Netspar Network for Studies on Pensions, Aging and Retirement

Hana Voňková

Patrick Hullegie

Is the Anchoring Vignettes Method Sensitive to the Domain and Choice of the Vignette?

Discussion Paper 01/2010 - 004

January 14, 2010

Is the anchoring vignettes method sensitive to the domain and choice of the vignette?

Hana Voňková* and Patrick Hullegie†

January 14, 2010

Abstract

The answers to self-assessment questions depend on the objective situation as well as on the reporting behavior of the individual. Anchoring vignettes have been introduced as a tool to correct self-assessments for heterogeneity in reporting behavior. The aim of this paper is to study whether the vignette method is sensitive to the domain and the choice of the vignette and is consistent over time. For this we use data from both waves of the Survey of Health, Ageing and Retirement in Europe (SHARE) for three domains of health: cognition, breathing and mobility. For cognition our results indicate that the method is sensitive to the choice of the vignette: one vignette brings the self-assessments closer to the objective situation, while two others do not. When possible results are found to be consistent over time. The breathing vignette collected in wave 2 brings the self-assessments closer to the objective situation. However, our results based on data from wave 1 also indicate that this might be sensitive to the choice of the vignette. Most positive are our results for mobility, which reveal that all vignettes bring the self-assessments closer to the objective situation.

JEL Classification: C81, I12
Keywords: Anchoring vignettes, heterogeneous reporting behavior, CHOPIT model

*Netspar, CentER, Tilburg University. *Address:* Tilburg University, Department of Econometrics and OR, PO Box 90153, 5000 LE Tilburg, The Netherlands. *E-Mail:* H.Vonkova@uvt.nl.

†Netspar, CentER, Tilburg University. *Address:* Tilburg University, Department of Economics, PO Box 90153, 5000 LE Tilburg, The Netherlands. *E-Mail:* P.G.J.Hullegie@uvt.nl.

1 Introduction

Self-assessments of health, work disability, job/life satisfaction and other concepts have frequently been used by researchers to study individual attitudes and behavior. These types of questions are commonly asked in socioeconomic surveys and are often used to study differences between countries or to compare different groups within a country. To give an example, consider the typical survey question that asks respondents to self-assess their health: *"Would you say your health is...,"* with answers ranging from *"very bad"* to *"very good"*. The answers to this and other self-assessment questions depend on the objective situation as well as on the individual's reporting behavior.

The following two examples illustrate that direct comparison of self-assessments might be misleading in case interest is in the objective situation. First, Aboriginals in Australia have been found to report better self-assessed health than the overall Australian population while all other indicators show that Aboriginals are at a serious disadvantage with respect to health (Bago d'Uva, Van Doorslaer, et al., 2008). Second, Sen (2002, p. 860) considers the well known evidence from different states of India: "The state of Kerala has the highest level of literacy (nearly universal for the young) and longevity (a life expectancy of about 74 years) in India. But is also has, by a very wide margin, the highest rate of reported morbidity among all Indian states (this applies to age specific as well as total comparisons). At the other extreme, states with low longevity, with woeful medical and educational facilities, such as Bihar, have the lowest rates of reported morbidity in India."

Furthermore, Kapteyn et al. (2007) write that "it is often hypothesized that men report themselves in better health than objective circumstances would warrant, that as people age they adjust their norms of what constitutes as good health downward, and that some of the socioeconomic status (SES) health gradient reflect different health thresholds by SES rather than true health differences." See also Bago d'Uva, O'Donnell, and Van Doorslaer (2008), which argues that

differences in reporting behavior might also come from differences in education. If more highly educated individuals are better informed about treatment options they might have stricter norms. With cross-country comparisons there is the additional concern that response categories may have different connotations. To quote Jürges (2007, p. 164): "For instance *'excellent'* is a term that is used in everyday parlance in the Anglo-Saxon world, but Germans would often consider *'ausgezeichnet'* as an ironic exaggeration, in particular if used in the context of health."

If differences in reporting behavior are systematic and interest is in the objective situation, direct comparison of the self-assessments might be misleading as it is impossible to know how much of the answers reflects objective differences among people and how much reflects heterogeneity in reporting behavior (also referred to as "differential item functioning" (DIF)).

King et al. (2004) have introduced anchoring vignettes as a tool to correct self-assessments for differences in reporting behavior. An anchoring vignette is a short description of aspects of a hypothetical person's life which are relevant to the domain of interest. Because the situation described in the vignette is the same for every respondent, anchoring vignettes help to identify heterogeneity in reporting behavior. However, this requires identifying assumptions: (1) *response consistency,* which is the assumption that individuals use the same subjective thresholds in rating the vignette person as they use when rating themselves; (2) *vignette equivalence,* which is the assumption that the level of the variable represented in the vignette is understood in the same way by all respondents.

The use of anchoring vignettes has grown rapidly and has not been restricted to health economics. See Van Soest et al. (2007) for other fields of application and corresponding references. An important question that has not been studied thoroughly is why DIF-adjusted self-assessments are better than unadjusted self-assessments and whether this is always the case, if at all.

The aim of this paper is to study whether the anchoring vignettes method is sensitive to the domain and the choice of the vignette and is consistent over time. For this we use data from both waves of the Survey of Health, Ageing and Retirement in Europe (SHARE) for three domains of

health: cognition, breathing, and mobility. The SHARE data contain, besides self-assessments and vignette evaluations, also objective measures. These data are required for our approach, which "tests" whether the DIF-adjusted self-assessments are "closer" to the objective measure than the unadjusted self-assessments.

The remainder of this paper is organized as follows. Section 2 discusses the literature. The data are described in Section 3 and the model in Section 4. The results are discussed in Section 5. Finally Section 6 concludes.

2 Related literature

The first comparison of unadjusted and DIF-adjusted self-assessments to an objective measure is, as far as we are aware, reported by King et al. (2004). They use self-assessments and vignette evaluations on visual acuity collected by the WHO for China and Slovakia. While the self-assessments do not show a significant difference of visual acuity between Chinese and Slovakian respondents, the measured test for vision suggests that the Chinese respondents have a substantially worse vision. The authors find that, in contrast with the unadjusted self-assessment, the DIF-adjusted self-assessments produce results that go in the same direction as those of the measured test.

Van Soest et al. (2007) introduce a more formal test for the *response consistency* assumption. Their proposed method tests whether the distribution of DIF-adjusted self-assessments in "closer" to the distribution of an objective measure than the unadjusted distribution. Van Soest et al. (2007) apply the method to data on drinking behavior collected from a web-based survey of students at a large Irish university. The results suggest that allowing for heterogeneous reporting behavior substantially improves the fit of the model as well as the correlation between the self-assessments and objective measure.

The paper by Gupta et al. (2008) takes a similar approach as Van Soest et al. (2007) and

seeks to test the validity of the *response consistency* assumption using data from the first wave of SHARE. They mainly focus on work disability and find that DIF-adjusted self-assessments are not more in line with the objective information than the unadjusted self-assessments.

A final paper related to ours is the one by Jürges and Winter (n.d.). They use data from a survey experiment to study whether the vignette evaluations are sensitive to vignette age and sex. The authors find that vignette evaluations are sensitive to the sex of the vignette person. By contrast, for the full sample Jürges and Winter (n.d.) find that the age of the vignette person, as implied by the name, does not effect the vignette evaluations. But for the sample of older respondents (80+) vignette evaluations are found sensitive to the age of the vignette person.

3 Data

We use data from both waves of the Survey of Health, Ageing and Retirement in Europe (SHARE) to address the question whether the vignette method is sensitive to the domain and the choice of the vignette and is consistent over time. SHARE is "a multidisciplinary and cross-national panel database of micro data on health, socioeconomic status and social and family networks of more than 40,000 individuals aged 50 or over." (www.share-project.org) In 2004 data were collected in eleven countries for the SHARE baseline wave.[1] Data for Israel were collected in 2005-06. The Czech Republic, Ireland and Poland joined SHARE in 2006 and contributed data for the second wave collected in 2006-07. For both waves the core information, like individual characteristics, was collected by a computer-assisted personal interview. In both waves self-assessments and associate anchoring vignettes were collected for six different domains of health using a self-completion drop-off part. For each of these six domains three vignettes were collected in the first wave. By contrast, the second wave collected one vignette per domain, which was chosen out of the three from the first wave. The drop-off questionnaire for the self-assessments and vignette

[1]The countries are: Austria, Belgium, Denmark, France, Germany, Greece, Italy, the Netherlands, Spain, Sweden, and Switzerland.

evaluations was only collected for a subsample of the overall SHARE sample for the following countries, where the number(s) in parentheses indicate(s) the wave(s) for which the data are available: Belgium (1,2), Czech Republic (2), Denmark (2), France (1,2), Germany (1,2), Greece (1,2), Italy (1,2), the Netherlands (1,2), Poland (2), Spain (1,2) and Sweden (1,2).[2] The data from the vignette subsamples of wave 1 and wave 2 are used to validate the vignette method for three domains of health: cognition, breathing, and mobility.

To keep as much information as possible we use different samples for each health domain.[3] These samples are subsamples of the vignette sample for the relevant wave, and might not be representative samples of the population aged 50 and over, as is SHARE. However, how representative these samples are for the population aged 50 and over is not a critical issue for this paper. The descriptive statistics for the self-assessments, vignette evaluations and objective measures, as described below, are based on the subsamples for the respective domains. The descriptive statistics of the covariates are described for the vignette samples of both waves. It will be noted when they substantially differ from those for the subsamples.

3.1 Self-assessments and vignette ratings

To begin, consider as an example the self-assessment question for cognition

> "*Overall in the last 30 days how much difficulty did you have with concentrating or remembering things?"* (none, mild, moderate, severe, extreme)

The self-assessment questions for the other domains are formulated similarly. See Appendix A for the exact wording, which is the same in both waves. As noted already, the vignette collected in the second wave of SHARE was chosen out the three vignettes collected in the first wave.

[2]Self-assessments and vignette evaluations were collected for Greece in both waves, but were not available in the release of the wave 2 data we use.

[3]Our analysis requires the availability of an objective measure for each domain. For example, for the mobility domain we use the result of a stand-up test. Only respondents younger than 75 were eligible for this test. For the cognition domain we use objective measures for which all respondents are asked to participate.

Each vignette describes aspects of a hypothetical person's life relevant to the domain. To give an example, consider the cognition vignette collected in both waves:

> "*(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. Overall in the last 30 days, how much difficulty did (Lisa) have with concentrating or remembering things?"* (none, mild, moderate, severe, extreme)

The exact wording of all three anchoring vignettes collected for each of the domains of cognition, breathing and mobility is given in Appendix A.

Next, we discuss descriptive statistics for the self-assessments and vignette evaluations. For the first wave, Table 1 shows the self-assessments and vignette evaluations for cognition. For all countries, except Belgium and the Netherlands, most respondents report to have no problems with concentrating or remembering things, followed by mild as the second most given answer. In all countries, but on Greece and Sweden, most respondents consider the hypothetical person described in vignette c1 to have a mild cognitive problem. In Greece most respondents think the vignette person has no problem, whereas the majority of the Swedish respondents judge the person to have a moderate problem. The evaluations of vignette c2 indicate that the respondents consider the hypothetical person described in this vignette to have more serious cognitive problems than the person of vignette c1. The person in vignette c3 is judged to have the most serious cognitive problems.

Table 2 provides the self-assessments and vignette evaluations for breathing based on data from wave 1. In all countries the majority of respondents report to have no breathing problem, followed by mild as the second most given answer. The hypothetical person in vignette b1 is judged to have moderate breathing problems by most respondents in all countries but on France and Spain. In France most respondents think the person has no problems, whereas in Spain the majority thinks the person has severe breathing problems. The evaluations of of vignette b2

reveal that the person described in that vignette to have more serious breathing problems than the person of vignette b1. The person described in vignette b3 is judged to have the most serious breathing problems.

Self-assessments and vignette evaluations for mobility based on wave 1 data are given in Table 3. The majority of respondents in all countries report to have no problems with moving around. The hypothetical person described in vignette m1 is considered to have either a mild or moderate problem by the majority of respondents in all countries. The person of of vignette m2 is judged to have more serious problems, as most of the respondents judge this person to have either moderate of severe problems moving around. The person described in vignette m3 is judged to have the most serious mobility problems.

Self-assessments and vignette evaluations from SHARE wave 2 are shown in Table 4 for cognition, breathing and mobility.

For the cognition domain, the self-assessments are very similar across time, for those countries that are in both waves. The exception is Sweden, where respondents report to have less serious cognitive problems. The hypothetical person in vignette c1, the vignette collected in both waves, is considered to have a mild cognitive problem by the majority of respondents in all countries. The vignette evaluations are fairly similar across time for those countries that are in both waves except for Sweden. There, the vignette evaluation became milder.

Next, consider the self-assessments and vignette evaluations for breathing in wave 2. The number of respondents within a category decreases when the category describes a more serious problem. So, most respondents report to have no breathing problem. The vignette person is considered to have a moderate breathing problem by the majority of respondents in all countries, but on Poland, Spain and Sweden. The majority of respondents in these three countries think the person has severe problems.

Finally consider the self-assessments and vignette evaluations for mobility in wave 2. In all countries but the Czech Republic, the vast majority of respondents report to have no mobility

problem. In the Czech Republic relatively many respondents think of themselves as having either a mild (42 %, average is 23.7%) or moderate (20%, average is 11.2%) problem with walking around. The vignette person is judged to have either a mild or moderate problem with walking around by the majority of respondents in all countries.

In our empirical analyses, we always combine the two categories "severe" and "extreme" for both self-assessments and vignette evaluations, because especially in the latter category there are few observations.

3.2 Objective measures

To study whether the vignette method is sensitive to the domain and the choice of the vignette and is consistent over time, we "test" whether the DIF-adjusted self-assessments are closer to an objective measure than the unadjusted self-assessments. So, essential to our approach is the availability of an objective measure. Here we discuss the objective measures used per domain. The exact wording of the questions used for the objective measures can be found in Appendix A.

For cognition we use the following four objective measures: immediate and delayed verbal memory, verbal fluency and numerical ability.[4] Immediate and delayed verbal memory were assessed using a 10-word learning task. Respondents were being read a list of 10 words by an interviewer. Then, respondents were asked to immediately recall as many words as possible, and as well after a short delay during which they answered other questions assessing cognitive functioning. Verbal fluency was examined by letting respondents name as many different animals as they could think of in 1 minute. Numerical ability was assessed using four different questions involving simple calculations. To give an example, consider the first question asked to all respondents:

[4]The variables used to measure cognitive functioning in SHARE are very similar to those used in other, well-known, surveys such as the English Longitudinal Study of Ageing (ELSA), the Asset and Health Dynamics Among the Oldest Old (AHEAD) study and the Health and Retirement Study (HRS). See e.g., Mehta et al. (2003); Llewellyn et al. (2008); Herzog and Wallace (1997).

"If the chance of getting a disease is 10 per cent, how many people out of 1000 (one thousand) would be expected to get the disease?"

All objective measures described above are available in both waves. Table 5 shows the percentage of respondents recalling a different number of words after a delay for the second wave of SHARE.[5] It discloses that in Italy, Poland, and Spain relatively many individuals recall, after a delay, no words or only one or two. The respondents in all other countries perform better, as most of them recall three up to six words. In Denmark, the Netherlands, and Sweden relatively many respondents are able to recall seven or eight words.

In our empirical analyses, immediate recall is mapped into 5 different categories. Delayed recall and numeracy into 4 categories, and verbal fluency (the number of animals mentioned) into 5 different groups.[6]

As objective measure for breathing we use the result of a so-called peak flow test, which measures how fast respondents can exhale while breathing out as hard and fast as possible.[7] Respondents are asked to do the test twice. If two measurements are available for a respondent we take the maximum value, otherwise we take the single measurement available as value. The unit of measurement of the peak-flow test is liters/minute and it ranges from 60 to 880.[8] Note that this measure is only available for SHARE wave 2. Table 6 provides descriptive statistics for the peak flow test. It discloses that the lung capacity of the Italian and Spanish respondents is relatively low. This is also true, but to a lesser extent, for the Polish and Czech respondents. The respondents in Denmark, the Netherlands, and Sweden have the largest average lung capacity.

As objective measure for mobility SHARE wave 2 respondents are asked to fold their arms

[5]The other three objective measures lead to similar conclusions. There are also no substantial differences between both waves.

[6]We merge the categories containing few observations.

[7]This test has been widely used in clinical practice to assess airflow obstruction and for monitoring patients with asthma (Nunn & Gregg, 1989; Quanjer et al., 1997).

[8]Interviewers were instructed to record a value of 30 if the respondent's measurement was less than 60, and to record a value of 890 if the respondent exhaled more than 880 liters per minute. In our study we observe this for a negligible (less than 2 percent) number of observations.

across their chest and to keep it like this while standing up. Our objective measure for mobility then is the time in seconds needed for five stands. Only respondents younger than 75 are eligible to do the test. From this group, only individuals who think it is safe to do a practice-test (stand-up once), who are able to do the practice-test according to instructions, and who think it is safe to stand-up five times are asked to do the test. We only observe the time in seconds for those respondents who are able to complete the test within one minute.[9] Descriptive statistics for the stand up test are given in Table 6. They reveal that 50 percent of the respondents completes the test within 10 seconds, but that the average time needed to finish the test is larger. The Italians and Spanish are relatively slow, whereas the Danes, Germans, and Swedish are relatively fast.

3.3 Covariates

The anchoring vignettes method models the "true" level of health and reporting heterogeneity using a vector of covariates. The model will be discussed in more detail in Section 4. In our paper we include the following covariates: country, age in groups of 5 years, gender, low/mid/high education, living alone, suffering from a long-term illness, never/sometimes/often engaged in physical activity. The education variable is based on the International Standard Classification of Education. Note that the long-term illness variable does not distinguish between the number and severity of chronic illnesses. Further information regarding the "construction" of our covariates can be found in Appendix B. Important to note here is that reference groups differ across our empirical analyses, but we always followed the rule that the group with most observations is taken as the reference group. Descriptive statistics of the covariates for waves 1 and 2 are given in Tables 7 and 8 respectively. Recall that they are given for the vignette samples and not for the subsamples of cognition, breathing, and mobility. However, it is noted when there are substantial differences.

[9] If the test was not completed within one minute we only observe that the respondent needed more than one minute and not the exact time. In our study we observe this for a negligible (0.5 percent) number of observations.

Consider first the descriptive statistics for wave 1. The second column of Table 7 shows that there is a large variation in the number of observations between countries: for France there are 885, while for Sweden 417. The percentage of male respondents, however, varies less and is in all countries approximately 45 percent. In Germany, nearly no respondent is low educated, whereas in Italy and Spain more than 50 percent is.[10] The number of respondents reporting to suffer from a long-term illness is 30.3 percent and 38.8 percent in Greece and the Netherlands respectively, but around 56 percent in Germany, Spain and Sweden. Of the Greek respondents 44 percent and of the Dutch 49.8 percent report to be often engaged in physical activity. Both numbers are significantly higher than those for the other countries. The average age of the respondent is around 62. The descriptive statistics for the cognition subsample are similar to those of the vignette sample.

Next, consider the descriptive statistics for the second wave as shown in Table 8. Also here there is a large variation in the number of respondents: 1164 in Germany and only 388 in France. The percentage of male respondents is, as in wave 1, around 45 percent. In the Czech Republic, Denmark, Germany and Poland few respondents have low education, whereas in Italy and Spain this is around 50 percent. There is also some variation in the percentage of respondents not living alone: in the Czech Republic and France this is 62.9 percent and 64.3 percent respectively, whereas in Italy and the Netherlands this is around 83 percent. In Poland 69.8 percent of the respondents report to suffer from a long-term illness, whereas in Italy this is 37.6 percent. The number of respondent reporting to be often engaged in physical activity in the Netherlands, 55.6 percent, is significantly higher than in other countries. The respondent is on average 64 years old. It should be noted that the wave 2 subsamples for cognition and breathing are similar to the vignette sample just described, but that the wave 2 mobility sample differs substantially. This is likely to be due to the selection rules used for the stand-up test: only those younger than 75 are

[10]The descriptive statistics for education in the vignette sample are similar to those for the overall SHARE sample, which is a representative sample for the population aged 50 years and older in all participating countries. This is true for both wave 1 and wave 2.

eligible to take the stand-up test.[11] The mobility subsample contains fewer observations for each country. Respondents are slightly better educated, live less often alone, suffer less often from a chronic illness, are more often engaged in physical activity and are on average younger.

4 Model

The approach taken in this paper to validate the anchoring vignettes method requires the availability of an objective measure. We follow Van Soest et al. (2007), which extends the compound hierarchical ordered probit (CHOPIT) model, by also modeling the objective measure.

4.1 Model for self-assessments

The self-assessment, y_{si}, of individual i is modeled as an ordered response equation with latent variable

$$y_{si}^* = \mathbf{x}_i'\beta_s + \varepsilon_{si},$$

where $\mathbf{x}_i$ is a vector of covariates including a constant term, and β_s a vector of parameters. The error term, ε_{si}, is assumed to be normally distributed with mean zero and variance σ_s^2, and independent of the covariates $\mathbf{x}_i$. The reported and observed responses, y_{si}, are generated by the following mechanism

$$y_{si} = k \Leftrightarrow \tau_{si}^{k-1} < y_{si}^* \leq \tau_{si}^k, \qquad k = 1, \ldots, K$$

[11] Of those that are eligible 17 per cent does not perform the test because for them it is either unsafe to do or they are not able to stand-up once (practice test).

where $-\infty = \tau_{si}^0 < \tau_{si}^1 < \ldots < \tau_{si}^K = +\infty$. The thresholds are modeled as

$$\begin{aligned}
\tau_{si}^1 &= \mathbf{x}_i'\gamma_s^1 + u_i, \\
\tau_{si}^k &= \tau_{si}^{k-1} + \exp\left(\mathbf{x}_i'\gamma_s^k\right), \quad k = 2, \ldots, K-1,
\end{aligned}$$

where $\mathbf{x}_i$ is a vector of covariates, and γ_s^k, for $k = 1, \ldots, (K-1)$, are vectors of parameters. The random effect, u_i, is assumed to be normally distributed with mean zero and variance σ_u^2, and independent of the covariates $\mathbf{x}_i$.

The idea that reporting behavior varies across individuals is formalized by modeling the thresholds to be individual-specific. The latent variable, y_{si}^*, can be interpreted as the true level of health as perceived by the individual. Note that using only self-assessments, the parameter vectors β_s and γ_s^1 are not separately identified, but the parameter vectors γ_s^k, for $k > 2$, are. That is, using only self-assessments we are not able to "decompose" the self-assessments in a part that is due to differences in "true" health (β) and a part due to heterogeneity in reporting behavior (γ_s^k, $k = 1, \ldots, K-1$).

4.2 Model for vignettes

Although in SHARE wave 1 three vignettes were collected for each domain, we only estimate models using a single vignette. The reason is that we want to study whether the method is sensitive to the choice of the vignette. The discussion below is based on the availability of a single vignette.

Under the assumption that there is an actual level of health, ϑ, associated with the hypothetical person described in the vignette, vignettes can be used to correct self-assessments for heterogeneity in reporting behavior. The assumption that the actual level of health of the hypothetical person described in the vignette is the same for every individual formalizes the vignette equivalence assumption. Each respondent perceives the actual level of health only with random

error, i.e.,

$$y_{vi}^{*} = \vartheta + \varepsilon_{vi},$$

where the error term, ε_{vi} is assumed to be normally distributed with mean zero and variance σ_v^2 and independent of the covariates $\mathbf{x}_i$. The observed vignette evaluations are generated by the following mechanism

$$y_{vi} = k \Leftrightarrow \tau_{vi}^{k-1} < y_{vi}^{*} \leq \tau_{vi}^{k}, \qquad k = 1, \ldots, K$$

where $-\infty = \tau_{vi}^{0} < \tau_{vi}^{1} < \ldots < \tau_{vi}^{K} = +\infty$. The thresholds are modeled similarly as in the self-assessment model, i.e.,

$$\begin{aligned}
\tau_{vi}^{1} &= \mathbf{x}_i' \gamma_v^{1} + u_i, \\
\tau_{vi}^{k} &= \tau_{vi}^{k-1} + \exp\left(\mathbf{x}_i' \gamma_v^{k}\right), \qquad k = 2, \ldots, (K-1),
\end{aligned}$$

where the term u_i is assumed to be the same in the thresholds of the self-assessment and vignette model. It introduces unobserved individual heterogeneity and implies that the vignette evaluation is correlated with the self-assessment (conditional on the covariates $\mathbf{x}_i$).

The response consistency assumption is formalized by assuming: $\tau_{si}^{k} = \tau_{vi}^{k}$, for $k = 1, \ldots, (K-1)$. In terms of the parameters this amounts to assuming $\gamma_s^{k} = \gamma_v^{k}$, for $k = 1, \ldots, (K-1)$.

4.3 Model for objective measure

To study whether the anchoring vignette method brings self-assessments closer to the objective situation we make use of measures of the objective situation. The four objective measures for cognition: immediate and delayed verbal memory, numeracy and verbal fluency, are all discrete

variables. In that case we model our objective measure as follows

$$y_{oi}^{*} = \mathbf{x}_i'\beta_o + \varepsilon_{oi},$$
$$y_{oi} = l \Leftrightarrow \tau_o^{l-1} < y_{oi}^{*} \leq \tau_o^{l},$$

where $-\infty = \tau_o^0 < \tau_o^1 < \ldots < \tau_o^L = +\infty$, are unknown thresholds that are the same for all individuals. The thresholds are modeled as

$$\tau_o^1 = \exp\left(\gamma_o^1\right),$$
$$\tau_o^l = \tau_o^{l-1} + \exp\left(\gamma_o^l\right), \qquad l = 2, \ldots, (L-1).$$

The objective measures for breathing and mobility, the result of the peak flow test and the stand-up test, respectively, are continuous variables, and are modeled as follows:[12]

$$y_{oi} = \mathbf{x}_i'\beta_o + \varepsilon_{oi}.$$

In both cases, discrete and continuous, the error term ε_{oi} is assumed to be independent of the covariates, $\mathbf{x}_i$, the unobserved heterogeneity term, u_i, and the error term of the vignette model, ε_{vi}. However, ε_{oi} is allowed to be correlated with the error term of the self-assessment model, ε_{si}, because the covariates might not capture all variation in "true" health, y_{si}^{*} and y_{oi}^{*}. The distribution of $(\varepsilon_{si}, \varepsilon_{oi})$ is assumed to be bivariate normal with mean zero, variances σ_s^2 and σ_o^2, and correlation ρ.

[12]Note that the objective measures of breathing and mobility are both, in principle, affected by censoring. However, recall that the number of censored observations is negligible and therefore we do not model the censoring.

4.4 Likelihood

The likelihood contribution of each individual i conditional on the unobserved heterogeneity, u_i, can be written as the product of a joint normal probability for the self-assessment and the objective measure, and a single normal probability for the vignette. In case of a discrete objective measure, the unconditional likelihood contribution for individual i is given by

$$\int \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{m=1}^{M} P\left(y_{si}=k, y_{oi}=l|\varphi,u_i\right)^{\mathbb{I}(y_{si}=k,y_{oi}=l)} P\left(y_{vi}=m|\varphi,u_i\right)^{\mathbb{I}(y_{vi}=m)} f\left(u_i\right) du_i,$$

where $f(.)$ is the normal density function with variance σ_u^2, and $\mathbb{I}(.)$ the indicator function. The vector of parameters is $\varphi = \left(\beta', \sigma_s^2, \gamma_s', \sigma_u^2, \vartheta, \gamma_v', \sigma_v^2, \beta_o', \tau_o', \sigma_o^2, \rho\right)'$.

4.5 Identification

We use three different models to study whether the vignette method is sensitive to the domain and the choice of the vignette, and is consistent over time. Each of these three models can be considered as a "special case" of the model discussed in the previous subsections, and every model has a different set of identifying assumptions.

CHOPIT model King et al. (2004) have introduced the CHOPIT model, which combines the self-assessment and vignette part of the model discussed before, i.e. the objective part is not included. For identification reasons the constant term of the β_s equals zero and the variance of error term in the self-assessment part is normalized to one, i.e., $\beta_{s,1} = 0, \sigma_s^2 = 1$. Assuming *vignette equivalence* and *response consistency* the CHOPIT model can separately identify β_s and $\gamma^1 = \gamma_s^1 = \gamma_v^1$. This can intuitively be understood as follows: the parameter vectors γ^k, for $k = 1, \ldots, (K-1)$, are identified by the vignette evaluations, and the parameter vector β_s then is identified by the self-assessments. Therefore, the CHOPIT model corrects the self-assessments for heterogeneity in reporting behavior.

Model A (No DIF, No RC) Reporting behavior is assumed to be homogeneous in this model, therefore $\tau_{si}^k = \tau_s^k$ and $\tau_{vi}^k = \tau_v^k$, for $k = 1, \ldots, (K-1)$, and $\sigma_u^2 = 0$. The model does not impose response consistency, that is, it allows for the possibility that $\tau_s^k \neq \tau_v^k$ for $k = 1, \ldots, (K-1)$. However, for identification reasons $\tau_s^1 = \tau_v^1 = 1$. The variances of the error terms in both the self-assessment and vignette model are normalized to one, i.e., $\sigma_s^2 = \sigma_v^2 = 1$. In case of an discrete objective measure, the first two thresholds of the objective measure are equal to one and two, i.e., $\tau_o^1 = 1, \tau_o^2 = 2$, for identification reasons.

Model B (DIF, RC) This model assumes that reporting behavior is heterogeneous across individuals and therefore the thresholds are individual-specific. In addition it assumes that response consistency holds, i.e., $\tau_{si}^k = \tau_{vi}^k$, for $k = 1, \ldots, (K-1)$. Furthermore the model normalizes the constant term in the parameter vector of the first threshold to one, i.e., $\gamma_{s,1}^1 = \gamma_{v,1}^1 = 1$. The variance of the error term in the self-assessment model is normalized to one, $\sigma_s^2 = 1$. In case of an discrete objective measure, the first two thresholds of the objective measure are equal to one and two, i.e., $\tau_o^1 = 1, \tau_o^2 = 2$, for identification reasons.

5 Results

The anchoring vignettes method intends to make self-assessments more comparable across individuals by correcting for heterogeneity in reporting behavior. The aim of this paper is to study whether the vignette method is sensitive to the domain and the choice of the vignette, and is consistent over time.

As a first step in validating the vignette method, we investigate whether different vignettes lead to similar DIF-adjusted self-assessments. Recall, that in the first wave of SHARE three vignettes were collected for each health domain. We estimate the CHOPIT-model using one vignette at a time and compute three DIF-adjusted self-assessments for each domain. If dif-

ferent vignettes would lead to similar DIF-adjusted self-assessments the correlation coefficient between any two DIF-adjusted self-assessments (each based on a different vignette) would be close to one. We compute the correlations between the predicted systematic parts, $\hat{y}^*_{si} = \mathbf{x}_i'\hat{\beta}_s$, of the three CHOPIT-models. These correlations are, however, uninformative about the question whether DIF-adjusted self-assessments are "closer" to the objective situation than unadjusted self-assessments.

As a second step we therefore estimate the models A and B discussed before. These models are also estimated for each domain separately using one vignette at a time. For every model we compute the correlation coefficients between the predicted systematic parts of (y^*_{si}, y^*_{oi}) and between the simulated values of (y^*_{si}, y^*_{oi}).[13]

Cognition First, we report the correlations between different DIF-adjusted self-assessments, each based on one vignette. See Table 9. The correlations indicate that vignettes c2 and c3 lead to similar DIF-adjusted self-assessments, whereas those based on vignette c1 are different from the other two. All this reveals is that for cognition the DIF-adjustment is sensitive to the choice of the vignette.

Second, we estimate the models A and B, separately using data from wave 1 and wave 2. For wave 1, the models are estimated for each combination of one of the four objective measures and one of the three vignettes. For wave 2, the models are estimated for each of the four objective measures using the single vignette that is collected. Table 10 and 11 provide a summary of results based on data from wave 1 and wave 2, respectively.

Consider first the results for wave 1. We only discuss them for delayed recall, as they are consistent with the other objective measures except verbal fluency.[14] The model that corrects for

[13]The predicted systematic parts are: $\hat{y}^*_{si} = \mathbf{x}_i'\hat{\beta}_s$ and $\hat{y}^*_{oi} = \mathbf{x}_i'\hat{\beta}_o$. The simulated values are obtained as follows: using the estimates of σ_s^2, σ_o^2, and ρ we simulate values from the bivariate normal distribution. These simulated values of ε_{si} and ε_{oi} are then added to the predicted systematic parts to obtain simulated values for y^*_{si}, and y^*_{si}.

[14]The results based on delayed recall and data from wave 1 are consistent with the results based on immediate recall and numeracy and wave 1 data. They are also consistent with the results based on all four objective measures and data from wave 2. The only exception are the results based on verbal fluency and wave 1 data. In this latter

reporting behavior heterogeneity (model B) gives a correlation between the predicted systematic parts of (y^*_{si}, y^*_{oi}) of 0.52 compared to around 0.76 for the model that does not make this correction (model A). The correlations between the simulated values of (y^*_{si}, y^*_{oi}) are 0.22 and 0.26 for model B and A respectively. So, both correlation coefficients are lower for model B than for model A. We therefore conclude that DIF-adjusted self-assessments based on vignette c1 are more different from the objective situation than the unadjusted self-assessments.

Adjusting self-assessments using vignette c2 leads to a different conclusion. For both models the correlation between the predicted systematic parts is comparable; 0.75 for model A and 0.74 for model B. The correlation between the simulated values increases from 0.26 for model A to 0.30 for model B. On the basis of these correlations we conclude that the DIF-adjusted self-assessments based on vignette c2 are about as close to the objective situation as the unadjusted self-assessments.

Consider next the results when vignette c3 is used. The correlation between the predicted systematic parts increases from 0.76 for model A to 0.83 for model B, and the correlation between the simulated values increases from around 0.26 for model A to 0.34 for model B. Here we conclude that the DIF-adjusted self-assessments based on vignette c3 are closer to the objective measure than the unadjusted self-assessments are.

Finally, consider the results based on data from wave 2. Since the vignette collected in the second wave is chosen out of the three from the first wave, we can study whether the results are consistent over time. Vignette c1 is the vignette collected in both waves. If the results are consistent over time we expect to conclude that the vignette method for vignette c1 does not help. The first set of results provided in Table 11 are for cognition using data from wave 2. Here the results are consistent across the four objective measures and lead to the same conclusion as before: DIF-adjusted self-assessments based on vignette c1 are more different from the objective

case results indicate that for all three vignettes, DIF-adjusted self-assessments are "closer" to the objective situation than the unadjusted self-assessments. We found that these results are sensitive to the inclusion of observations from Greece, as leaving out those observations gives results that are in line with those reported for the other objective measures in both waves

situation than the unadjusted self-assessments. So, the results for vignette c1 are found to be consistent over time.

To summarize, our results reveal that for cognition the vignette method is sensitive to the choice of the vignette. Tables 12, 13, 14 and 15 give the parameter estimates of β_o and β_s of model B for vignette c1, c2 and c3 and wave 1 data and vignette 1 and wave 2 data, respectively.

could offer an explanation. The differences in the correlations are for an important part caused by the country dummies and gender dummy, as for these variables either of the two following cases occurs relatively often: (1) one of the two parameter estimates is significantly different from zero, while the other is not; (2) both are significantly different from zero, but with opposite signs.

Breathing First, we discuss the correlations between different DIF-adjusted self-assessments using data from wave 1. These correlations are reported in Table 9, and they reveal that the DIF-adjusted self-assessments based on vignettes b2 and b3 are very different from the one based on vignette b1. Vignettes b2 and b3 lead to DIF-adjusted self-assessments that are much more similar. Thus, we conclude that for breathing the DIF-adjustments are sensitive to the choice of the vignette.

Second, we investigate whether the DIF-adjusted self-assessments are closer to the objective variable than the unadjusted self-assessments. We do this using data from wave 2, for which an objective measure is available and vignette b1 is collected. For wave 1 there is no objective measure available. The results of the models A and B are reported in Table 11.

The correlation coefficient between the predicted systematic parts of (y^*_{si}, y^*_{oi}) equals 0.45 for model A and 0.53 for model B.[15] The correlation coefficient between the simulated values of (y^*_{si}, y^*_{oi}) equals 0.25 for model A and 0.27 for model B. Although perhaps low, both correlation

[15]The reason for the low correlations between the self-assessment and objective measure is that the parameter estimates of certain country and age dummies and the gender dummy show the same discrepancy as described earlier for cognition.

coefficients still increase when the self-assessments are adjusted for heterogeneity in reporting behavior. So, correcting for reporting behavior heterogeneity brings the self-assessments of wave 2 closer to the objective situation.

Mobility Finally, consider the results for mobility. We first give the correlations between different DIF-adjusted self-assessments using one vignette at a time and data from wave 1. Table 9 reports the correlations, which are high and approximately the same. Therefore, if the method works for one of the vignettes it is likely that it will work for the other two vignettes as well.

The results of the models A and B based on data from wave 2, for which an objective measure is available and vignette m1 is collected, are given in Table 11. The correlation between the predicted systematic parts of (y^*_{si}, y^*_{oi}) increases from 0.54 (model A) to 0.63 (model B), and the correlation between the simulated values of (y^*_{si}, y^*_{oi}) increases from 0.18 (model A) to 0.20 (model B). So, both correlation coefficients increase when allowing for heterogeneity in reporting behavior. Based on data from wave 2 we conclude that the vignette method helps for mobility.

6 Conclusions

The vignette method intends to bring self-assessments closer to the objective situation by adjusting them for heterogeneity in reporting behavior. The aim of this paper is to study whether the method is sensitive to the domain and the choice of the vignette, and is consistent over time. For this purpose we use SHARE data and focus on three different domains of health: cognition, breathing and mobility.

Our results clearly reveal that the method is sensitive to the choice of the vignette for cognition: DIF-adjusted self-assessments based on vignette c1 are more different from the objective situation than unadjusted self-assessments; for vignette c2 we conclude that the vignette method does not bring the self-assessments closer to the objective situation; the conclusions for vignette

c3 is that the self-assessments are brought closer to the objective situation. Vignette c1, which is collected in both waves of SHARE, leads to conclusions that are consistent over time. The conclusions for cognition are the same irrespective of the objective measure used, except verbal fluency for wave 1.

For the breathing vignette used in SHARE wave 2, vignette b1, our results indicate that the vignette method helps. However, our results also show that there is no guarantee that it would work with one of the two other breathing vignettes collected in the first wave of SHARE.

Most positive are the results for mobility. There the vignette method helps for the vignette collected in wave 2. Moreover, the vignette method is unlikely to be sensitive to the choice of the vignettes used in wave 1.

Although our results indicate that the vignette method is sensitive to the domain and choice of the vignette, this should not be taken as a reason to reject this method. Rather, future research should aim to improve the method. Our results suggests several ideas for future research.

First, for the cognition domain we found that different vignettes lead to different results. The vignette describing a hypothetical person with the most extreme cognitive problems (vignette c3) brings the DIF-adjusted self-assessments closer to the objective situation than the other two vignettes describing milder problems. This suggests that the level of health of the vignette person matters, at least in this case. More research should be done to find out, not only, how the level of health of the vignette person matters, but also how to formulate vignettes in general.

Second, despite the fact that the DIF-adjusted self-assessments are sometimes closer to the objective situation than unadjusted self-assessments, one might argue that they are not that much closer. Furthermore, our results show that the vignette method is sensitive to the choice of the vignette, at least for the domains of cognition and breathing. The reason may be that the CHOPIT model is incorrectly specified, in particular the *response consistency* assumption may not hold for all vignettes. More research should be done to find out whether or not the *response consistency* assumption is tenable. Also, relaxing distributional assumptions might be an interesting area of

future research aimed at improving the existing methods.

Acknowledgments

This paper uses data from SHARE release 2.0.1 (wave 1) and 1.1.0 (wave 2) as of December 2008. SHARE data collection in 2004-2007 was primarily funded by the European Commission through its 5th and 6th framework programmes (project numbers QLK6-CT-2001- 00360; RII-CT- 2006-062193; CIT5-CT-2005-028857). Additional funding by the US National Institute on Aging (grant numbers U01 AG09740-13S2; P01 AG005842; P01 AG08291; P30 AG12815; Y1-AG-4553-01; OGHA 04-064; R21 AG025169) as well as by various national sources is gratefully acknowledged (see http://www.share-project.org for a full list of funding institutions).

We would like to thank Teresa Bago d'Uva, Peter Kooreman and Arthur van Soest for helpful comments and the people of various SHARE country teams for answering our (data related) questions.

Table 1: Self-assessment and vignette evaluations for cognition - SHARE wave 1. Numbers are proportions in each category.

	B	F	DE	GR	IT	NL	ES	SE
Self-assessment								
none	33.88	39.15	44.33	52.59	42.12	42.69	44.13	56.60
mild	45.17	35.91	36.08	31.47	35.06	47.95	23.91	21.83
moderate	19.31	21.45	16.08	13.29	15.76	6.82	22.17	12.44
severe	1.46	3.24	3.51	2.66	5.41	1.95	9.57	8.38
extreme	0.18	0.25	0.00	0.00	1.65	0.58	0.22	0.76
Vignette c1								
none	17.85	16.08	23.30	41.40	27.53	21.83	16.74	5.58
mild	64.30	53.37	49.48	39.44	43.76	68.81	38.26	24.11
moderate	15.30	26.06	24.33	15.80	20.24	8.19	33.26	46.19
severe	2.37	3.87	2.27	3.36	7.76	1.17	11.52	23.60
extreme	0.18	0.62	0.62	0.00	0.71	0.00	0.22	0.51
Vignette c2								
none	2.19	4.74	8.25	11.47	6.35	1.36	4.35	0.51
mild	26.59	31.17	33.40	34.41	31.53	17.35	27.83	6.09
moderate	51.55	51.62	44.74	37.62	42.12	50.88	48.04	20.81
severe	18.76	11.60	13.20	15.94	18.59	25.15	19.13	57.87
extreme	0.91	0.87	0.41	0.56	1.41	5.26	0.65	14.72
Vignette c3								
none	1.28	2.37	2.89	3.08	4.00	1.17	0.43	0.25
mild	9.84	9.23	8.66	13.57	15.29	5.26	4.35	1.78
moderate	33.15	39.28	27.01	27.83	32.47	31.77	28.26	8.88
severe	45.90	44.39	50.72	44.06	40.24	39.38	60.87	58.63
extreme	9.84	4.74	10.72	11.47	8.00	22.42	6.09	30.46

Country abbreviations: Belgium (B), France (F), Greece (GR), Germany (DE), Italy (IT), the Netherlands (NL), Spain (ES), Sweden(SE).

Table 2: Self-assessment and vignette evaluations for breathing - SHARE wave 1. Numbers are proportions in each category.

	B	F	DE	GR	IT	NL	ES	SE
Self-assessment								
none	63.93	60.86	65.32	67.56	73.82	70.29	74.34	38.85
mild	25.14	22.09	20.16	24.58	16.04	23.50	15.57	29.32
moderate	9.11	13.50	10.28	5.90	6.60	3.88	7.02	21.55
severe	1.46	3.44	3.43	1.69	3.07	1.75	3.07	8.02
extreme	0.36	0.12	0.81	0.28	0.47	0.58	0.00	2.26
Vignette b1								
none	18.94	37.06	2.42	1.26	5.66	2.14	1.10	0.75
mild	34.43	32.52	14.52	24.58	28.54	26.60	7.02	15.54
moderate	33.70	24.66	49.60	43.96	38.21	43.11	34.21	43.86
severe	10.93	5.40	31.85	26.54	25.00	23.11	50.66	36.34
extreme	2.00	0.37	1.61	3.65	2.59	5.05	7.02	3.51
Vignette b2								
none	1.82	2.94	3.83	0.70	5.19	2.72	0.66	0.75
mild	4.92	2.21	7.66	5.34	8.96	4.27	3.51	7.02
moderate	23.68	18.53	23.99	18.54	21.46	20.97	15.79	14.79
severe	51.37	66.01	55.24	46.07	45.05	40.97	56.58	49.37
extreme	18.21	10.31	9.27	29.35	19.34	31.07	23.46	28.07
Vignette b3								
none	2.00	3.31	3.63	0.56	5.66	3.30	0.66	0.75
mild	1.64	1.72	3.43	1.26	4.25	1.75	1.32	3.76
moderate	5.46	5.28	8.06	10.81	11.56	6.60	16.01	7.02
severe	47.91	61.60	45.36	36.94	38.92	21.36	44.96	49.87
extreme	42.99	28.10	39.52	50.42	39.62	66.99	37.06	38.60

Country abbreviations: Belgium (B), France (F), Greece (GR), Germany (DE), Italy (IT), the Netherlands (NL), Spain (ES), Sweden(SE).

Table 3: Self-assessment and vignette evaluations for mobility - SHARE wave 1. Numbers are proportions in each category.

	B	F	DE	GR	IT	NL	ES	SE
Self-assessment								
none	55.35	66.91	46.26	74.30	58.55	57.93	52.49	38.36
mild	26.32	15.13	27.07	15.36	20.37	24.86	19.96	38.36
moderate	12.34	13.78	18.99	5.31	11.01	11.47	17.79	17.90
severe	4.54	3.69	7.27	3.63	7.26	4.40	8.68	4.60
extreme	1.45	0.49	0.40	1.40	2.81	1.34	1.08	0.77
Vignette m1								
none	11.43	9.23	5.86	9.64	21.78	4.02	3.04	14.58
mild	43.92	32.60	26.67	33.66	36.53	43.21	21.48	40.92
moderate	36.84	47.60	49.70	44.69	30.91	38.62	54.45	34.27
severe	7.44	9.84	16.97	11.59	9.84	11.85	20.39	9.72
extreme	0.36	0.74	0.81	0.42	0.94	2.29	0.65	0.51
Vignette m2								
none	2.18	2.71	3.43	1.26	4.68	1.91	1.30	1.28
mild	13.43	7.75	11.52	17.04	11.24	9.94	8.68	15.86
moderate	41.20	39.85	35.96	38.97	30.21	33.65	44.25	46.04
severe	35.93	45.88	43.84	36.45	43.79	39.77	40.56	35.04
extreme	7.26	3.81	5.25	6.28	10.07	14.72	5.21	1.79
Vignette m3								
none	1.81	2.34	1.21	0.56	4.22	1.53	0.65	0.00
mild	3.81	8.24	7.07	5.03	12.18	2.49	5.21	2.56
moderate	35.75	37.02	26.26	22.91	20.37	29.06	24.95	14.83
severe	43.56	47.72	56.77	41.06	51.52	39.39	59.87	59.08
extreme	15.06	4.67	8.69	30.45	11.71	27.53	9.33	23.53

Country abbreviations: Belgium (B), France (F), Greece (GR), Germany (DE), Italy (IT), the Netherlands (NL), Spain (ES), Sweden(SE).

Table 4: Self-assessments and vignette evaluations for cognition, breathing and mobility - SHARE wave 2. Numbers are proportions in each category.

	Self-assessment										Vignette evaluation									
	B	CZ	DK	F	DE	IT	NL	PO	ES	SE	B	CZ	DK	F	DE	IT	NL	PO	ES	SE
Cognition																				
none	30.88	39.21	54.66	35.41	43.38	37.89	39.88	37.38	43.34	41.68	25.73	29.83	32.49	21.81	26.76	28.78	32.26	12.52	15.51	27.21
mild	48.54	43.05	30.36	42.49	38.93	38.91	50.10	28.55	28.63	36.72	61.87	57.74	53.95	55.81	53.33	47.58	62.12	41.99	50.70	48.81
moderate	17.54	14.24	13.06	20.11	14.40	15.86	7.62	24.86	19.88	18.36	10.18	11.19	12.85	19.26	16.80	19.68	4.81	35.73	25.05	20.52
severe	2.69	3.05	1.92	1.70	2.93	6.31	1.80	7.00	6.56	3.02	2.11	1.02	0.71	3.12	2.67	3.82	0.80	8.84	8.75	3.02
extreme	0.35	0.45	0.00	0.28	0.36	1.03	0.60	2.21	1.59	0.22	0.12	0.23	0.00	0.00	0.44	0.15	0.00	0.92	0.00	0.43
Breathing																				
none	60.51	56.39	78.87	56.33	66.83	76.81	66.80	64.09	71.75	63.47	6.23	1.33	3.74	3.00	3.23	6.19	2.68	4.44	0.91	4.01
mild	27.38	27.35	14.09	25.00	20.06	16.28	27.63	16.02	17.31	24.05	5.33	28.31	26.57	23.33	28.80	32.92	25.77	11.58	12.53	24.94
moderate	9.41	13.01	5.55	16.00	9.03	4.96	3.71	11.58	7.74	8.69	1.93	50.12	41.62	50.33	39.92	38.58	53.20	30.89	27.33	28.06
severe	2.08	3.13	1.39	2.33	3.80	1.59	1.24	6.37	2.73	3.34	5.53	18.67	25.93	23.00	26.62	20.88	16.29	50.39	53.53	35.19
extreme	0.61	0.12	0.11	0.33	0.29	0.35	0.62	1.93	0.46	0.45	0.98	1.57	2.13	0.33	1.43	1.42	2.06	2.70	5.69	7.80
Mobility																				
none	59.20	33.28	76.96	82.71	51.00	67.19	55.83	55.27	64.16	63.17	9.29	4.18	8.23	8.41	5.72	12.24	5.10	8.55	6.14	5.71
mild	27.20	42.44	14.94	8.88	31.59	21.09	33.50	19.37	19.11	20.32	55.56	48.87	37.97	39.25	29.48	46.61	43.20	27.64	26.28	34.60
moderate	11.77	19.77	6.84	8.41	14.30	10.16	7.52	18.52	12.63	11.11	28.86	39.87	39.75	40.65	43.28	33.07	39.08	39.60	45.39	30.48
severe	1.66	4.18	1.27	0.00	3.11	1.56	2.67	6.27	3.75	5.08	6.30	7.07	13.92	11.68	20.27	7.81	11.89	23.65	21.84	27.30
extreme	0.17	0.32	0.00	0.00	0.00	0.00	0.49	0.57	0.34	0.32	0.00	0.00	0.13	0.00	1.24	0.26	0.73	0.57	0.34	1.90

Note: For cognition, breathing and mobility SHARE wave 2 collected vignettes c1, b1 and m1, respectively. Country abbreviations: Belgium (B), Czech Republic (CZ), Denmark (DK), France (F), Germany (DE), Italy (IT), the Netherlands (NL), Poland (PO), Spain (ES), Sweden(SE).

Table 5: Delayed recall - SHARE wave 2

Number of words	B	CZ	DK	F	DE	IT	NL	PO	ES	SE
0	8.54	12.77	5.06	5.38	4.53	11.45	4.61	16.21	10.74	3.24
1	9.12	7.57	3.85	9.92	4.80	11.01	3.81	11.60	15.11	3.46
2	10.64	13.67	7.69	13.88	11.20	16.89	11.02	17.50	17.89	9.50
3	20.12	20.56	15.79	20.96	19.64	18.50	13.63	21.36	21.67	18.14
4	21.99	21.58	19.64	18.70	20.09	19.38	19.04	19.34	18.89	19.87
5	14.27	13.79	20.95	16.43	20.53	9.99	15.83	8.29	8.55	21.81
6	8.89	6.33	13.97	9.07	10.67	6.46	16.23	3.68	4.77	13.39
7	4.44	2.60	7.79	3.68	5.87	3.38	9.02	1.47	2.19	6.26
8	1.29	0.79	3.14	1.70	1.87	1.47	4.01	0.55	0.20	3.24
9	0.70	0.11	1.82	0.28	0.53	0.44	2.00	0.00	0.00	0.65
10	0.00	0.23	0.30	0.00	0.27	1.03	0.80	0.00	0.00	0.43
mean	3.51	3.17	4.32	3.53	3.96	3.16	4.40	2.66	2.79	4.25
std.dev	1.96	1.91	2.01	1.87	1.86	2.08	2.13	1.79	1.75	1.86

Note: The first 11 rows give the percentage of respondents able to recall the given the number of words after a short delay. The last 2 rows give the mean number of words recalled after a delay and the standard deviation (std.dev). Country abbreviations: Belgium (B), Czech Republic (CZ), Denmark (DK), France (F), Germany (DE), Italy (IT), the Netherlands (NL), Poland (PO), Spain (ES), Sweden(SE).

Table 6: Descriptive statistics for the peak flow test (breathing) and the stand-up test (mobility) - SHARE wave 2

	Peak flow test			Stand-up test		
Country	median	mean	std.dev	median	mean	std.dev
Belgium	330	345.66	149.77	10.78	11.76	5.38
Czech Rep.	320	326.21	132.10	10.09	10.95	4.29
Denmark	390	394.32	145.88	9.19	9.69	3.26
France	350	358.40	174.67	10.00	10.89	6.06
Germany	350	361.42	149.19	9.50	10.98	6.09
Italy	280	294.63	145.09	10.65	12.59	7.44
Netherlands	390	403.24	149.06	10.73	11.75	5.28
Poland	305	325.34	153.01	10.01	11.17	4.60
Spain	270	328.53	225.51	11.00	12.62	6.52
Sweden	420	434.09	141.69	9.44	9.87	3.65

Note: The unit of measurement for the peak flow test is liters/minute and it ranges from 60 to 880. The unit of measurement for the stand-up test is time in seconds. We only observe the exact time for those respondents who are able to complete the test in 1 minute.

Table 7: Descriptive statistics for covariates - SHARE wave 1

Country	N	Male (%)	Education mid (%)	Education high (%)	Not alone (%)	Long-term illness (%)	Sport sometimes (%)	Sport often (%)	Mean age	Sd age
Belgium	567	43.7	50.7	25.2	73.9	47.8	24.2	27.3	62.57	9.72
France	885	42.7	36.3	19.5	68.0	48.8	21.0	32.0	63.90	10.38
Germany	508	43.5	74.7	24.5	77.6	57.1	27.4	41.7	63.44	9.14
Greece	720	45.8	37.2	20.1	73.8	30.3	44.0	34.2	60.82	10.61
Italy	445	43.8	36.4	6.5	73.4	45.6	21.3	26.7	63.53	9.34
Netherlands	538	47.6	64.1	21.6	84.2	38.8	16.7	49.8	62.24	9.38
Spain	464	41.8	26.8	10.2	72.0	55.2	12.9	26.5	64.64	10.53
Sweden	417	47.2	33.6	30.4	77.2	56.1	27.8	38.4	64.14	9.63

Table 8: Descriptive statistics for the covariates - SHARE wave 2

Country	N	Male (%)	Education mid (%)	Education high (%)	Not alone (%)	Long-term illness (%)	Sport sometimes (%)	Sport often (%)	Mean age	Sd age
Belgium	896	45.6	51.7	22.2	72.6	43.5	21.6	27.4	65.40737	10.040828
Czech Rep.	923	40.2	88.9	11.0	62.9	53.3	28.0	25.2	64.61388	9.994165
Denmark	1029	44.1	61.0	38.9	74.7	46.0	19.1	37.8	64.33625	9.937531
France	388	44.1	37.4	23.6	64.3	50.5	26.9	26.4	65.42268	10.359267
Germany	1164	45.7	69.7	29.6	76.2	54.5	28.7	36.5	65.05412	9.302206
Italy	697	45.8	41.2	9.1	83.0	37.6	24.4	24.1	64.82783	8.804007
Netherlands	527	47.6	59.2	27.2	82.3	38.8	19.0	55.6	61.73435	9.893810
Poland	566	43.3	81.3	17.3	73.9	69.8	21.4	25.6	62.99293	9.814201
Spain	518	45.4	29.7	13.2	79.3	51.9	19.0	34.8	64.29344	10.413817
Sweden	478	45.6	36.2	33.1	77.6	52.9	24.0	41.6	66.35774	10.102393

Table 9: Correlation matrix - SHARE wave 1

cognition		$\hat{y}_{c1}$	$\hat{y}_{c2}$	$\hat{y}_{c3}$
	$\hat{y}_{c1}$	1.0000	0.7553	0.7529
	$\hat{y}_{c2}$		1.0000	0.9237
	$\hat{y}_{c3}$			1.0000
breathing		$\hat{y}_{b1}$	$\hat{y}_{b2}$	$\hat{y}_{b3}$
	$\hat{y}_{b1}$	1.0000	0.5595	0.4351
	$\hat{y}_{b2}$		1.0000	0.9374
	$\hat{y}_{b3}$			1.0000
mobility		$\hat{y}_{m1}$	$\hat{y}_{m2}$	$\hat{y}_{m3}$
	$\hat{y}_{m1}$	1.0000	0.9633	0.8959
	$\hat{y}_{m2}$		1.0000	0.9371
	$\hat{y}_{m3}$			1.0000

Note: Numbers are correlation coefficients between two DIF-adjusted self-assessments, each computed by estimating the CHOPIT-model using 1 vignette at a time (denoted in the subscript).

Table 10: Summary of results for cognition - SHARE wave 1

Vignette	No. of parameters	Loglikelihood	AIC	Corr($\hat{y}^*_{si}, \hat{y}^*_{oi}$)	Corr($\varepsilon_{si}, \varepsilon_{oi}$)	Corr(y^*_{si}, y^*_{oi})
Immediate recall						
Vignette c1						
A	53	-16,209.12	32,462.24	0.7007	0.1243	0.2361
B	116	-15,689.02	31,422.04	0.4537	0.1215	0.1883
Vignette c2						
A	53	-16,255.96	32,555.91	0.7007	0.1243	0.2361
B	116	-15,770.65	31,585.30	0.6808	0.1307	0.2786
Vignette c3						
A	53	-15,197.01	30,438.02	0.7007	0.1243	0.2361
B	116	-14,894.04	29,832.08	0.7777	0.1291	0.3209
Delayed recall						
Vignette c1						
A	52	-15,137.61	30,319.22	0.7547	0.1485	0.2567
B	115	-14,618.11	29,280.22	0.5227	0.1474	0.2172
Vignette c2						
A	52	-15,184.44	30,412.89	0.7548	0.1485	0.2567
B	115	-14,698.00	29,440.00	0.7420	0.1587	0.3023
Vignette c3						
A	52	-14,125.50	28,295.00	0.7547	0.1485	0.2567
B	115	-13,821.86	27,687.73	0.8304	0.1555	0.3378
Numeracy						
Vignette c1						
A	52	-15,129.56	30,303.13	0.7486	0.0754	0.2102
B	115	-14,608.73	29,261.46	0.5372	0.0659	0.1650
Vignette c2						
A	52	-15,176.40	30,396.80	0.7486	0.0754	0.2102
B	115	-14,692.70	2,9429.40	0.7192	0.0733	0.2501
Vignette c3						
A	52	-14,117.46	28,278.91	0.7486	0.0754	0.2102
B	115	-13,815.77	27,675.54	0.7813	0.0746	0.2869
Verbal fluency						
Vignette c1						
A	53	-16,022.07	32,088.13	0.5045	0.1186	0.1968
B	116	-15,502.38	31,048.75	0.6163	0.1133	0.2242
Vignette c2						
A	53	-16,070.72	32,185.45	0.5045	0.1186	0.1968
B	116	-15,587.28	31,218.56	0.7333	0.1217	0.2996
Vignette c3						
A	53	-15,009.33	30,062.66	0.5045	0.1186	0.1968
B	116	-14,709.81	29,463.60	0.7444	0.1206	0.3202

Note: Model A assumes there is no heterogeneity in reporting behavior, and no response consistency (No DIF, No RC). Model B assumes there is heterogeneity in reporting behavior, and response consistency holds (DIF, RC). Estimations are based on data of 4343 respondents.

Table 11: Summary of results for cognition, breathing and mobility- SHARE wave 2

Domain	No. of parameters	Loglikelihood	AIC	Corr$(\hat{y}^*_{si}, \hat{y}^*_{oi})$	Corr$(\varepsilon_{si}, \varepsilon_{oi})$	Corr(y^*_{si}, y^*_{oi})
Cognition						
Immediate recall						
A	57	-25,010.81	50,069.62	0.7780	0.2063	0.3156
B	126	-24,558.24	49,164.48	0.4000	0.2064	0.2380
Delayed recall						
A	56	-23,223.06	46,494.13	0.7873	0.2059	0.3131
B	125	-22,770.11	45,588.21	0.4302	0.2064	0.2430
Numeracy						
A	56	-23,538.05	47,124.09	0.7052	0.1378	0.2433
B	125	-23,082.89	46,213.79	0.4028	0.1299	0.1767
Verbal fluency						
A	57	-24,880.49	49,808.97	0.7280	0.110	0.2389
B	126	-24,421.47	48,890.93	0.3737	0.1047	0.1556
Breathing						
A	57	-20,646.37	41,342.74	0.4463	0.1775	0.2456
B	129	-20224.53	40499.06	0.5268	0.1843	0.2748
Mobility						
A	51	-24,781.54	49,607.07	0.5453	0.1311	0.1837
B	114	-24,516.58	49,077.16	0.6297	0.1323	0.1969

Note: Model A assumes there is no heterogeneity in reporting behavior, and no response consistency (No DIF, No RC). Model B assumes there is heterogeneity in reporting behavior, and response consistency holds (DIF, RC). Estimations for cognition, breathing and mobility are based on data of 6895, 6393 and 4788 respondents, respectively.

Table 12: Parameters estimates for cognition: delayed recall and vignette c1 - SHARE wave 1

	model A				model B							
	β_o		β_s		β_s		γ^1		γ^2		γ^3	
covariates	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values
Constant	2.3984	39.3193	3.0045	40.6622	3.2282	25.9253	1.0000	-	0.0659	0.7030	0.1594	2.8574
Belgium	0.0216	0.3880	-0.0798	-1.2977	-0.2299	-2.7704	-0.2851	-2.8335	-0.0173	-0.1984	0.1987	4.3698
Germany	0.1111	1.8580	0.0477	0.7092	-0.0461	-0.5169	0.0625	0.6106	-0.0937	-1.0789	-0.1281	-2.4236
Greece	0.1133	2.1588	0.1895	3.1868	-0.2495	-3.2818	-0.0991	-1.1641	-0.2367	-2.9217	-0.2149	-4.2825
Italy	-0.3495	-5.7197	0.0407	0.6064	-0.1090	-1.2319	0.2901	3.5140	-0.4372	-5.1496	-0.1480	-2.6425
Netherlands	0.2347	4.0391	0.0895	1.3799	-0.1452	-1.6790	-0.2229	-1.9262	-0.4258	-3.5523	0.3290	7.0085
Spain	-0.4026	-6.6884	0.0293	0.4442	0.1518	1.7720	0.4780	5.9095	-0.2123	-2.9220	-0.3383	-5.6098
Sweden	0.3523	5.6780	0.2010	2.8339	0.8019	8.6411	1.0568	12.5279	-0.2397	-3.1888	-0.4478	-6.0038
Male	-0.2019	-6.3628	0.1073	3.0293	0.1329	2.8613	-0.0046	-0.0962	0.0565	1.2624	-0.0247	-0.8881
Age < 50	0.0893	1.3076	-0.0376	-0.4797	-0.0181	-0.1787	-0.0032	-0.0289	-0.0110	-0.0964	0.0394	0.6444
Age 55 to 60	-0.1184	-2.4764	0.0250	0.4547	0.0051	0.0716	-0.1958	-2.4946	0.1823	2.4435	0.0175	0.4018
Age 60 to 65	-0.2097	-4.0907	-0.0772	-1.3248	-0.0551	-0.7243	-0.1264	-1.5945	0.0613	0.7792	0.1458	3.2802
Age 65 to 70	-0.3801	-7.0501	-0.1912	-3.1753	-0.1137	-1.4141	-0.0657	-0.8234	0.0974	1.2546	0.1180	2.4825
Age 70 to 75	-0.4532	-7.6464	-0.2394	-3.6664	-0.1650	-1.9280	-0.1947	-2.3252	0.2688	3.5093	0.0867	1.7194
Age 75 to 80	-0.6265	-8.8247	-0.3185	-4.1424	-0.2763	-2.7619	-0.1184	-1.1712	0.1348	1.4328	0.1107	1.6865
Age > 80	-1.0369	-12.1891	-0.5114	-5.8866	-0.4621	-3.9078	-0.0813	-0.8391	0.1019	1.0449	0.0540	0.7059
Education low	-0.4145	-10.5895	-0.1843	-4.2687	-0.0755	-1.3178	0.2501	4.2255	-0.0798	-1.4760	-0.1086	-3.1156
Education high	0.2879	6.9441	0.1389	2.9412	0.1269	2.0603	0.1113	1.6488	-0.1156	-1.7344	-0.0330	-0.9261
Phys. act. sometimes	0.1206	2.9571	0.1401	3.0490	0.1109	1.8549	-0.1170	-1.8441	0.0586	1.0148	0.0563	1.5391
Phys. act. often	0.0513	1.3541	0.0942	2.2142	0.0656	1.1574	-0.0735	-1.3047	0.0059	0.1100	0.0568	1.7314
Alone	-0.1153	-3.1260	-0.1384	-3.3857	-0.1178	-2.2290	0.1344	2.5806	-0.0595	-1.2029	-0.0880	-2.6351
Illness long	-0.1055	-3.2841	-0.3841	-10.6429	-0.3589	-7.5814	0.0513	1.0508	-0.0046	-0.1010	-0.0070	-0.2478

	model A		model B	
Threshold parameters	coeff	t-values	coeff	t-values
γ_s^1	1.0000	-		
γ_s^2	-0.0437	-1.2476		
γ_s^3	0.0327	1.4768		
γ_v^1	1.0000	-		
γ_v^2	-0.0293	-1.0099		
γ_v^3	0.2735	15.1808		
γ_o^1	0.0000	-	0.0000	-
γ_o^2	0.0000	-	0.0000	-
γ_o^3	0.1136	3.3128	0.1142	3.2693
Variances				
σ_s^2	1.0000	-	1.0000	-
σ_u^2	0.0000	-	0.3453	13.0264
σ_v^2	1.0000	-	0.7529	34.0589
σ_o^2	0.9195	43.0112	0.9196	42.4812
ρ	0.1485	8.1440	0.1474	7.5498
Vignette dummy	coeff	s.e.	coeff	s.e.
ϑ	2.5201	0.0296	2.5687	0.1044

Note: the estimated β_o parameters are similar in model A and B.

Table 13: Parameters estimates for cognition: delayed recall and vignette c2 - SHARE wave 1

	model A				model B							
	β_o		β_s		β_s		γ^1		γ^2		γ^3	
covariates	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values
Constant	2.3984	39.3194	3.0045	40.6621	3.3928	29.9814	1.0000	-	0.2028	2.9768	0.1950	3.0062
Belgium	0.0216	0.3878	-0.0798	-1.2977	0.1986	2.1907	0.2329	3.4704	-0.0548	-0.9559	0.1576	2.7072
Germany	0.1111	1.8579	0.0478	0.7093	-0.1573	-1.6160	0.0792	1.0761	-0.1909	-2.9863	-0.1437	-2.2926
Greece	0.1133	2.1589	0.1895	3.1868	-0.0429	-0.5032	0.1432	2.2210	-0.3152	-5.2604	-0.1150	-1.9710
Italy	-0.3495	-5.7197	0.0407	0.6064	-0.0065	-0.0684	0.2797	3.9498	-0.3195	-4.9154	-0.0265	-0.4061
Netherlands	0.2346	4.0390	0.0896	1.3799	0.7488	7.5827	0.5974	8.7346	-0.1742	-2.8407	0.2426	3.8784
Spain	-0.4026	-6.6884	0.0293	0.4442	0.0176	0.1924	0.3055	4.4086	-0.2002	-3.3544	-0.2196	-3.1368
Sweden	0.3523	5.6779	0.2010	2.8338	0.9351	8.3633	1.5190	18.6135	-0.6429	-6.7649	-0.4573	-4.7436
Male	-0.2019	-6.3627	0.1073	3.0293	0.0204	0.3995	-0.0154	-0.4214	-0.0326	-0.9540	-0.0629	-1.8077
Age < 50	0.0893	1.3076	-0.0376	-0.4797	-0.1372	-1.1994	-0.0410	-0.5037	-0.0145	-0.1854	-0.0676	-0.9095
Age 55 to 60	-0.1184	-2.4764	0.0250	0.4547	0.0610	0.7451	0.0026	0.0468	0.0007	0.0124	0.0466	0.8608
Age 60 to 65	-0.2097	-4.0907	-0.0772	-1.3248	-0.0113	-0.1354	-0.1554	-2.5303	0.1874	3.3084	0.0622	1.1232
Age 65 to 70	-0.3801	-7.0500	-0.1912	-3.1752	-0.1343	-1.5391	-0.0253	-0.4183	0.1289	2.2408	-0.0393	-0.6597
Age 70 to 75	-0.4532	-7.6463	-0.2394	-3.6664	-0.2313	-2.4522	-0.1829	-2.7338	0.1837	2.9500	0.0737	1.1198
Age 75 to 80	-0.6265	-8.8247	-0.3185	-4.1424	-0.1823	-1.7219	-0.0211	-0.2725	0.1780	2.5071	0.0059	0.0756
Age > 80	-1.0369	-12.1889	-0.5114	-5.8865	-0.5289	-4.4012	-0.1254	-1.4690	0.1219	1.5714	-0.0220	-0.2466
Education low	-0.4145	-10.5895	-0.1843	-4.2689	-0.2955	-4.7702	0.0387	0.8683	-0.0200	-0.4785	-0.2150	-5.0392
Education high	0.2879	6.9441	0.1389	2.9412	0.1866	2.6395	0.0128	0.2651	0.0274	0.5933	0.0017	0.0373
Phys. act. sometimes	0.1206	2.9573	0.1401	3.0491	0.1032	1.5440	-0.0735	-1.5563	0.0189	0.4342	0.0462	1.0028
Phys. act. often	0.0513	1.3541	0.0942	2.2141	0.0653	1.0544	-0.0731	-1.6830	0.0002	0.0049	0.0714	1.7078
Alone	-0.1153	-3.1262	-0.1384	-3.3859	-0.2018	-3.5127	0.0511	1.2160	-0.0586	-1.5154	-0.0927	-2.2558
Illness long	-0.1056	-3.2842	-0.3841	-10.6429	-0.3881	-7.3161	0.0468	1.2636	-0.0161	-0.4633	-0.0411	-1.1474

	model A		model B	
Threshold parameters	coeff	t-values	coeff	t-values
γ_s^1	1.0000	-		
γ_s^2	-0.0437	-1.2479		
γ_s^3	0.0327	1.4768		
γ_v^1	1.0000	-		
γ_v^2	0.1715	8.9751		
γ_v^3	0.1523	5.7152		
γ_o^1	0.0000	-	0.0000	-
γ_o^2	0.0000	-	0.0000	-
γ_o^3	0.1136	3.3126	0.1141	3.2700
Variances				
σ_s^2	1.0000	-	1.0000	-
σ_u^2	0.0000	-	0.2827	8.9110
σ_v^2	1.0000	-	0.7873	33.8202
σ_o^2	0.9195	43.0107	0.9197	42.5004
ρ	0.1485	8.1440	0.1587	8.2474
Vignette dummy	coeff	s.e.	coeff	s.e.
ϑ	1.7304	0.0210	1.9276	0.0206

Note: the estimated β_o parameters are similar in model A and B.

Table 14: Parameters estimates for cognition: delayed recall and vignette c3 - SHARE wave 1

	model A				model B							
	β_o		β_s		β_s		γ^1		γ^2		γ^3	
covariates	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values
Constant	2.3984	39.3206	3.0045	40.6626	3.4627	28.7653	1.0000	-	0.2940	3.6277	0.1501	1.9913
Belgium	0.0216	0.3880	-0.0798	-1.2978	0.0869	0.7636	0.1067	1.4250	-0.0765	-1.1434	0.2174	3.2334
Germany	0.1111	1.8581	0.0477	0.7090	-0.1239	-1.0209	0.2533	3.0957	-0.3310	-4.1994	-0.1318	-1.6740
Greece	0.1133	2.1588	0.1895	3.1867	0.0146	0.1398	0.1800	2.5544	-0.3502	-5.2045	-0.0067	-0.0984
Italy	-0.3495	-5.7197	0.0407	0.6063	-0.2694	-2.3843	0.0397	0.4995	-0.3580	-4.8580	0.0141	0.1869
Netherlands	0.2347	4.0392	0.0895	1.3798	0.5239	4.0590	0.3684	4.6501	-0.2025	-2.5700	0.3033	4.2603
Spain	-0.4026	-6.6884	0.0292	0.4440	0.3024	2.5820	0.6017	7.2667	-0.1684	-2.3414	-0.3153	-3.4770
Sweden	0.3523	5.6781	0.2010	2.8338	0.7771	5.4302	1.3716	12.2120	-0.6676	-6.0652	-0.4305	-4.1446
Male	-0.2019	-6.3631	0.1073	3.0292	-0.0117	-0.1831	-0.0825	-1.9069	0.0102	0.2473	-0.0686	-1.6854
Age < 50	0.0893	1.3075	-0.0376	-0.4798	-0.1422	-0.9886	-0.1007	-1.0828	0.0803	0.8706	-0.1190	-1.3114
Age 55 to 60	-0.1184	-2.4765	0.0250	0.4546	0.0716	0.6987	-0.0231	-0.3479	0.0867	1.2875	-0.0113	-0.1801
Age 60 to 65	-0.2097	-4.0907	-0.0772	-1.3248	-0.0561	-0.5248	-0.0576	-0.8122	0.0325	0.4545	0.0829	1.2871
Age 65 to70	-0.3801	-7.0501	-0.1912	-3.1752	-0.1640	-1.4996	-0.0491	-0.6697	0.1267	1.7738	-0.0446	-0.6469
Age 70 to 75	-0.4532	-7.6464	-0.2394	-3.6664	-0.1900	-1.6109	-0.2251	-2.8039	0.2840	3.9083	0.0289	0.3815
Age 75 to 80	-0.6265	-8.8247	-0.3185	-4.1424	-0.3575	-2.6750	-0.1188	-1.2614	0.1336	1.5707	-0.0787	-0.8480
Age > 80	-1.0369	-12.1892	-0.5114	-5.8864	-0.6049	-4.1447	-0.2037	-1.9138	0.1498	1.6393	-0.1100	-1.0258
Education low	-0.4145	-10.5896	-0.1843	-4.2688	-0.5442	-7.0702	-0.1249	-2.3779	-0.1101	-2.2775	-0.2205	-4.4086
Education high	0.2879	6.9441	0.1388	2.9409	0.3368	3.6323	0.2620	4.4762	-0.0767	-1.2718	0.0143	0.2640
Phys. act. sometimes	0.1206	2.9570	0.1401	3.0489	0.2368	2.8252	0.0024	0.0423	0.0657	1.2547	0.0726	1.3655
Phys. act. often	0.0512	1.3540	0.0942	2.2144	0.0033	0.0427	-0.1236	-2.3715	-0.0228	-0.4588	0.0962	1.9870
Alone	-0.1153	-3.1260	-0.1384	-3.3860	-0.2820	-3.9365	-0.0338	-0.6778	-0.0478	-1.0382	-0.1038	-2.1369
Illness long	-0.1055	-3.2840	-0.3841	-10.6429	-0.3475	-5.2896	0.1393	3.1310	-0.0753	-1.8006	-0.0481	-1.1677

	model A		model B	
Threshold parameters	coeff	t-values	coeff	t-values
γ_s^1	1.0000	-		
γ_s^2	-0.0437	-1.2478		
γ_s^3	0.0327	1.4768		
γ_v^1	1.0000	-		
γ_v^2	-0.0045	-0.1843		
γ_v^3	-0.2018	-4.0735		
γ_o^1	0.0000	-	0.0000	-
γ_o^2	0.0000	-	0.0000	-
γ_o^3	0.1136	3.3127	0.1142	3.3303
Variances				
σ_s^2	1.0000	-	1.0000	-
σ_u^2	0.0000	-	0.1882	4.3567
σ_v^2	1.0000	-	1.0691	29.1000
σ_o^2	0.9195	43.0110	0.9197	43.0093
ρ	0.1485	8.1441	0.1555	8.3040
Vignette dummy	coeff	s.e.	coeff	s.e.
ϑ	0.7644	0.0192	0.9324	0.0799

Note: the estimated β_o parameters are similar in model A and B.

Table 15: Parameters estimates for cognition: delayed recall and vignette c1 - SHARE wave 2

	model A				model B							
	β_o		β_s		β_s		γ^1		γ^2		γ^3	
covariates	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values
Constant	2.4411	49.3476	3.1867	52.2536	3.1586	31.5578	1.0000	-	-0.1582	-1.9384	0.1973	4.6166
Belgium	-0.1062	-2.3255	-0.2120	-4.1046	-0.3742	-5.5750	-0.3445	-4.0998	0.0518	0.6407	0.2155	5.7417
Czech Rep.	-0.3369	-7.5129	-0.0336	-0.6596	-0.2133	-3.1722	-0.3213	-3.7240	0.0457	0.5666	0.1337	3.5700
Denmark	0.1354	3.1316	0.1959	3.8972	0.0342	0.5202	-0.3830	-4.0563	0.2236	2.8016	-0.0180	-0.4673
France	-0.0831	-1.3504	-0.1280	-1.8390	-0.0906	-0.9233	-0.2479	-2.2459	0.2436	2.5202	0.0878	1.7297
Italy	-0.1930	-3.7328	-0.1905	-3.2432	-0.2678	-3.4478	0.0315	0.3736	-0.0733	-0.9024	-0.0629	-1.3934
Netherlands	0.1834	3.3835	-0.0727	-1.1849	-0.3708	-4.6828	-0.4211	-3.4096	-0.2142	-1.5864	0.3133	7.2819
Poland	-0.6618	-12.4662	-0.2916	-4.9702	0.2011	2.6096	0.5460	7.1115	0.0912	1.3029	-0.2234	-4.4805
Spain	-0.3977	-6.9206	-0.1221	-1.8748	0.1418	1.6611	0.3687	4.1337	-0.0625	-0.7228	-0.0998	-1.8906
Sweden	0.2911	5.2144	-0.0336	-0.5310	-0.0421	-0.5296	-0.0561	-0.5852	0.1282	1.4961	-0.0957	-1.9709
Male	-0.2640	-10.5906	0.0513	1.8356	0.1471	4.0868	0.0513	1.1724	0.0814	2.0776	-0.0365	-1.7770
Age < 50	0.2284	2.9122	0.0247	0.2730	-0.0070	-0.0563	0.2222	1.7339	-0.1855	-1.3021	-0.1332	-1.8971
Age 50 to 55	0.1128	2.8898	0.0032	0.0702	-0.0594	-1.0283	-0.0875	-1.1664	0.0066	0.0934	0.0256	0.7841
Age 60 to 65	-0.0754	-1.9621	-0.0407	-0.9161	0.0279	0.4840	-0.0238	-0.3246	0.0699	1.0582	0.0351	1.0807
Age 65 to 70	-0.2345	-5.7467	-0.2114	-4.5401	-0.1480	-2.4500	-0.0647	-0.8522	0.1143	1.7083	0.0507	1.4684
Age 70 to 75	-0.4021	-9.0649	-0.3000	-6.0028	-0.1908	-2.8951	0.1196	1.6257	-0.0408	-0.5898	0.0549	1.5215
Age 75 to 80	-0.7283	-13.8470	-0.4562	-7.9610	-0.3793	-5.0295	0.0444	0.5356	0.0769	1.0226	-0.0362	-0.8283
Age > 80	-0.8983	-15.7206	-0.6857	-11.2571	-0.5630	-7.2060	0.0658	0.7495	0.0618	0.7969	0.0613	1.3525
Education low	-0.3362	-8.6113	-0.0812	-1.8662	0.0314	0.5615	0.1905	3.0807	-0.0494	-0.8594	-0.0559	-1.7016
Education high	0.3399	11.1246	0.1357	3.9216	0.0182	0.4004	-0.1321	-2.0307	-0.0299	-0.5175	0.0373	1.4710
Phys. act. sometimes	0.1405	4.5094	0.1324	3.7418	0.1648	3.5759	-0.0935	-1.6472	0.0837	1.6434	0.0628	2.4178
Phys. act. often	0.1491	5.0726	0.1442	4.2984	0.1478	3.3562	-0.1345	-2.5247	0.0838	1.7088	0.0689	2.7738
Alone	-0.0918	-3.1831	-0.0304	-0.9346	-0.0446	-1.0648	0.0474	0.9683	-0.0335	-0.7436	-0.0347	-1.4311
Illness long	-0.0793	-3.2128	-0.3964	-14.0716	-0.4376	-11.8630	0.0133	0.2909	0.0182	0.4364	-0.0559	-2.6984

	model A		model B	
Threshold parameters	coeff	t-values	coeff	t-values
γ_s^1	1.0000	-		
γ_s^2	-0.0137	-0.4878		
γ_s^3	0.1313	7.9145		
γ_v^1	1.0000	-		
γ_v^2	-0.0068	-0.2436		
γ_v^3	0.3927	29.2871		
γ_o^1	0.0000	-	0.0000	-
γ_o^2	0.0000	-	0.0000	-
γ_o^3	0.0697	2.5556	0.0707	2.5662
Variances				
σ_s^2	1.0000	-	1.0000	-
σ_u^2	0.0000	-	0.3793	19.1580
σ_v^2	1.0000	-	0.7526	43.3007
σ_o^2	0.9056	53.3265	0.9059	52.7075
ρ	0.2059	14.5904	0.2064	13.5226
Vignette dummy	coeff	s.e.	coeff	s.e.
ϑ	2.8395	0.0292	2.6044	0.0226

Note: the estimated β_o parameters are similar in model A and B.

Table 16: Parameters estimates for breathing and vignette b1 - SHARE wave 2

	model A				model B							
	β_o		β_s		β_s		γ^1		γ^2		γ^3	
covariates	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values
Constant	-3.3664	-12.7142	2.5898	6.2115	2.5153	4.4095	1.0000	-	-0.4417	-1.1086	-0.0238	-0.0561
Belgium	0.0054	0.1407	-0.0771	-1.3064	-0.1626	-2.0592	-0.2811	-6.0521	0.1526	2.6002	0.1320	2.2526
Czech Rep.	-0.1253	-3.3432	-0.1704	-2.9887	-0.0298	-0.3946	-0.1566	-3.5598	0.2678	4.9400	0.1702	2.7309
Denmark	0.1643	4.5716	0.3585	5.9521	0.3817	4.7105	-0.0147	-0.3612	0.0626	1.1145	-0.0242	-0.3765
France	0.1121	2.0926	-0.1958	-2.4227	-0.0433	-0.3787	-0.1121	-1.7956	0.3365	4.5853	0.0129	0.1489
Italy	-0.2821	-6.2722	0.3436	4.5790	0.2867	3.0072	-0.1001	-1.9120	0.0213	0.2953	0.0176	0.2384
Netherlands	0.1070	2.4088	-0.0678	-0.9649	0.1391	1.4044	-0.2269	-4.2737	0.3362	5.2671	0.2317	3.2362
Poland	-0.1349	-3.1031	-0.0290	-0.4341	0.1443	1.6362	0.4526	9.3377	-0.1358	-1.8891	-0.4950	-5.4849
Spain	-0.0568	-1.1530	0.2254	2.8256	0.7031	6.2292	0.5760	10.1828	-0.1135	-1.2750	-0.0805	-0.7951
Sweden	0.4812	10.4635	-0.0493	-0.6943	0.0440	0.4657	0.2539	5.0080	-0.2915	-3.6533	0.0099	0.1349
Male	0.6207	22.5658	-0.0537	-1.2430	0.0172	0.3006	0.1554	4.8821	-0.0658	-1.5785	-0.0413	-0.9248
Age < 50	0.1732	2.6571	0.1387	1.2080	0.0991	0.6638	-0.0378	-0.4862	-0.0642	-0.5979	0.0313	0.2786
Age 50 to 55	0.1167	3.6107	0.0357	0.6662	0.0340	0.4735	0.0424	1.1417	-0.0813	-1.5781	0.0142	0.2593
Age 60 to 65	-0.0750	-2.3433	-0.0321	-0.6192	0.0057	0.0828	0.0170	0.4595	0.0058	0.1170	0.0233	0.4456
Age 65 to 70	-0.2057	-5.9537	-0.1168	-2.1314	-0.0350	-0.4703	0.0198	0.4976	0.0642	1.2426	0.0307	0.5321
Age 70 to 75	-0.4180	-11.3233	-0.1935	-3.3692	-0.1090	-1.4111	0.0219	0.5105	0.0547	0.9879	0.0380	0.6142
Age 75 to 80	-0.5698	-13.0567	-0.3659	-5.5591	-0.3324	-3.7820	-0.0461	-0.9050	0.1405	2.2482	-0.0286	-0.3774
Age > 80	-0.6975	-14.7455	-0.3919	-5.6047	-0.3192	-3.4357	-0.0213	-0.3899	0.1762	2.7318	-0.0748	-0.9607
Education low	-0.0676	-2.0430	-0.0597	-1.1601	-0.1541	-2.2868	-0.0128	-0.3327	-0.0657	-1.3155	-0.0439	-0.8529
Education high	0.1662	6.6719	0.1357	3.3218	0.1679	3.0325	0.0318	1.1153	-0.0095	-0.2474	0.0133	0.3177
Phys. act. sometimes	0.1028	3.9474	0.2252	5.5465	0.2191	3.9390	-0.0752	-2.4322	0.0325	0.8315	0.0589	1.3930
Phys. act. often	0.1916	7.7935	0.3470	8.8164	0.3722	6.9026	0.0030	0.1065	-0.0373	-0.9907	0.0647	1.5725
Alone	-0.0763	-3.1565	-0.0281	-0.7489	-0.0423	-0.8447	-0.0014	-0.0504	-0.0005	-0.0141	-0.0099	-0.2456
Illness long	-0.0726	-3.5133	-0.6058	-18.4179	-0.5863	-12.6973	0.0494	2.0463	0.0223	0.7140	-0.0290	-0.8259
Height/100	1.8753	11.7954	0.3818	1.5304	0.6843	2.0849	0.2424	1.3546	0.0671	0.2817	-0.0795	-0.3129

	model A		model B	
Threshold parameters	coeff	t-values	coeff	t-values
γ_s^1	1.0000	-		
γ_s^2	-0.2801	-7.0409		
γ_s^3	-0.1842	-7.5494		
γ_v^1	1.0000	-		
γ_v^2	0.0668	4.0186		
γ_v^3	0.2380	10.3863		
Variances				
σ_s^2	1.0000	-	1.0000	-
σ_u^2	0.0000	-	0.2697	10.7015
σ_v^2	1.0000	-	0.6292	29.7016
σ_o^2	0.7940	113.0374	0.7940	155.2211
ρ	0.1775	11.4724	0.1843	12.5227
Vignette dummy	coeff	s.e.	coeff	s.e.
ϑ	1.5443	0.0165	1.9037	0.2995

Note: the estimated β_o parameters are similar in model A and B. Because the objective measure is a continuous variable there are no thresholds in the objective part of the model.

Table 17: Parameters estimates for mobility and vignette m1 - SHARE wave 2

	model A				model B							
	β_o		β_s		β_s		γ^1		γ^2		γ^3	
covariates	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values	coeff	t-values
Constant	10.2631	37.6973	0.5567	8.6877	0.5278	7.5545	1.0000	-	-0.0655	-1.0767	-0.2171	-3.0472
Belgium	0.6714	2.3664	-0.1666	-2.5368	0.2014	2.4264	0.2986	4.7167	0.1871	3.2578	0.0311	0.4253
Czech Rep.	-0.2514	-0.9018	0.3357	5.4664	0.4647	5.6770	-0.0636	-0.9899	0.3143	5.7693	0.1802	2.8651
Denmark	-1.1451	-4.4328	-0.6087	-9.5125	-0.3782	-4.6748	0.2504	4.2667	-0.0714	-1.2049	0.0150	0.2410
France	-0.0859	-0.2142	-0.8509	-7.7109	-0.5834	-4.1251	0.2952	3.0739	-0.1426	-1.3923	0.1333	1.3277
Italy	1.3363	3.9144	-0.3573	-4.3539	0.0216	0.2052	0.3716	4.9814	-0.0043	-0.0589	0.0090	0.1017
Netherlands	1.1338	3.6045	-0.0282	-0.3917	0.0712	0.7475	0.0238	0.3190	0.2212	3.5025	-0.0002	-0.0026
Poland	0.0459	0.1382	-0.0906	-1.2030	0.0299	0.3091	0.2545	3.5702	-0.3348	-4.2640	-0.1111	-1.4694
Spain	1.6303	4.3697	-0.2452	-2.7699	-0.1751	-1.5432	0.1327	1.5433	-0.2436	-2.5721	-0.0002	-0.0024
Sweden	-1.1432	-3.2972	-0.2144	-2.6251	-0.1323	-1.2565	0.1525	1.8293	-0.0924	-1.1383	-0.3637	-4.1083
Male	-0.7236	-4.7555	-0.0937	-2.5967	-0.1783	-3.8769	-0.0782	-2.2424	-0.0021	-0.0664	-0.0370	-1.0054
Age < 50	-1.2467	-2.8404	-0.1805	-1.6345	-0.0506	-0.3694	0.2027	2.1583	-0.1692	-1.6820	-0.2209	-1.9159
Age 50 to 55	-0.3599	-1.6331	-0.0201	-0.3804	-0.0259	-0.3836	0.0021	0.0417	-0.0414	-0.8640	-0.0225	-0.4182
Age 60 to 65	0.4910	2.2239	0.0043	0.0827	0.0364	0.5491	0.0061	0.1241	0.0635	1.4061	0.0220	0.4070
Age 65 to 70	0.5654	2.3730	0.0626	1.1223	0.0861	1.2346	0.0023	0.0435	0.0313	0.6331	0.0303	0.5326
Age > 70	1.6823	6.4139	0.1025	1.6852	0.1037	1.3221	-0.0589	-0.9713	0.1240	2.3715	0.1098	1.7654
Education low	0.1571	0.6012	0.0669	1.0808	0.0579	0.7290	0.0109	0.1932	-0.0245	-0.4645	-0.0050	-0.0759
Education high	-0.5882	-3.3274	-0.1939	-4.4970	-0.2900	-5.2604	-0.0916	-2.1532	-0.0016	-0.0408	-0.0274	-0.6376
Phys. act. never	1.1375	6.3417	0.2271	5.3642	0.2906	5.4802	0.0780	1.9305	-0.0569	-1.5281	0.0634	1.4671
Phys. act. sometimes	0.5761	3.0319	0.0178	0.3911	0.0280	0.4708	-0.0015	-0.0345	0.0297	0.7518	0.0451	0.9736
Alone	0.1499	0.8046	-0.0182	-0.4156	0.0111	0.2005	0.0416	1.0067	-0.0312	-0.8259	-0.0780	-1.7212
Illness long	0.7989	5.2136	0.8040	22.0554	0.8067	17.0687	-0.0272	-0.7792	-0.0025	-0.0786	0.1001	2.6738

	model A		model B	
Threshold parameters	coeff	t-values	coeff	t-values
γ_s^1	1.0000	-		
γ_s^2	-0.0827	-3.2678		
γ_s^3	-0.0431	-1.1114		
γ_v^1	1.0000	-		
γ_v^2	0.3239	16.0719		
γ_v^3	0.1102	5.4864		
Variances				
σ_s^2	1.0000	-	1.0000	-
σ_u^2	0.0000	-	0.1846	4.8206
σ_v^2	1.0000	-	0.6773	34.1089
σ_o^2	5.1187	97.8321	5.1187	181.8809
ρ	0.1311	8.1014	0.1323	7.3719
Vignette dummy	coeff	s.e.	coeff	s.e.
ϑ	2.4606	0.0272	2.1562	0.0649

Note: the estimated β_o parameters are similar in model A and B. Because the objective measure is a continuous variable there are no thresholds in the objective part of the model.

A Self-assessment and vignette questions

Here the self-assessment and vignette questions of both SHARE waves are given. The vignette asked in wave 2 is chosen out the three asked in the first wave and is always the first vignette given here. The vignette of wave 2 is slightly different from the one of wave 1 since the last sentence differs. In all cases the possible answer categories are 'none', 'mild', 'moderate', 'severe', and 'extreme'.

Cognition

Self-assessment question

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

Vignettes

c1: (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. Overall in the last 30 days, how much difficulty did (Lisa) have with concentrating or remembering things? (Wave 2: In your opinion, how much difficulty does (Lisa) have with concentrating or remembering things?)

c2: (Sue) is keen to learn new recipes but finds that she often makes mistakes and has to reread several times before she is able to do them properly. Overall in the last 30 days, how much difficulty did (Sue) have with concentrating and remembering things?

c3: (Eve) cannot concentrate for more than 15 minutes and has difficulty paying attention to what is being said to her. Whenever she starts a task, she never manages to finish it and often forgets what she was doing. She is able to learn the names of people she meets.

Overall in the last 30 days, how much difficulty did (Eve) have with concentrating or remembering things?

Objective measures

- Immediate recall: Now, I am going to read a list of words from my computer screen. We have purposely made the list long so it will be difficult for anyone to recall all the words. Most people recall just a few. Please listen carefully, as the set of words cannot be repeated. When I have finished, I will ask you to recall aloud as many of the words as you can, in any order. Is this clear?

- Delayed recall: A little while ago, I read you a list of words and you repeated the ones you could remember. Please tell me any of the words you can remember now? (This question is directly asked after the final question assessing numerical ability.

- Verbal fluency: Now, I would like you to name as many different animals as you can think of. You have one minute to do this.

- Numeracy: Next, I would like to ask you some questions which assess how people use numbers in everyday life.

    1) If the chance of getting a disease is 10 per cent, how many people out of 1,000 would be expected to get the disease?

    2) If the respondent's answer to question 1 was incorrect, the next question is: In a sale, a shop is selling all items at half price. Before the sale, a sofa costs 300 (local currency). How much will its cost in the sale?

    3) If the respondent's answer to question 1 was correct, the next question is: A second hand car dealer is selling a car for 6,000 (local currency). This is two-thirds of what it costs new. How much did the car costs new?

4) If the answer to question 3 is correct, the next question will be: Let us say you have 2,000 (local currency) in a savings account. The accounts earns ten per cent interest each year. How much will you have in the account at the end of two years?

Breathing

Self-assessment question

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

Vignettes

b1: (Mark) has no problems with walking slowly. He gets out of breath easily when climbing 20 meters uphill or a flight of stairs. In the last 30 days, how much of a problem did Mark have because of shortness of breath? (Wave 2: How much of a problem does (Mark) have because of shortness of breath?)

b2: (Paul) suffers from respiratory infections about once every year. He 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. In the last 30 days, how much of a problem did (Paul) have because of shortness of breath?

b3: (Henri) has been a heavy smoker for 30 years and wakes up with a cough every morning. He gets short of breath even while resting and does not leave the house anymore. He often needs to be put on oxygen. In the last 30 days, how much of a problem did (Henri) have because of shortness of breath?

Objective measure (peak-flow test)

- The next test that I am going to ask you to perform will measure how fast you can expel air from your lungs. It is important that you blow as hard and as fast as you can. I would

like you to perform the test two times. When we are ready to begin, I will ask you to stand up. Take as deep a breath as possible. Open your mouth and close your lips firmly around the outside of the mouthpiece, and then blow as hard and as fast as you can into the mouthpiece

Mobility

Self-assessment question

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

Vignettes

m1: (Rob) is able to walk distances of up to 200 meters without any problems but feels tired after walking one kilometer or climbing more than one flight of stairs. He has no problems with day-to-day activities, such as carrying food from the market. Overall in the last 30 days, how much of a problem did (Rob) have with moving around? (Wave 2:In your opinion, how much of a problem does Rob have with moving around?)

m2: (Kevin) does not exercise. He cannot climb stairs or do other physical activities because he is obese. He is able to carry groceries and do some light household work. Overall in the last 30 days, how much of a problem did (Kevin) have moving around?

m3: (Tom) has a lot of swelling in his legs due to his health condition. He has to make an effort to walk around his home as his legs feel heavy. Overall in the last 30 days, how much of a problem did (Tom) have moving around?

Objective measure

- The next test measures the strength and endurance in your legs. I would like you to fold your arms across your chest and sit so that your feet are on the floor; then stand up keeping

your arms folded across your chest. The respondent is then asked whether s/he thinks it is safe to stand up five from a chair five times without using their arms. When the answer is affirmative the respondent is asked to do the test and the interviewer records the time (in seconds) used for five stands.

B Description of variables

Table 18: Description of covariates

Covariate	Description
Age	Depending on the wave, each respondent's age is calculated as 2004 or 2007 minus the year or birth, which information is provided by SHARE. We include dummies for age groups.
Education	Based on the International Standard Classification of Education (ISCED 97) we categorize education in three dummies: low, middle and high. We define ISCED levels 0 and 1 as low education, 2 and 3 as middle and 4,5 and 6 as high education. The lowest educational group is used as the reference group in the analyzes.
Gender	We include a dummy for being male.
Not alone	SHARE contains information on the marital status of its respondents. Possible answers are: (1) married and living together with spouse, (2) registered partnership, (3) married, living separated from spouse, (4) never married, (5) divorced, (6) widowed. We include a dummy for whether a respondent is living alone or not, where we define "not living alone" if marital status is reported as either (1) or (2).
Physical activity	SHARE contains information on the frequency of physical activity, such as sports or heavy housework, of its respondents. Possible answers are (1) more than once a week, (2) once a week, (3) one to three times a month, (4) hardly ever, or never. We define that a respondents is engaged *often* in physical activity if the answer is (1), *sometimes* if the answer is (2) or (3) and *never* if the answer is (4). The group of respondents that reports to be engaged in physical activity "hardly ever or never engaged" is taken as the reference group.
Illness long	The SHARE questionnaire contains the following question: "Some people suffer from chronic or long-term health problems. By long-term we mean it has troubled you over a period of time or is likely to affect you over a period of time. Do you have any long-term health problems, illness, disability or infirmity?" The answer can be either yes or no.

References

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

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

Gupta, N. D., Kristensen, N., & Pozzoli, D. (2008). *External validation of the use of vignettes in cross-country health studies.* IZA Discussion paper No. 3989.

Herzog, A. R., & Wallace, R. B. (1997). Measures of cognitive functioning in the AHEAD study. *Journal of Gerontology Series B*, *52B*, 37-48.

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

Jürges, H., & Winter, J. (n.d.). *Are anchroing vignettes ratings sensitive to vignette age and sex.*

Kapteyn, A., Smith, J. P., & Van Soest, A. (2007). Vignettes and self-reports of work disability in the united states and the netherlands. *American Economic Review*, *97*, 461 - 473.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*, 191-207.

Llewellyn, D. J., Lang, I. A., Langa, K. M., & Huppert, F. A. (2008). Cognitive function and psychological well-being: findings from a population-based cohort. *Age and Ageing*, *37*(6), 685-689.

Mehta, K. M., Yaffe, K., Langa, K. M., Sands, L., Whooley, M. A., & Covinsky, K. E. (2003). Additive effects of cognitive function and depressive symptoms on mortality in elderly community-living adults. *Journal of Gerontology Series A*, *58A*, 461-467.

Nunn, A. J., & Gregg, I. (1989). New regression equations for predicting peak expiratory flow in adults. *British Medical Journal*, *298*(6680), 1068-1070.

Quanjer, P. H., Lebowitz, M., Gregg, I., Miller, M., & Pedersen, O. (1997). Peak expiratory flow: conclusions and recommendations of a working party of the European Respiratory Society. *European Respiratory Journal*, *24*, 2S-8S.

Sen, A. (2002). Health: perception versus observation. *British Medical Journal*, *324*(7342), 860-861.

Van Soest, A., Delaney, A., Harmon, C., Kapteyn, A., & Smith, J. P. (2007). *Validating the use of vignettes for subjective threshold scales.* IZA Discussion paper No. 2860.