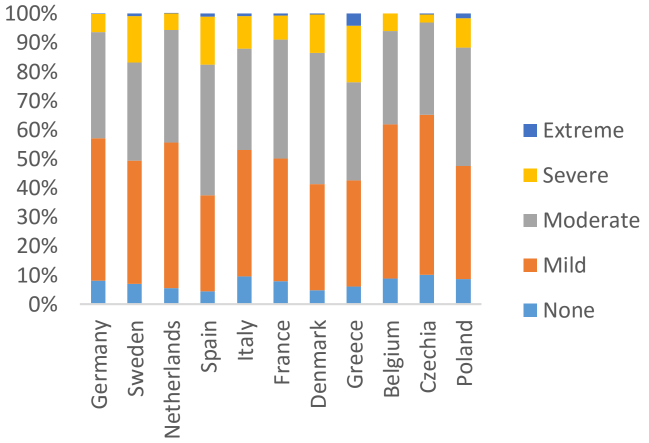
Breathing Vignette



Depression Self Rating



Depression Vignette


Germany
Sweden
Netherlands
Spain
Italy
France
Denmark
Greece
Belgium
Czechia
Poland
Extreme
Severe
Moderate
Mild
None