Netspar

Network for Studies on Pensions, Aging and Retirement

Netspar THESES

Sandra Vriend

Testing the Vignettes Method

Correcting Self-Reported Health Measures

MSc Thesis 2010

Testing the Vignettes Method

Correcting Self-Reported Health Measures

Thesis MSc Economics

Author:
Sandra Vriend
Student number:
1620479

Supervisors:
Prof. dr. Bas van der Klaauw
Prof. dr. Maarten Lindeboom

Vrije Universiteit Amsterdam
Faculty of Economics and Business Administration
Department of Economics

Date:
July 2010

Abstract

The vignettes method is increasingly used to correct for reporting heterogeneity present in self-reported health measures. The general idea is that the reporting behaviour of a specific respondent can be anchored using the way in which that respondent evaluates the health status of hypothetical vignette persons. After anchoring the reporting behaviour of each individual, reporting heterogeneity is corrected for and thus cleansed (self-reported) health measures can be obtained. However, the validity of this vignettes method rests on the two crucial assumptions of vignette equivalence and response consistency. Violation of these assumptions may reduce the power of the vignettes method as a correction tool. In this thesis we test for the validity of this method by running regressions in which an outcome measure that is expected to be affected by true health in the end is regressed on a number of individual characteristics, the corrected health measures obtained from the vignettes method and the uncorrected observed self-reports. Evaluating the relative significance of the corrected and uncorrected health measures then allows one to assess the quality of using anchoring vignettes as a tool to correct for reporting heterogeneity. We do not find (strong) positive results on the validity of the vignettes method. While a large number of our test regressions are inconclusive, we do find negative results in a considerable amount of test regressions. The most positive result that we find is both corrected and uncorrected health measures having a similar significance in the test regressions. Our results do not reject the vignettes method as a correction tool per se, but it seems to be that one should be cautious in using the method. Further research is needed to improve the vignettes method thereby enhancing health measurement.

Keywords: (Anchoring) vignettes, vignettes approach, health measurement, self-reported health measures, reporting heterogeneity, response category cut-point shift

Acknowledgements

First, I would like to thank my supervisors Bas van der Klaauw and Maarten Lindeboom for their excellent supervision, inspiring advice and suggestions and helpful comments. They were always willing to help me if I had questions or encountered problems in doing my research. I would also like to thank Henri de Groot, not only for being interested in this thesis but also for the pleasant and useful conversations that I have had with him during my study. Finally, a special word of thanks to my family and friends who have always supported me while writing this thesis.

"This paper uses data from SHARE release 2.3.0, as of November 13th 2009. 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)."

Table of contents

Abstract **2**

Acknowledgements **2**

1. Introduction **5**

2. The usage of health measures in economics **6**

3. The problems with measuring health **8**

3.1 Problems with self-reported health measures 8

3.2 Resulting biases in estimation procedures 12

3.3 Potential solutions to the problems with self-reported health measures 13

4. The vignettes approach **16**

4.1 What are anchoring vignettes? 16

4.2 Correcting for DIF using the vignettes method 17

5. Models for self-reported health measures **20**

5.1 Ordinal data, ordered response models and ordered probit models 20

5.2 Incorporating the vignettes method in statistical models 21

6. Potential problems with the usage of vignettes **25**

7. Data description **28**

7.1 The SHARE data set 28

7.1.1 The individual CAPI modules 29

7.1.2 The drop-off questionnaire 31

7.1.3 The vignette questionnaire 31

7.1.4 The generated variables modules 32

7.2 Constructing a master dataset 33

7.2.1 The sample 33

7.2.2 Data selection and construction of variables 34

7.3 Data description and descriptive statistics 36
7.3.1 A description of the vignette evaluations 36
7.3.2 Covariates and socio-economic variables 50
7.3.3 Objective health measures and health limitations 55

8. Testing the appropriateness of the vignettes approach 57
8.1 Generating corrected health measures 57
8.1.1 A modified version of the HOPIT model 58
8.1.2 Estimating the modified version of the HOPIT model 60
8.1.3 A simulation exercise 66
8.1.4 Corrected latent health 67
8.2 The testing procedure 72
8.3 Test results 75
8.3.1 The outcome variables in the test regressions 75
8.3.2 Estimated regression equations 76
8.3.3 Conclusions 80

9. Conclusion 82

References 84

Appendices 88
Appendix 1 Objective health measures and health limitations 88
Appendix 2 HOPIT estimation results 90
Appendix 3 The simulation exercise 96
Appendix 4 Estimation results for the test regressions 98

1. Introduction

In various fields of study self-assessments on subjective measures are used. In this respect one could think of self-assessments on work disability (Kapteyn, Smith & Van Soest, 2009), health, political efficacy, life satisfaction and job satisfaction (Bago d'Uva et al., 2009). What all these self-assessments have in common is that the evaluations of individuals may be subjective. The subjectivity of these assessments may cause incomparability problems. Self-assessments may be incomparable between individuals, over time, across countries and across different groups within a population or country (e.g., Murray et al., 2002). Several factors may account for these comparability problems. For instance socio-economic status of individuals may affect the way in which a particular outcome is evaluated (Bago d'Uva et al., 2008). In addition, gender, age and educational attainment of individuals may play a role in reporting behaviour (Murray et al., 2002). Moreover, socio-demographic factors and culture may influence the way in which individuals evaluate their own status (Tandon et al., 2002). Self-reports are also found to suffer from some other limitations; these will be discussed in detail in the sections to come. Various attempts have been undertaken to solve these problems with self-reported measures. Sometimes more objective measures of the outcome under consideration have been used in order to obtain more reliable results. But these objective measures may suffer from other limitations, as Bound (1991) describes for health measures in particular. Another attempt to correct for the problems with self-assessments is the use of the so-called vignettes method. But, no consensus has been reached yet on the quality of this method for correcting self-reported measures.

The purpose of this thesis is to test whether the vignettes approach is an appropriate method to use in order to correct for differences in reporting behaviour in the case of self-reported health measures. The problem statement that will be covered is:

> *How appropriate is the usage of vignettes for correcting self-reported health measures so as to obtain reliable measures for the health status of individuals that can be used in studying effects of health on various outcomes?*

This problem statement is divided into the following four sub questions:

1. *What are the problems associated with using self-reported health measures of individuals?*
2. *What is the vignettes approach and how does it try to correct self-reported health measures?*
3. *What are the potential problems of using the vignettes approach for determining the true health status of individuals?*
4. *How can we test the appropriateness of the vignettes approach for correcting self-reported health measures?*

The first sections of this thesis explain the theoretical framework in the light of which the vignettes method will be tested. Section 2 briefly goes into the different usages of health measures in economics. Thereafter, section 3 discusses the problems associated with measuring health, which were the ones initiating the application of the vignettes method to the field of health and labour economics. Section 4 presents the vignettes method. Modelling self-reported health measures and including the vignettes therein is the topic of section 5. Section 6 examines potential problems of the vignettes method that may deteriorate the quality of the corrected health measures. Section 7 describes the data that will be used in testing the vignettes method. Subsequently, section 8 discusses the empirical tests and their results. Finally, section 9 concludes.

2. The usage of health measures in economics

The usage of health measures in economics is widespread. There are some hotly debated topics, for example in the light of the current ageing of the society, that make adequate measurement of the health status of individuals important. Studies that use health measures adopt either objective health indicators or more subjective self-assessments of the health status or both of them. Regarding objective health measures one could think of medical examinations; in case of self-reported health measures an individual is asked to rate his own health on some (discrete) scale. These self-reports can be on a specific domain of health, like mobility or cognition, or they can cover the general health status. This section discusses the different usages of various types of health measures.

Health measures can be informative on a lot of different kinds of outcomes. For instance, Bago d'Uva et al. (2008) refer to self-reported health measures that are found to be a pretty good predictor of future mortality (e.g., Idler & Benyamini, 1997). Also, the utilization of medical care and health services is determined by the health status of individuals (Butler et al., 1987). In the light of the ageing of the society it is interesting to look at the effects of an ageing society on morbidity and mortality and therefore on medical care use.

Another important field of study in which health measures are used is (socio-economic) inequality and inequity in care use. Inequity in health care delivery has for instance been studied by Wagstaff & Van Doorslaer (2000)[1]. Inequality in care use does not need to point at inequity per se; to conclude that inequity exists one has to look at the underlying source of the inequality in care use. One has to determine to what extent inequality in care use is due to differences in the socio-economic status of individuals and to what extent this inequality is the result of differences in the true health status of individuals. On the one hand, it may be that individuals with the same health status but from different socio-economic backgrounds do not get equal care, which shows true inequality and inequity in care use. On the other hand, it may be that individuals from different socio-economic groups just have different levels of health. Then, differences in care use do not represent inequity. To see which of these two approaches underlies the potential inequality, one has to correct for the needs of various individuals, i.e., look at needs-standardized care use. Depending on which of these two approaches is the true cause of inequality in care use, policy implications may be very different. Nevertheless, for both strands of reasoning appropriate measurement of the health status is important for the accuracy of the conclusions reached.

Related to this, a large body of literature has evaluated the relationship between self-reported health and socio-economic status over time. For instance, past socio-economic status and past health shocks incurred during childhood are found to affect the health situation of an individual later in life (Van den Berg & Lindeboom, 2007). The results on the relationship between past socio-economic status on health later in life and, the other way around, the effect of past health shocks on current socio-economic status may also be affected by the accuracy of the health measures used. In this case, comparability problems of self-reported health measures over time come into play. When people adapt their valuation of health over time, self-reported health measures during childhood are not comparable to self-reported health measures of the elderly. One has to bear this in mind while working with self-reported health measures.

[1] A broad literature on inequity in health care utilization has developed. See Wagstaff & Van Doorslaer (2000) for other relevant references.

Next to the applications of health measures in studies on medical care use and socio-economic inequality, there is a wide range of applications in labour supply and retirement models. A broad literature has developed looking at the relative importance of financial incentives and health status in the decision to retire. Dwyer & Mitchell (1999) indicate that the health status of an individual can significantly influence among others the preferences for work and the time horizon that an individual faces. Therefore health is an important factor in the decision to retire. Not only can early retirement be induced by bad health, it is also possible that financial incentives in the retirement system are such that early retirement becomes attractive. However, no consensus has been reached yet on the relative importance of financial incentives and health in the retirement decision. The extent to which health is found to affect the retirement decision may depend on the health measure that is used. Bound et al. (1999) show that health is indeed an important determinant of labour force exit and early retirement. Bound, Stinebrickner & Waidmann (2007) and Banks, Emmerson & Tetlow (2007) argue that there is a possible interaction between health and financial incentives in the decision to retire: workers suffering from deteriorating health may want to retire, although they will only be able to do this when they have enough financial resources available. This would then imply that financial incentives have differential effects on individuals in poor health and individuals in good health.

In labour supply models there is also an important effect of the incentives in the social security system, or more specifically the disability insurance system, on the labour supply decisions of individuals. Social security can be a financially attractive option for individuals, and this might encourage them to report that they have health problems limiting the amount of work they can do (e.g., Gruber & Wise, 1998; Garcia-Gomez, von Gaudecker & Lindeboom, 2009). We will come back to this type of reporting behaviour in the next section since this is one of the major problems associated with the usage of self-reported health measures.

To conclude, health measures are thus applied in a wide range of topics ranging from the utilization of medical care and health services to the effect on retirement and labour supply. Since adequate estimates of the role of health on these different outcomes do critically depend on the appropriateness of the health measures used, it is important to look at the potential problems associated with both objective as well as subjective health measures. This will be the topic of the next section.

3. The problems with measuring health

The incorporation of health in the various applications discussed in the previous section is hindered by a considerable amount of difficulties. First, health may be an endogenous variable in labour supply models since health and work are jointly determined (Lindeboom, 2005). In addition, self-reported health may be determined by an individual's labour market status (e.g., Bound, 1991; Kerkhofs & Lindeboom, 1995). There are also a lot of other problems distinguished affecting the quality of self-reports. The most important difficulties with measuring health will be the topic of this section.

3.1 Problems with self-reported health measures

In a survey individuals are often asked to rate their own health status using some ordered categorical response scale (Salomon, Tandon & Murray, 2001). A typical survey question designed to determine the general health status of an individual looks like "How would you rate your health?" (Bound et al., 1999; Lindeboom, 2005) or "How is your health in general?" (Lindeboom & Van Doorslaer, 2004). These questions are then answered on a categorical response scale. This scale usually consists of five response categories like "excellent", "very good", "good", "fair" or "poor" (Crossley & Kennedy, 2002). A response scale ranging from "very good" to "very poor" is also possible (Lindeboom & Van Doorslaer, 2004). Which response scale is used differs somewhat across surveys.

The previous questions on health ask the individual for an evaluation of his general health status. However, as has been mentioned by Lindeboom (2005), general health is probably not an appropriate measure to use in labour supply models; work related health may do a better job. Related to that, Murray et al. (2000) indicate that the multidimensionality of true health may also leave room for discussion on which domains of health are important to measure in order to approximate the actual health status correctly. They state that there are several domains that affect everyone's conception of the health status, among others mobility, hearing, vision, pain, affect, cognition and communication. Therefore, next to the aforementioned general health questions, surveys often evaluate some specific health domains, like mobility, pain and cognition, separately on a similar categorical response scale. Furthermore, respondents may be asked to indicate whether they suffer from some work limiting health problem *("Does your health limit you in the kind or amount of work that you can do?"* (Lindeboom, 2005, p. 4)) or whether they have difficulties with certain activities[2] (e.g., Murray et al., 2002).

However, these self-reported health measures are not perfect; they measure the underlying health status of individuals with some error. For instance, Butler et al. (1987) state that self-assessed health measures perceived health, which may differ from actual health. According to Sadana et al. (2002) such differences between perceived and true health reflect among others variations in cultural and gender norms, knowledge and information. Besides, they state that self-reported health (i.e., the health that individuals report in an interview) may differ from perceived health (i.e., the health status based on the knowledge and beliefs of an individual).

Also, Bound, Stinebrickner & Waidmann (2007) state that the usage of self-rated health suffers from serious problems. First, self-reported health measures are discrete, while the underlying

[2] For such questions the five point response scale may look like "no difficulty", "mild difficulty", "moderate difficulty", "severe difficulty", and "extreme/cannot do" (e.g., Murray et al., 2002).

true health status is a continuous measure. Secondly, the major problem with self-reported health measures is that individuals differ in their valuation of health. This implies that two individuals with exactly the same true health status may differ in their self-reported health levels. So self-reported health measures are a subjective judgement which makes it difficult to compare the outcomes between individuals (Bound et al., 1999). This incomparability of self-reported health across populations or sub populations has received different terminology in the literature. Lindeboom & Van Doorslaer (2004) mention various terminology for these comparability problems, for instance: "*state dependent reporting bias*" (Kerkhofs & Lindeboom, 1995), "*response category cut-point shift*" (e.g., Murray et al., 2002), "*reporting heterogeneity*" (e.g., Shmueli, 2003) or "*differential item functioning (DIF)*" (e.g., King et al., 2004).

State-dependent reporting bias is an endogeneity problem of self-reported health measures. This is found in particular in labour supply and retirement models. The term state-dependent reporting refers to the fact that differences in labour market status may affect the health level that an individual reports. The reasoning is that the labour market status of an individual and whether he is enrolled in disability insurance may determine whether he reports himself being in poor health or not. Also, individuals who do not receive early retirement benefits may have a financial incentive to report poor health or the existence of work limitations in order to be eligible for receiving disability insurance benefits (Bound et al., 1999; Bound, Stinebrickner & Waidmann, 2007). Thus health and work are interrelated, i.e., self-reported health and labour market status are not completely independent of each other (Lindeboom, 2005). Therefore health has to be treated endogenously in labour supply models.

A specific form of state-dependent reporting is found in the so-called *justification bias* (Kapteyn, Smith & Van Soest, 2009): individuals who do not participate in the labour market may rationalize their inactivity by reporting the existence of health problems (e.g., Bound, 1991; Bound et al., 1999, Lindeboom, 2005). When individuals use health as a justification for leaving the labour force early, subjective health assessments may measure the preference for leisure instead of the true health status (Dwyer & Mitchell, 1999). Dwyer & Mitchell, referring for instance to Chirikos & Nestel (1984) and Anderson & Burkhauser (1985), indicate that the existence of such a bias has been confirmed by several studies in the past. On the other hand, Dwyer & Mitchell (1999) mention that other studies, for instance Stern (1989), have shown that self-reported measures of health could be used. The existence of justification bias in self-reported health measures may cause self-reports not to measure the actual health status of a person accurately. These measures then do not capture the true effect of health on, in this case, labour supply or retirement decisions. This results in an inappropriate, biased estimate of the effect of health, as will be explained further in section 3.2.

More generally, the problem with the comparability of self-reported health measures is termed differential item functioning (DIF), response category cut-point shift or reporting heterogeneity as has been mentioned before. This terminology will be used interchangeably. Categorical response scales for measuring health have a number of response thresholds. These are the levels of health for which an individual will shift from one response category to another (Murray et al., 2000). Differential item functioning means that individuals may differ in the response scales that they use while assessing their own health status. So, different individuals may use different threshold levels (often called *cut-points* in the literature) despite having the same true health status. Thus, while a specific actual health level may be translated as "good" health by one individual, another individual

with the same actual health may rate his health only as "fair" since he has a higher threshold value for being in good health. The answers to the discrete self-assessed health question are then incomparable across individuals (e.g., Bago d'Uva et al., 2008; Kapteyn, Smith & Van Soest, 2007). As long as this variation in reporting behaviour is random, it may not be that much of a problem. However, systematic differences in reporting behaviour are much more of a concern (Bago d'Uva et al., 2008). According to Bago d'Uva et al. (2008) differences in health inequalities according to self-reported and objective measures of health point at the possibility of systematic differences in reporting behaviour. Such comparability problems may exist both within populations, for example between different socioeconomic groups, as well as across different populations or countries. Besides these same problems may exist over time, since expectations for health can change over time resulting in different response category cut-points over time (Salomon, Tandon & Murray, 2004; Voňková & Hullegie, 2010).

These comparability issues may result when individuals understand health questions differently and use the available response scales in a different manner (Salomon, Tandon & Murray, 2001). So there is "*individual variation in the mapping from an unobserved continuous latent scale into a set of discrete categorical responses*" (Salomon, Tandon & Murray, 2001, p. 2). On the level of populations, Murray et al. (2000) mention the role of cultural expectations for domains of health in determining the response category cut-points of individuals from different populations. Factors influencing threshold variation within populations or cultural groups may be age[3], gender, language and experiences with being ill (Lindeboom & Van Doorslaer, 2004). Language may play a role since a word, despite of being translated in a correct manner, may have a different connotation in one language as compared to the meaning it has in another language (Jürges, 2007). Besides, reporting heterogeneity between individuals may stem from differences in education levels (Bago d'Uva, O'Donnell & Van Doorslaer, 2008). Finally, Bago d'Uva et al. (2008) indicate that financial incentives to report poor health, the extent to which survey questions are comprehensible, the conceptions of health in general and the expectations for own health may cause reporting heterogeneity to exist. The way in which differences in expectations of health may affect reporting behaviour is illustrated by Murray et al. (2000). They mention the example of an individual with a chronic disease who rates his mobility as excellent when he is able to walk around the house, while an athlete will probably only report excellent mobility when he is able to run 10 kilometres. The perception of health by an individual may therefore differ significantly from his actual health because of the role that health expectations play in individual health perceptions.

A typical, often cited example illustrating the comparability problems across groups is that of the Aboriginals in Australia. It is found that Aboriginals in Australia report better self-assessed health than the Australian population does. This is in sharp contrast with all other major indicators of morbidity and mortality showing that the Aboriginal population does much worse with respect to health than the general Australian population (e.g., Murray et al., 2002; Lindeboom & Van Doorslaer, 2004). This example therefore suggests that self-assessments on health of Aboriginals and the Australian population in general may not be comparable. Another example that is often quoted is that of the Kerala region in India. Although this region is found to have the highest life expectancy,

[3] Age may be an important factor influencing cut-points. Individuals may shift their expectations of health when they get older (Murray et al., 2000): ageing is generally associated with deteriorating health and this may be incorporated in the expectations of the elderly, such that some fixed true health level may be translated in a higher response category, i.e., better health, when they get older.

the lowest mortality rates among children and the highest rates of literacy of all regions in India, it also reports the highest level of morbidity of all regions (e.g., Murray et al., 2002; Lindeboom & Van Doorslaer, 2004; Iburg et al., 2002). Also empirically the severity of the problems with comparing self-reported health measures has been shown. Iburg et al. (2002) have investigated the existence of cut-point shift. They find significant differences in individual reporting thresholds as a function of gender, race and income. For instance, males are found to have lower cut-points than females such that the pattern of women reporting worse health that is frequently observed may reflect something else than true health differences between males and females (Iburg et al., 2002). This illustrates the seriousness of the comparability problems. Using uncorrected self-reported health measures may lead to conclusions that are far apart from reality and therefore solutions are required that correct for comparability issues and bring self-reports closer to the actual health status of individuals.

There are several ways to visualize the comparability problems with self-reported health measures. The gist in all these representations is the same: individuals may use diverse reporting thresholds in answering self-assessment questions and therefore individuals may translate the same level of the underlying unobserved outcome into different categorical responses. One way to represent this is shown in Figure 1. Such representations have been used by among others Murray et al. (2000), Tandon et al. (2002), Murray et al. (2002) and Salomon, Tandon & Murray (2004). The figures they show are for a specific domain of health, namely mobility. We have drawn a similar figure for the general health status.

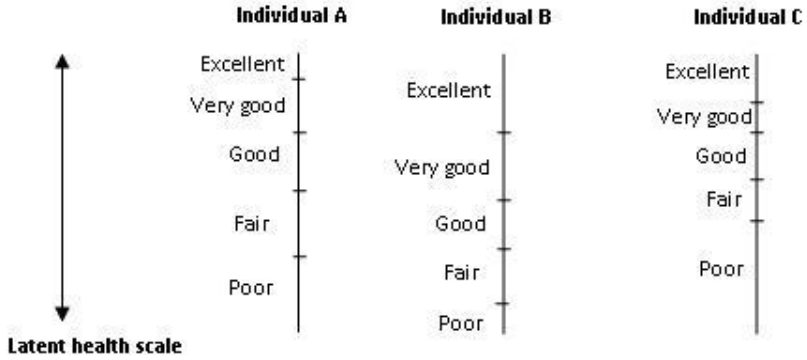

Figure 1: Mapping latent health status into individual categorical responses on health status.

Figure 1 shows the way in which three individuals map some unobserved latent general health level into a categorical response. For the same level of latent health, the three individuals respond very differently to the self-assessment question. Individual A must be in better true health in order to report excellent health than persons B and C. Individual B is much more optimistic on his health status than individuals A and C, since he is inclined to report better health for relatively low true latent health levels. Finally, individual C is quite extreme in his self-reports. For a relatively large range of true health levels he reports being in poor health or in excellent health, while he is much less likely to report categorical health levels in between these two extremes. Only for a rather small part of the latent health scale, he reports being in very good, good or fair health. The fact that the same latent health level leads to different categorical responses illustrates the existence of differential item functioning or response category cut-point shift. Stated differently, Figure 1 shows how individuals (or populations) vary in their mapping of unobserved latent health into a categorical self-reported health assessment. Therefore, self-reports are incomparable and cannot be used

directly to infer differences in the true health level across individuals. In order to be able to compare the categorical responses between the three individuals, the different response category cut-points must be made in line with each other. Such a translation can be done using anchoring vignettes. The gist of such an adjustment procedure is that the categorical response scale of one individual is used as the benchmark while the scales of the other individuals are adjusted to fit this benchmark scale. A detailed discussion of this method is postponed to section 4.

In the literature two types of reporting heterogeneity or response category cut-point shift are distinguished. Lindeboom & Van Doorslaer (2004) label these two types cut-point shift and index shift. They define cut-point shift as the situation in which the reporting thresholds (cut-points) are affected differently by the response behaviour, leading to a change of the relative positions of the reporting thresholds. On the other hand, in the case of index shift there is a parallel shift in the reporting thresholds resulting from reporting heterogeneity, such that the relative position of the thresholds remains unaltered. Some different labels have been assigned to these types of response category cut-point shift; for instance, systematic cut-point shift (e.g., Salomon, Tandon & Murray, 2004) or parallel cut-point shift is sometimes used instead of index shift.

The distinction between these two types of reporting heterogeneity is important in deciding which model to use. A standard model that is often used in case of ordinal data is the ordered probit model. In this model it is not possible to cope with cut-point shift, although some adaptations can be made in order to correct for index shift (Salomon, Tandon & Murray, 2004). In case of cut-point shift a different model is required. Usually the HOPIT model is employed in these instances. The modelling of self-reported health measures will be discussed extensively in section 5.

To conclude, the major problem associated with self-reported health measures is the fact that self-reports depend on both the objective situations as well as on the reporting behaviour of individuals and their unique subjective response thresholds (e.g., Voňková & Hullegie, 2010; Van Soest et al., 2007). Reporting heterogeneity between individuals may result in either cut-point shift or index shift, which may require the use of other types of models then the ones employed normally. These comparability problems can cause significant biases in estimation procedures that use self-reported health measures without correction either as independent or dependent variables. This will be the topic of section 3.2.

3.2 Resulting biases in estimation procedures

The aforementioned problems with self-reported health measures may result in significant biases in estimation procedures that use these measures as independent or dependent variable. The direction of these biases, i.e., underestimation or overestimation of coefficients, depends on the type of bias. Different types of biases and their direction will be discussed in the remainder of this section.

Firstly, the justification bias will have a significant effect on the estimated role of health in labour supply decisions. When health self-assessments contain justification bias, it will lead to biased health effects in the direction of poorer reported health driving retirement (Dwyer & Mitchell, 1999). Thus, the effect of health will be overestimated and the effect of economic variables (e.g., income and social security or disability insurance benefits) diminishes when self-reported health measures are used instead of objective health measures (Bound, 1991). This means that individuals who like to

work will report better health and therefore undervalue potential work-limiting disabilities. On the other hand, individuals with substantial preferences for leisure will exaggerate their work limitations and retire earlier (Dwyer & Mitchell, 1999). So endogeneity problems of self-reports will lead to overestimation of the effect of health on individual labour force participation (Bound et al., 1999).

Secondly, according to Bound (1991) there is no reason to expect that subjective judgements on health are comparable across individuals, i.e., reporting heterogeneity may exist. This lack of comparability represents measurement error which may lead to an underestimation of the effect of health on labour supply decisions (Bound, 1991; Bound et al., 1999). Underestimation of the health effect on labour market outcomes will also influence the estimated effects of economic variables on participation. The reason is that economic variables are correlated with health variables. Therefore a bias in the effect of economic variables on labour force participation will also result, independent of the accuracy of the estimated health effects (Bound, 1991)[4]. Related to this, Crossley & Kennedy (2002) find that there is substantial measurement error and uncertainty in the self-reported health measures leading to unreliable responses. The presence of measurement error in self-reported health may in turn lead to inconsistent estimates when these health measures are incorporated as independent variables.

Thirdly, as Crossley & Kennedy (2002) indicate, the answers of individuals to self-reported health question depend on the nature of the survey (i.e., written or verbal survey methods) and the sequence of prior questions. So self-reported health measures may also be incomparable across surveys. The extent to which a typical survey design is able to approximate true health may then affect the quality of self-reported health measures and the magnitude of the biases in the estimation procedures that adopt these health measures.

Finally, the difference between general health and work related health causes a downward bias in the effect of health on participation decisions (Lindeboom, 2005). Using an objective health measure like a physical exam does not necessarily capture the work limitation resulting from specific health problems. Although health limitations may be identified by these objective health measures, it does not necessarily take into account work limitations, such that these objective measures cannot explain labour force withdrawal or inactivity (Bound, 1991). So using these objective measures may cause general health instead of work related health being measured. This may in turn result in a downward bias in the estimated health effect.

In general one can conclude from the literature that accurate estimation of models including (self-reported) health as either the dependent variable or as an independent variable requires certain methods to correct for the several biases discussed in this section. Besides comparison of health across populations or groups within populations also requires correction methods that account for differences in response scales used by individuals. In the literature a couple of potential solutions to the problems with self-reported health measures have been developed and some of these, and their shortcomings, will be discussed in the next section.

3.3 Potential solutions to the problems with self-reported health measures

Several studies have used more objective health measures instead of the heterogeneous health self-reports in order to circumvent the problems with these self-reports. Objective health measures may

[4] For those interested a more thorough and technical discussion of the different biases resulting from the usage of subjective self-reported health measures in labour supply and retirement models can be found in Bound (1991).

be compared to self-assessed measures and from this comparison the existence of justification bias or state-dependent reporting may be confirmed or rejected (Lindeboom, 2005). Objective health measures can be for instance physician-assessed health measures or other medical reports. Nevertheless, these physician-assessed measures are sub-optimal, since also physicians are found to have some expectations that differ by age, gender and race (Murray et al., 2000). Besides, there is often no accurate objective measure available (Lindeboom, 2005; Bago d'Uva et al., 2008).

Other studies have used measured tests as objective measure for a particular health domain. For the cognition domain one of the measured tests used is immediate and delayed verbal memory. In this case individuals are shown a list of 10 words and are asked to immediately recall as many words as possible. This same question is also posed to them after a short delay (e.g., Bago d'Uva et al., 2009; Voňková & Hullegie, 2010). Another example of such a measured test is the walking speed test for the mobility domain. Individuals are then asked to walk a distance at their usual pace. The time it takes to do this is used as a measure of their mobility (Bago d'Uva et al., 2009). The use of measured test is probably not hindered that much by the influence of expectations as other objective and subjective health measures (Murray et al., 2000). However, this method is also not perfect, since only a part of the true health status in the domain under consideration is measured. The quality of measured tests as an approximation of true health may differ across tests.

As an alternative to just incorporating either self-reports or objective measures of health, one may also test for reporting heterogeneity by examining variation in reporting behaviour conditional on some objective health measure. Yet this method has also been linked to several problems, like the limited availability of accurate objective measures and the fact that when self-reports do contain some information on the true health status, conditional on objective measures, then this information will be lost (Bago d'Uva et al., 2008). Another approach is to instrument subjective self-reported health measures by objective health measures. It is however found that this method tends to underestimate the effect of economic and financial variables in labour supply decisions (Bound, 1991), although it estimates the health effect correctly (Kerkhofs & Lindeboom, 1995). The bias when using objective measures to instrument the subjective ones may be even worse than the bias when using self-reports alone (Bound, 1991).

One may be inclined to think that objective measures are closer to the true health status and that they therefore tend to alleviate the biases associated with the usage of self-reported health measures. In fact, the incorporation of objective health measures in estimation procedures critically rests on the assumption that objective health measures approximate true health better than self-reports do (Butler et al., 1987). However, numerous problems associated with the incorporation of objective measures instead of self-reports have been distinguished. Bound (1991), looking at applications in labour supply models, indicates that objective measures may measure something else then work capacity. When objective measures tend to gauge health instead of work capacity, they may not be appropriate to use in explaining labour supply decisions. More specifically, when objective health proxies measure health instead of capacity to work, errors-in-variables bias[5] may result (Kerkhofs & Lindeboom, 1995; Bound, 1991). The biases resulting from the usage of objective

[5] Errors-in-variables bias is a bias in the estimates of certain regression coefficients resulting from measurement error in the independent variables (Stock & Watson, 2007).

measures may be substantial and probably even worse than those resulting when using self-reports (Bound, 1991)[6].

So the several alternatives to using self-reported health measures without correction so as to alleviate or circumvent the incomparability issues associated with self-reported health measures are also found to suffer from serious limitations. Therefore still other methods are used to solve for these comparability issues. One important alternative method in this respect is that of the anchoring vignettes. We will turn to the discussion of the usage of anchoring vignettes as a correction tool in the next section.

[6] A more thorough discussion of the different biases and their implications for the statistical models can be found in Bound (1991).

4. The vignettes approach

As has been described in the previous section, there are a couple of problems involved with the usage of self-reported health measures. Especially the existence of reporting heterogeneity may seriously affect the trustworthiness of results obtained using these self-reported measures without applying an appropriate correction procedure. A method to correct self-reported (health) measures that has recently been introduced by King et al. (2004) is that of the anchoring vignettes. This method tries to make a distinction between differences in self-reported health resulting from differences in actual health and differences resulting from varying norms or expectations for health (Salomon, Tandon & Murray, 2004). The application of the vignettes method as a correction procedure is not bound to health economics alone. Anchoring vignettes are used in all kinds of fields of study, for example in studies on political efficacy (King et al., 2004), health (Bago d'Uva et al., 2008) and work disability (e.g., Kapteyn, Smith & Van Soest, 2009). For the present purpose we will only focus at the usage of anchoring vignettes for correcting self-reported health measures. In this section the vignettes method will be explained extensively. Both a description of the vignettes used in surveys as well as the way in which they allow one to correct for reporting heterogeneity will be discussed.

4.1 What are anchoring vignettes?

An anchoring vignette is a brief description of the health status of a hypothetical individual. A respondent is asked to evaluate a hypothetical person similar to the way in which he evaluates his own health situation. The respondent may be told to assume that the vignette persons have the same age and background as him. This is for instance the case in wave 2 of the SHARE project (www.share-project.org, 2010). The questions asking the respondent to rate the vignettes are almost identical to those asked for the self-assessment of health. Moreover, the same categorical response scale is used both for the self-assessment and for the vignette evaluation (e.g., Kapteyn, Smith & Van Soest, 2007, 2009; Voňková & Hullegie, 2010; King et al., 2004; Murray et al., 2002; Tandon et al., 2002). The vignette usually describes the ability level of the hypothetical individual on a given health domain (Murray et al., 2002). Some health domains for which vignettes are used are cognition (e.g., Voňková & Hullegie, 2010; Bago d'Uva et al., 2009), breathing (e.g., Voňková & Hullegie, 2010), mobility (e.g., Bago d'Uva et al., 2009; Voňková & Hullegie, 2010; Tandon et al., 2002; Murray et al., 2000), pain (e.g., Kapteyn, Smith & Van Soest, 2009; Bago d'Uva et al., 2008), affect (e.g., Bago d'Uva et al., 2008;), self-care (e.g., Murray et al., 2002; Bago d'Uva et al., 2008), cardio-vascular diseases (Kapteyn, Smith & Van Soest, 2009) and vision (e.g., Murray et al., 2000). An increasing number of household surveys, like the English Longitudinal Study of Ageing (ELSA), the Health and Retirement Study (HRS), the Survey of Health, Ageing and Retirement in Europe (SHARE) and the WHO's World Health Surveys (WHS), now contain a series of vignettes (Bago d'Uva et al., 2009).

A concrete example of a mobility vignette shown to individuals younger than 65 years old[7] in the wave 2 SHARE questionnaire is (www.share-project.org, 2010):

[7] For this particular vignette and self-assessment on mobility the questions are the same in the questionnaire for respondents aged 65 and older (www.share-project.org, 2010).

(Rob) is able to walk distances of up to 200 metres without any problems but feels tired after walking one kilometre or climbing more than one flight of stairs. He has no problems with day-to-day activities, such as carrying food from the market. In your opinion, how much of a problem does (Rob) have with moving around?

The corresponding self-assessment question for the mobility domain is formulated as:

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

The response categories from which a respondent can choose for both the self-assessments and the vignette evaluations are "none", "mild", "moderate", "severe" and "extreme". As one can see, the wording of the self-assessment and vignette questions is almost the same. Several vignettes representing varying levels of severity of health problems can be provided for one health domain. This is for instance the case in the SHARE dataset. The order in which vignettes are shown to the individuals may be randomized (King et al., 2004). Then individuals' evaluations will not reflect logical ordering considerations alone, thereby making response consistency more likely to hold (Salomon, Tandon & Murray, 2001).

The example above shows a vignette for a particular health domain. Theoretically it may also be possible to establish a vignette for the general health level of an individual, but it may be difficult to specify a complete description of a person's general health status while keeping it short at the same time. It is however important for vignettes to be short, since otherwise individuals will possibly not be able to rate the vignette appropriately. Individuals will lose their focus if vignettes are too long and this may deteriorate their understanding of the health status described. Also important to note is that the degree of reporting heterogeneity may differ across health domains. This may be the case since one domain of health may be more easy to grasp and more easy to understand than others. A very concrete health domain, like mobility, may be much easier for all individuals to evaluate than a more abstract, less tangible domain like pain or affect.

4.2 Correcting for DIF using the vignettes method

The anchoring vignettes discussed in the previous section can be used to correct for reporting heterogeneity. The general idea behind using anchoring vignettes to do the correction is that differences in the evaluation of a typical vignette person represent differences in reporting behaviour of individuals since the health status of the vignette person is fixed for every respondent (e.g., Bago d'Uva et al., 2008; Voňková & Hullegie, 2010; Kapteyn, Smith & Van Soest, 2007, 2009; Van Soest et al., 2007; King et al., 2004). Thus, the vignettes can be used to anchor reporting behaviour of an individual, and when this is known, the underlying true health differences can be discovered. Once one knows, from the outcomes of the vignette evaluations, which differences in reporting behaviour exist, one will be able to correct for this. Such a correction procedure can be done graphically or in a statistical, parametric, model. We will graphically explain the correction procedure in this section. The statistical models including vignettes will be discussed in section 5.

A graphical explanation of the correction procedure can be linked to Figure 1 in section 3.1; the basics of the graphical explanation and this figure are the same. Such a graphical explanation of the vignettes method and the way it can be used to correct for reporting heterogeneity has been shown by Van Soest et al. (2010)[8], by King et al. (2004) for political efficacy and by Kapteyn et al. (2010)[8] for work limitations. Our graphical representation shown in Figure 2 is very similar to that used by Van

[8] Presentations during the Netspar Mini Theme Conference on Anchoring Vignettes of April 28th 2010.

Soest et al. (2010), Kapteyn et al. (2010) and King et al. (2004) referred to earlier. It can be interpreted to represent vignettes and self-reports on mobility for instance.

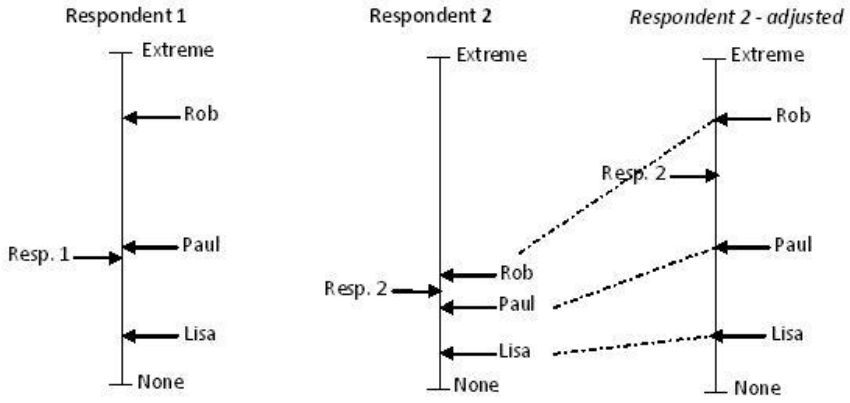

Figure 2: A graphical approach to correcting for reporting heterogeneity using vignettes.

Figure 2 consists of three parts. The leftmost part of the figure shows the responses of respondent 1 to the self-assessment question and three vignettes. The middle part shows the same for respondent 2. The vignette persons evaluated by respondents 1 and 2 are exactly the same. The rightmost part shows the responses of respondent 2 corrected for differential item functioning. Evaluating only the first two parts of the figure, one would be inclined to conclude that respondent 1 has more severe limitations than respondent 2. However, then reporting heterogeneity is not accounted for. But, reporting heterogeneity seems to be present, since the scale used by respondent 2 is much more compact than that of respondent 1. Making both scales comparable requires stretching the scale used by respondent 2 in such a way that both respondents evaluate the vignette persons in the same way, which is reasonable since the true health status of these vignette persons is the same. Such an adjustment of the scale of respondent 2 is shown in the rightmost part of the figure. The evaluation of the severity of the limitations of Rob, Lisa and Paul by respondent 2 is made in line with the evaluation of these persons by respondent 1. The self-assessment of respondent 2 must be adjusted accordingly: respondent 2 rates himself between Rob and Paul, but somewhat closer to Paul. This must still be the case in the adjusted scale. The space between Rob and Paul is now larger than in the middle panel and therefore the space between respondent 2 and Paul and respondent 2 and Rob must also become larger. The relative spaces between respondent 2 and Paul and respondent 2 and Rob must remain equal. After correcting, the self-reports on the severity of the limitations of respondents 1 and 2 are comparable. Comparing the leftmost and rightmost panels of Figure 2 leads to the conclusion that respondent 1 has less severe limitations than respondent 2, which is in sharp contradiction with the conclusion obtained earlier from comparing the self-reports before the correction procedure. Thus, adopting anchoring vignettes to correct for reporting heterogeneity alters the evaluation of the relative health states of both respondents in this case (King et al., 2004; Kapteyn et al., 2010; Van Soest et al., 2010).

The validity of the vignettes method rests on two crucial assumptions: *vignette equivalence* and *response consistency* (e.g., King et al., 2004; Kapteyn, Smith & Van Soest, 2007, 2010; Bago d'Uva et al., 2008, 2009; Voňková & Hullegie, 2010). *Vignette equivalence* means that all respondents are thought to perceive the health level described in the vignette in the same way (e.g., Bago d'Uva et

al., 2009; Salomon, Tandon & Murray, 2004). So it should not matter which characteristics the individuals have for the way in which they perceive the health level of the vignette person (Salomon, Tandon & Murray, 2001). Or more generally, as King et al. (2004) formulate it: "*vignette equivalence is the assumption that the level of the variable represented in any one vignette is perceived by all respondents in the same way and on the same unidimensional scale, apart from random measurement error.*" (p. 194). The second assumption of *response consistency* implies that individuals classify the vignettes in the same way as they rate their own health. So, in terms of the mapping of latent true health into a categorical response, this assumption means that the mapping in case of evaluating one's own health is identical to the mapping used in evaluating the health of a hypothetical person (e.g., Bago d'Uva et al., 2008, 2009; Van Soest et al., 2007; King et al., 2004).

In practice, these two assumptions may be violated, thereby reducing the quality of the correction for DIF using anchoring vignettes. Occurrences in which these assumptions can be violated will be discussed in section 6 in which the limitations of the vignettes method will be examined. First, we will turn to the modelling of self-reported health measures and incorporating the vignettes method therein in section 5.

5. Models for self-reported health measures

So far, we have discussed self-reported health measures, the potential problems with these measures and the vignettes method as a solution for these problems. We will now turn to the modelling of self-reported health measures. Modelling such discrete self-reports is often done "*by assuming that the observed categorical variable is a discrete representation of an underlying unobserved true health level, measured on a continuous scale*" (Bago d'Uva et al., 2008, p. 356). This is generally done using ordered response models, more specifically ordered probit models (Bago d'Uva et al., 2008). This is however not possible when reporting heterogeneity exists, since the model cannot cope with variation in response category cut-points (Salomon, Tandon & Murray, 2004). Ordered response models can however be extended to incorporate anchoring vignettes so as to correct for such reporting heterogeneity. First, section 5.1 describes the standard theory on ordered response models and the application of this to modelling self-reported health measures. The extended ordered probit model incorporating anchoring vignettes, the so-called HOPIT model, will be discussed in section 5.2.

5.1 Ordinal data, ordered response models and ordered probit models

Asking individuals to evaluate a particular outcome measure, like health, on a discrete categorical response scale results in ordinal data[9]. This means that the outcomes are ordered (for instance from very bad to excellent health), but the numerical values do not have an additional meaning. Consecutive integer values are assigned to the different categorical responses. The so-called ordered response model then links the observed self-reported outcomes to the index function. The index function, Equation 1, is a function relating the unobserved latent outcome variable (y_i^*) to a number of explanatory variables (x_i') and an error term (ε_i) (Heij et al., 2004).

$$y_i^* = x_i'\beta + \varepsilon_i, \;\; E(\varepsilon_i) = 0 \tag{1}$$

Subsequently, the observed outcome (y_i) is related to the unobserved latent outcome variable using a number of thresholds as in Equation 2 (Heij et al., 2004).

$$\begin{aligned} &y_i = 1 \; if \; -\infty < y_i^* \leq \tau_1, \\ &y_i = j \; if \; \tau_{j-1} < y_i^* \leq \tau_j, \\ &y_i = m \; if \; \tau_{m-1} < y_i^* \leq \infty. \end{aligned} \tag{2}$$

where m indicates the number of response categories, $j = 2, \dots, m-1$ and the τ's are the threshold values. There are $m-1$ thresholds needed when there are m categorical responses. The parameters that need to be estimated are the β's and the thresholds. These can be estimated using maximum likelihood estimation. Therefore we need to specify the log-likelihood function, for which the probabilities of reporting a particular categorical response are required. When $F(\cdot)$ is the cumulative distribution function of the error term, these probabilities can be specified as in Equation 3 (Heij et al., 2004).

$$\begin{aligned} p_{ij} = P[y_i = j\,] &= P\big[\tau_{j-1} < y_i^* \leq \tau_j\,\big] = P\big[y_i^* \leq \tau_j\,\big] - P\big[y_i^* \leq \tau_{j-1}\big] \\ &= F\big(\tau_j - x_i'\beta\big) - F\big(\tau_{j-1} - x_i'\beta\big), \qquad j = 1, \dots, m \end{aligned} \tag{3}$$

Then the log-likelihood is defined by Equation 4 (Heij et al., 2004).

$$\log\big(L(\beta, \tau_1, \dots, \tau_{m-1})\big) = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} \log\big(p_{ij}\big) \tag{4}$$

[9] The basic econometric concepts in the first part of this sub section that are relevant for the modelling of self-reported health measures are described based on parts of Chapters 6.1 and 6.2 of Heij et al. (2004).

where y_{ij} is an indicator function that equals 1 if $y_i = j$ and 0 if $y_i \neq j$. A cumulative distribution function of the error term ε_i must be specified. When a standard normal distribution function is assumed, the model is called an ordered probit model (Heij et al., 2004).

The theory on ordered response models is often applied to self-reported health measures. Bago d'Uva et al. (2008) describe a model for homogeneous reporting behaviour using the ordered probit model. The (observed) self-reported categorical health measure of individual i can be denoted by Y_{ri}[10]. This is the answer to a survey question assessed on a categorical response scale; the number of values that Y_{ri} can take depends on the number of categorical responses allowed for in the survey design. Bago d'Uva et al. (2008) then assume that this self-report is generated by an underlying latent (unobserved) true health level, denoted Y_{ri}^*. They specify the latent true health by the index function shown in Equation 5[11].

$$Y_{ri}^* = X_i'\beta + \varepsilon_i, \qquad \varepsilon_i | X_i \sim N(0,1) \tag{5}$$

In Equation 5 X_i is a vector of covariates and ε_i is the error term; it is assumed that conditional on the covariates the error term follows a standard normal distribution. Furthermore, they describe the relationship between unobserved latent true health and the categorical health measure as in Equation 6.

$$Y_{ri} = k \quad \Leftrightarrow \quad \tau^{k-1} \leq Y_{ri}^* < \tau^k \tag{6}$$

where $k = 1, \dots, K$, $\tau^0 < \tau^1 < \cdots < \tau^{K-1} < \tau^K$ and $\tau^0 = -\infty, \tau^K = \infty$[12] (Bago d'Uva et al., 2008). The τ^k's are the cut-points and k indicates a specific categorical response with K being the total number of response categories. For instance, when individuals are asked to rate their health on a five-point scale, K equals 5 and the cut-points are $\tau^0, \tau^1, \tau^2, \dots, \tau^5$. Homogeneity in reporting behaviour is reflected in the fact that the cut-points are constant (Bago d'Uva et al., 2008), i.e., do not vary across individuals. When reporting heterogeneity exists, biased estimates of the β's will result. The parameters τ^k, $k = 1, \dots K-1$ and the β's must be estimated (Bago d'Uva et al., 2008).

Similar health reporting models have been provided by among others Kerkhofs & Lindeboom (1995) and Lindeboom & Kerkhofs (2009). These models differ in notation and mainly focus on the relationship between health and work. The model of Kerkhofs & Lindeboom (1995) is somewhat more general since no specific functional form is imposed. Moreover an objective health measure is added to instrument the latent health variable.

5.2 Incorporating the vignettes method in statistical models

The standard ordered probit model for self-reported health measures as discussed in the previous section can be extended to include anchoring vignettes to correct for reporting heterogeneity. In this extended model the thresholds or cut-points are individual-specific, i.e., they are allowed to vary across individuals. Such an adjustment to the standard ordered probit model has been provided by King et al. (2004). They have developed a model that takes into account the vignette evaluations and

[10] The description of the model used here differs slightly in notation from the model as described in Bago d'Uva et al. (2008). This is done in order to get notations in the different models in this thesis in line with each other for clarity.

[11] Some other ways of modelling the true health status and self-reports are also used. For instance, Tandon et al. (2002) define the latent variable as a normally distributed variable with mean μ_i and variance 1, where μ_i is a function of certain individual characteristics like age, gender and education. Also King et al. (2004) use a slightly different model. However, the model presented in this section is the one most common in the literature on vignettes in health economics.

[12] A generalization of this model in which the cut-points are a function of the covariates is also possible (see e.g. Bago d'Uva et al. (2008) for an explanation of such a generalized model).

adjusts the self-reported health measures accordingly. This model has been labelled the hierarchical ordered probit (HOPIT) model (e.g., Bago d'Uva et al., 2008), or compound hierarchical ordered probit (CHOPIT) model (Van Soest et al., 2007; Voňková & Hullegie, 2010). These two models are very much alike however. I will refer to it as the HOPIT model in the remainder of this thesis.

The basics of the HOPIT model are similar to those of the ordered probit model discussed in section 5.1. That is, the self-report is a mapping from a true latent health level to a categorical response, the result of this mapping depending on the cut-points. The cut-points are now individual-specific and can be determined using the information from the vignette evaluations. These vignette evaluations are indicative for the way in which an individual uses the categorical responses, and therefore they can be used to identify and adjust differences in response category cut-points between individuals. The HOPIT model consists of two parts: the vignette component and the self-assessment component (e.g., King et al., 2004). The vignette component determines the cut-points for a particular individual, while the self-assessment component takes the values of these cut-points as given.

The self-assessment component consists of three equations. The index function is described in Equation 7 and is very similar to that in the ordered probit model (Equation 5). The relationship between self-reports and latent health is described in Equation 8, which is the same as Equation 6 except for the fact that cut-points are individual-specific as indicated by the subscript i.

$$Y_{ri}^{*} = X_i'\beta + \varepsilon_{ri}, \qquad \varepsilon_{ri} \sim N(0, \sigma_{\varepsilon}^2), \quad \varepsilon_{ri}\ independent\ of\ X_i \tag{7}$$

$$Y_{ri} = k \quad \Leftrightarrow \quad \tau_i^{k-1} \leq Y_{ri}^{*} < \tau_i^{k} \tag{8}$$

where i $(i = 1, \dots, n)$ indicates the respondent, $k = 1, \dots, K$ indicates the response categories, X_i is a vector of individual characteristics, Y_{ri}^{*} is the true health status of a respondent, Y_{ri} is the self-report, ε_{ri} an error term and $\tau_i^0 < \tau_i^1 < \cdots < \tau_i^{K-1} < \tau_i^K$ and $\tau_i^0 = -\infty, \tau_i^K = \infty$ (e.g., Kapteyn, Smith & Van Soest, 2007; Van Soest et al., 2007; Bago d'Uva et al., 2008; Voňková & Hullegie, 2010)[13]. To this cut-point equations should be added. Several specifications for the cut-points have been used in the literature. Two distinct versions are shown in Equation 9 (Bago d'Uva et al., 2008, 2009) and Equation 10 (King et al., 2004; Kapteyn, Smith & Van Soest, 2007, 2009; Van Soest et al., 2007).

$$\tau_i^k = \gamma^k X_i \tag{9}$$

$$\tau_i^1 = \gamma^1 X_i + u_i,\ u_i \sim N(0, \sigma_u^2)\ and\ \tau_i^k = \tau_i^{k-1} + exp\left(\gamma^k X_i\right), k = 2, \dots, K-1 \tag{10}$$

In both equations $\tau_i^0 = -\infty, \tau_i^K = \infty$ holds. The error term (u_i) in Equation 10 is assumed to be independent of X_i and the other error terms in the model. In the specification of Equation 9, no error term has been included, while this is done in Equation 10. Including the error term in the specification means that an unobserved individual effect is incorporated (Kapteyn, Smith & Van Soest, 2007) (i.e., individual heterogeneity is allowed for in the cut-points (Kapteyn, Smith & Van Soest, 2009)). Besides, in Equation 9 the cut-points are a linear function of the characteristics, while Equation 10 defines the first threshold as a linear function, but the function for the other cut-points also contains an exponential part.

When only the self-assessment part of the model is used, there is an identification problem (e.g., Bago d'Uva et al., 2009; Kapteyn, Smith & Van Soest, 2007, 2009). The parameters β and γ^1 are not identified separately in this model (Kapteyn, Smith & Van Soest, 2007, 2009). For identification of the model additional information is needed (Bago d'Uva et al., 2009). This information can be provided by anchoring vignettes. Therefore the vignette component of the model should be added.

[13] The notation used here differs from that adopted in the papers referred to. This is done to keep it comparable to the ordered probit model described in section 5.1.

The vignette component can be specified in the following way. Let j denote a particular vignette where $j = 1, \dots, J$. The perceived health status of the vignette person j is denoted Y_{ji}^{v*} and the categorical response of respondent i[14] to vignette j is denoted by Y_{ji}^{v}[15]. Again, several ways exist for modelling the vignette component. King et al. (2004) visualize the assumption of *vignette equivalence*, as has been describe in section 4.2, in their model in the fact that the true health level of the vignette person does not depend on the respondent i evaluating the vignette (i.e., Y_j^{v*} does not have a subscript i) (King et al., 2004). On the other hand, Bago d'Uva et al. (2008, 2009) and Voňková & Hullegie (2010) take up the assumption of *vignette equivalence* by modelling the true health level of the vignette person as in Equation 11[16].

$$Y_{ji}^{v*} = \alpha_j + \varepsilon_{ij}^{v}, \quad \varepsilon_{ij}^{v} \sim N(0,1) \tag{11}$$

So, in this case *vignette equivalence* means that all individuals have the same perception on the latent health level of the vignette person, except for some random measurement error (Bago d'Uva et al., 2008). The observed categorical responses to the vignettes can be related to the true latent perceived health status of the vignette person with Equation 12.

$$Y_{ji}^{v} = k \quad \Leftrightarrow \quad \tau_i^{k-1} \leq Y_{ji}^{v*} < \tau_i^{k} \tag{12}$$

where again $k = 1, \dots, K$, $\tau_i^0 < \tau_i^1 < \cdots < \tau_i^{K-1} < \tau_i^K$ and $\tau_i^0 = -\infty, \tau_i^K = \infty$ (Bago d'Uva et al., 2008, 2009; Voňková & Hullegie, 2010). Besides, the cut-points are defined by Equation 13. They are again, as in the self-assessment component, functions of the covariates contained in the vector X_i.

$$\tau_i^k = \gamma_v^k X_i \tag{13}$$

The assumption of *response consistency* discussed in section 4.2 is formalized in the specification of the cut-points. These cut-points are not allowed to vary across the different vignettes j (Bago d'Uva et al., 2008), therefore the cut-point τ_i^k has no v in its superscript. Or stated differently, one has to impose that $\gamma_v^k = \gamma^k$ for $k = 1, \dots, K$, where the γ^k parameters are the same as those in Equation 9 (e.g., Bago d'Uva et al.,2010; Van Soest et al., 2007; King et al., 2004). Thus, the cut-points for evaluating one's own health are assumed to be the same as the cut-points used in evaluating the health status of a hypothetical vignette person.

The incorporation of anchoring vignettes in the model allows one to separately identify the β and γ coefficients. The vignettes are now used to identify the γ parameters in the functions of the cut-points. Besides, the vignettes can be used to identify the parameters α_j for $j = 1, \dots, J$. When these cut-point values have been fixed using the anchoring vignettes, they can be used in the self-assessment part, i.e., using the self-reports one can identify the β parameters (e.g., Kapteyn, Smith & Van Soest, 2007, 2009; Van Soest et al., 2007). The parameters can now be identified "*up to the usual normalization of scale and location*" (Kapteyn, Smith & Van Soest, 2007, p. 464). As with standard ordered response models this model can be estimated using maximum likelihood estimation. Theoretically, the usage of one vignette is enough to make the model identified. However, it may be better to include a larger number of vignettes in order to get a better correction for response category cut-point shift (King et al., 2004).

The basic HOPIT model can be extended further by including objective measures into the model. These objective measures can be used to determine the quality of the vignettes method as a

[14] It is also possible to let a subset of the respondents evaluate the vignettes. This is allowed for in the model described by King et al. (2004).
[15] The superscript v used here refers to the vignette component of the model.
[16] Notation as in Bago d'Uva et al. (2008) is used here.

correction tool since it is assumed that the correction must bring the self-assessments closer to the objective situation. Such an extension of the model is for instance provided by Van Soest et al. (2007). Objective measures have also been included by Voňková & Hullegie (2010). Besides, Bago d'Uva et al. (2009) present the incorporation of objective measures as an alternative to the vignettes method. The latter has also been done by Lindeboom & Kerkhofs (2009). They allow for a nonlinear relationship between latent true health and the objective measures.

The objective health status can be denoted by Y_{oi}^* , again X_i denotes a vector of covariates and ε_{oi} is an error term. The assumption made by Van Soest et al. (2007) and Voňková & Hullegie (2010) regarding this error term is that it is independent of the covariates, independent of the error term in the vignette part of the model and the unobserved heterogeneity term (u_i) in Equation 10. Finally, the error term ε_{oi} is allowed to be correlated with the error term from the self-assessment component of the model (ε_{ri}). The specification of the objective part of the model is now described by Equations 14 and 15 (Voňková & Hullegie, 2010).

$$Y_{oi}^* = X_i'\beta_o + \varepsilon_{oi} \tag{14}$$

$$Y_{oi} = k \quad \Leftrightarrow \quad \tau_o^{k-1} \leq Y_{oi}^* < \tau_o^k \tag{15}$$

where $k = 1, \dots, K$, $\tau_o^0 < \tau_o^1 < \cdots < \tau_o^{K-1} < \tau_o^K$ and $\tau_o^0 = -\infty, \tau_o^K = \infty$. The cut-points are now not allowed to vary across individuals (Van Soest et al., 2007; Voňková & Hullegie, 2010).

6. Potential problems with the usage of vignettes

Whether the vignettes method does a good job in correcting self-reported health measures still is an important question that researchers try to answer in several different ways. To a large extent the problems with the usage of anchoring vignettes as a correction tool are grounded in the assumptions needed in the models incorporating them. These assumptions of *vignette equivalence* and *response consistency* may fail in several instances and in that case the vignettes method may be inappropriate as a correction tool. Despite of the increasing number of studies evaluating the validity of the vignettes method, no consensus on its quality has been reached yet. Several problems may deteriorate the quality of the vignettes method. This will be the topic of this section.

The assumption of *vignette equivalence*, meaning that individuals all understand the underlying health level described by a vignette in the same way, may be violated when the vignette descriptions are incomplete or unclear leading to errors in the interpretation of the story told by the vignette. It may cause difficulties in understanding the health status described by it or individuals using their own imagination to complete or clarify the vignette descriptions. In that case one individual may perceive the described health status in a different way then another individual, such that *vignette equivalence* is not fulfilled. Variation in ratings of a particular vignette then cannot be attributed to response category cut-points shift (e.g., Bago d'Uva et al., 2008, 2009; King et al., 2004). In addition, individuals may have difficulties in understanding the underlying true health level of a vignette person in a good manner when their own health status is very incomparable to the health status of the vignette person. For example, when the vignette person is in very good health, he may not be able to understand what it is like to be in very bad health, such that he will have difficulties in evaluating the health status of someone with extreme difficulties in a consistent manner.

Response consistency means that individuals use the response scales in the same way independent of whose health status they are evaluating. This assumption may be violated when strategic considerations[17] play a role in evaluating one's own health status while these are absent when evaluating a hypothetical person's health (Bago d'Uva et al., 2008, 2009). In that case individuals use the response scales in a different way depending on which person's health they are evaluating. Since vignettes have been developed and incorporated so as to correct for systematic differences in reporting behaviour violation of this assumption will result in serious problems. If vignettes are evaluated systematically different compared to the own health status, they are not indicative for how individuals evaluate their own health and thus cannot be used to correct for reporting heterogeneity. Thus, if this assumption is not fulfilled the power of anchoring vignettes as a correction tool will be reduced significantly and probably the correction method would even become powerless (Bago d'Uva et al., 2009).

Several attempts to test the appropriateness of the vignettes method as a correction tool have yet been undertaken. Some of these studies specifically test if the two assumptions have been fulfilled. Others look at more objective measures and compare the corrected self-report to such an objective measure, assuming that successful correction would imply that the corrected self-reports are closer to the objective measures than the uncorrected self-reported health measures. These studies provide mixed results on the quality of the vignettes approach. Some find positive results, others find

[17] An example of such strategic considerations is the justification bias discussed in section 3.1.

positive results only in a couple of health domains and still others indicate negative results for the appropriateness of using anchoring vignettes as a correction method. We will discuss some of these studies in a bit more detail here.

The assumption of *vignette equivalence* can be evaluated informally by looking at the ranking of evaluated vignettes across individuals. Some variation in rankings can be expected because of random measurement error. However evaluating the consistency of individual rank orderings compared to the overall average rank ordering for all individuals gives some indication of the validity of the *vignette equivalence* assumption. Results using such an approach are overall positive on the validity of the *vignette equivalence* assumption (Salomon, Tandon & Murray, 2001, 2004). A formal test for *vignette equivalence* has been undertaken by Bago d'Uva et al. (2009). They state that a necessary condition for this assumption to hold is that there is "*no systematic variation with observed individual characteristics in the perceived difference in states corresponding to any two vignettes*" (Bago d'Uva et al., 2009, p. 3). This can be tested with the null hypothesis that there is no interaction between vignette dummies and the covariates relevant for the vignette. In testing this, one has to assume that individuals use the same response thresholds for the evaluation of all vignettes. The results obtained by Bago d'Uva et al. (2009) show that violation of the *vignette equivalence* assumption indeed seems to be the case for both domains of health investigated.

Testing the assumption of *response consistency* requires a different approach. Several studies incorporate objective measures in order to be able to test for this assumption (e.g., Bago d'Uva et al., 2009; Van Soest et al., 2007; Gupta et al., 2010). When such objective measures are available one can compare the cut-points resulting from the vignettes and one (or several) objective measure(s) (Bago d'Uva et al., 2009). Bago d'Uva et al. (2009) test for response consistency using the null hypothesis of equal cut-points in the vignette component and the self-assessment component of the HOPIT model. Besides, they test for equal distances between the cut-points. The latter is a weaker test for response consistency, but this one is appropriate when the assumptions necessary for the former test do not hold. It is found that the strict test of *response consistency* is rejected for all domains of health considered. The weaker test rejects the null hypothesis of *response consistency* for one domain but not for the other. Bago d'Uva et al. (2009) eventually conclude that they have found evidence against the validity of the vignettes method for both health domains considered. However, they also indicate that the adopted tests are very demanding. Despite of the violation of both assumptions, the vignettes method may be appropriate to determine the direction of the bias caused by reporting heterogeneity (Bago d'Uva et al., 2009).

Gupta et al. (2010) test for *response consistency* by relaxing this assumption in the standard CHOPIT model. They estimate a model without this assumption imposed, adding objective measures to make the model identified. They label this model the OChopit model, i.e., the objective-extended CHOPIT model. Besides, they estimate the standard CHOPIT model with vignettes and subjective self-assessments included. These models are then compared based on the likelihood values and a likelihood-ratio (LR) test is applied. The results show a better fit of the model without the assumption of *response consistency* imposed, so this assumption may be restrictive and not completely valid.

King et al. (2004) provide some supportive evidence for the validity of the vignettes method. They compare corrected and uncorrected self-assessments to an objective measure using self-reports and vignette evaluations on the health domain of visual acuity. They find that their approach seems to correct in an appropriate way since the corrected self-reported measure of visual acuity is closer to the objective measure than the uncorrected self-assessment.

Voňková & Hullegie (2010) look at whether the vignettes method is sensitive to the health domain that the vignette is about and the choice of a particular vignette. They find that the vignettes method is in fact sensitive to the domain and choice of the vignette. However, they only look at correlations in their study while it would also be a good idea to look at whole distributions and the extent to which they are comparable using different vignettes and compared to the objective measures. Another problem may be that the objective measures used are of questionable quality. For example for mobility the objective measure used is the time it takes an individual to stand up five times keeping his arms across his chest. But this may probably not be very indicative for the true mobility level of an individual.

Van Soest et al. (2007) also test for response consistency using objective measures, vignettes and self-assessments on drinking behaviour. They compare whether the distribution of corrected self-reports of drinking problems is closer to the objective one than the uncorrected self-reports. Van Soest et al. (2007) conclude that the vignettes method is able to correct the self-reported drinking problems and bring them closer to the adopted objective measures of drinking problems. They also indicate that this positive conclusion may be related to the availability of a clear objective measure closely related to the self-assessment question. However, when the objective situation is less clear, the vignettes method may be less successful in correcting for DIF (Van Soest et al., 2007).

Finally, Van Soest et al. (2010) conducted a survey in which individuals are first asked to evaluate their own health on a categorical scale. In addition, they should rate a couple of vignettes and answer a number of detailed, objective, questions on their own health. In the second round of the survey, individuals are shown a vignette that has been constructed based on the evaluation of their own health resulting from the objective part of the questionnaire of the first wave. If response consistency holds the evaluation of the vignette in the second wave should not be systematically different from the evaluation of the self-assessment part in the first wave. Van Soest et al. (2010) mention two assumptions underlying this test. First, the response scales should be the same in wave 1 and wave 2. Secondly, the objective health questions from wave 1 should be a complete description of the true health status of the respondent. The objective questions cover a number of topics for a specific health domain. This can lead to quite long vignette descriptions in the second wave, which may have a negative impact on the ability of respondents to understand the vignette and to evaluate it consistently. Important to note is that a negative result does not prove that response consistency fails; a couple of other explanations for negative results have been provided by Van Soest et al. (2010), for example the existence of order effects in vignette evaluations meaning that the evaluation of a particular vignette may be affected by previously evaluated vignette descriptions. Overall, Van Soest et al. (2010) conclude that the joint hypothesis of all assumptions holding at the same time is rejected most of the time, but nevertheless vignettes seem to do a good job in correcting for reporting heterogeneity.

In reaction to the potential problems with the vignettes method some improvements have been proposed recently. For instance, Hopkins & King (2010) suggest that self-reports follow the vignette evaluations in a survey. This may improve the quality of the method since it is found that the order of survey questions matters and that the meaning of self-reports may be clearer after evaluating vignettes such that a better self-report can be provided. When vignettes are provided first, they may lay the foundations for the frame in which the self-reports should be placed. It is found that asking the vignettes first and the self-assessment question thereafter would improve the quality of the correction method significantly (Hopkins & King, 2010).

7. Data description

Now that the need of the vignettes method, the method itself and its potential problems have been discussed, we can continue to develop a test of the appropriateness of the vignettes method as a tool for correcting self-reported health measures. The testing procedure developed and applied in this thesis will make use of the SHARE data. This section first describes the dataset in detail.

7.1 The SHARE data set

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a survey consisting of individuals aged 50 and over in several countries in Europe. It is a "*unique and innovative multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks of more than 45,000 individuals aged 50 or over*" (SHARE Tackling the Demographic Challenge, 2010). The individuals participating in the survey are representative for the non-institutionalized population aged 50 and older. Besides spouses of these individuals are interviewed if they are younger than 50 years old (www.share-project.org, 2010). The SHARE project has been harmonized with the Health and Retirement Study (HRS) conducted in the United States and the English Longitudinal Survey of Ageing (ELSA) in the United Kingdom (SHARE Tackling the Demographic Challenge, 2010).

Up to now, data for two waves of the survey are available: wave 1 in 2004 and wave 2 in 2006. The panel structure of the data allows one to look at the dynamic character of the ageing process (SHARE Tackling the Demographic Challenge, 2010). In the first wave twelve countries participated in the survey: Austria, Belgium, Denmark, France, Germany, Greece, Israel, Italy, the Netherlands, Spain, Sweden and Switzerland. The second wave covers a broader range of countries, also including the Czech Republic, Poland and Ireland. The total number of respondents in wave 1 is 31,115. In wave 2 not only individuals from the first wave have been re-contacted; also a "refresher" sample has been drawn in most countries (www.share-project.org, 2010). This has extended the sample size to 34,415 individuals. The sample sizes are largest in Belgium, France and Sweden. The least individuals have been interviewed in Switzerland and Denmark. The total number of women in the sample is considerably larger than the total number of men: 17,304 versus 13,811. In the first wave, the fraction of women is larger than that of men in all countries.

The data collected in the SHARE project include a broad range of variables; for instance health measures (e.g., self-reported health measures, health conditions, use of health care facilities and physical and cognitive functioning), psychological variables (e.g., well-being and life satisfaction) and economic variables (e.g., current work activity, job characteristics, sources and composition of current income, housing, education, wealth and consumption) (www.share-project.org, 2010). Besides, for the present purpose the anchoring vignettes are a very important part of the dataset.

The SHARE dataset consists of several modules: a computer-assisted personal interview (CAPI), a paper based questionnaire for the drop-off and vignettes questionnaires and generated variables modules (SHARE Guide to Release 2.3.0 Waves 1 & 2, 2009). The questionnaires are translated for each country, with some minor deviations. There are also deviations in the questionnaires across waves, details on that can be found at the SHARE website (www.share-project.org, 2010). Not each question of the CAPI interview needs to be answered by each

individual[18]. The main sample finishes the interview with the drop-off questionnaire; the vignette sample on the other hand ends with a special questionnaire containing anchoring vignettes (SHARE Guide to Release 2.3.0 Waves 1 & 2, 2009).

7.1.1 The individual CAPI modules

The main questionnaire consists of several computer-assisted personal interview (CAPI) modules. There are 20 modules in the first wave and three more in the second wave. In the remainder of this section the contents of the various CAPI modules will be discussed briefly[19], based on the various questionnaires[20] (www.share-project.org, 2010).

Each interview starts with a so-called coverscreen module. Data from this module are available both at the household as well as at the individual level. The questions in the coverscreen module are basic demographic questions. For instance name, gender, date of birth, composition of the household, age and whether someone is single or living with a spouse or partner are assessed in this questionnaire.

The demographic (DN) module asks for the month and year of birth of the respondent, country of birth, citizenship, highest educational degree obtained and further education and marital status. Next to that, some questions are asked about the respondent's natural parents; among others age, last job, the amount of personal contact with the parents and the health of the parents are considered. Also the family composition of the respondent is considered. More details on this can be found in the children (CH) module. This latter module provides detailed insight into the household composition, for instance the number of children, where they live, their age, education and occupation.

A couple of modules evaluate the health status of a respondent in detail. First of all, the physical health (PH) module asks the individual for a self-report on his health status on a scale ranging from "excellent" to "poor". In addition long-term health problems, work limiting conditions, limitations in activities due to health problems and doctor examinations of existing conditions are considered. Moreover, the respondent is asked to indicate from which symptoms he suffered for the past six months and the drugs he currently uses at least once a week. Finally, difficulties with a large number of daily activities and the help received for those activities are examined. Furthermore, emotional health is assessed in the mental health (MH) module. Here the respondent is asked whether he has felt sad or depressed in the last six months, whether he feels guilty or blames himself for something and whether he has problems with sleeping. Other aspects of the mental health state that are considered are irritability, concentration, tiredness and the history of the respondent with depression. Related to the health status of the individual, the behavioural risks (BR) module looks at current and past smoking behaviour and drinking behaviour in the last six months and the amount of physical activity the respondent engages in.

[18] A detailed description of who answers what part of the CAPI interview can be found in the SHARE Guide to Release 2.3.0 Waves 1 & 2 (2009).

[19] We will discuss only those modules that are most relevant for our purpose. Modules that have not been included in the discussion here are: social support (SP), activities (AC), expectations (EX), interviewer observations (IV) and the end-of-life interview (XT). A complete discussion of the various modules can be found in the SHARE documentation (www.share-project.org, 2010).

[20] Complete versions of the generic English questionnaires as well as the country-specific translated versions can be found at the SHARE website (www.share-project.org, 2010).

While the aforementioned health modules largely contain health self-assessments or information received from a doctor, several other modules look at more objective health measures. In these modules several measured tests are conducted that should give an objective assessment of some aspect of the health status. First, there is a grip strength (GS) module that includes a test in which the respondent is asked to squeeze a handle as hard as possible for a few seconds. Measurements are collected two times with a dynamometer for both the left and the right hand. Another objective health measure is the walking speed test conducted in the walking speed (WS) module. The respondent is asked to walk a short distance and the time needed for this is recorded. The test is conducted twice. The cognitive function (CF) module contains several tests to assess the respondent's cognitive ability. In the first test the interviewer reads ten words aloud and the respondent is asked to immediately recall as many words as possible. After that, his verbal fluency is tested by asking him to name as many animals as possible in one minute. This part of the module is followed up by the numeracy part in which the use of numbers by the respondent is assessed. This part of the questionnaire covers some basic mathematics. After these numeracy questions have been completed, the delayed recall ability of an individual is tested by asking the respondent to recall as many of the ten words mentioned earlier.

The chair stand (CS) module and the peak flow (PF) module have only been conducted in wave 2. The former is a test to measure the strength and endurance in the legs of the respondent. In this test the respondent is asked to fold his arms around his chest and then stand up from a chair five times while keeping his arms around his chest. Some studies have used this test as an objective assessment of the mobility of the respondent (e.g., Voňková & Hullegie, 2010). The peak flow test measures how fast a respondent can expel air from his lungs, while blowing as hard as he can; it is an assessment of the breathing ability of a respondent.

The health care (HC) module evaluates several forms of health care usage by the respondent. For instance, the usage of a medical doctor, i.e., a general practitioner or a specialist, in the last 12 months is assessed. Besides, dentist visits, hospital visits, stays at a nursing home, the receipt of home care and various types of surgeries are examined in detail. Costs and availability of care are considered, just as out of-pocket care expenses and health insurance coverage. Also more detailed questions about the health insurance of a respondent are posed.

The financial and economic situation is accounted for in a couple of modules. For instance, labour market status, housing ownership and financial transfers are considered in these modules. The module on employment and pensions (EP) looks at the current labour market status of the respondent, i.e., (self-)employed, unemployed, retired, sick or disabled or homemaker. If the respondent is currently employed, detailed questions about the current job are asked. Furthermore, job satisfaction and whether the job is physically demanding are considered. Moreover, the income of the respondent is evaluated in detail; the amount and frequency of payment from several income sources is taken into account. If the respondent is retired at the moment, he is asked for the reason of his retirement and how he feels about being retired. Similarly, an unemployed individual is asked how he became unemployed and the disabled are asked if work has been the cause of becoming disabled. Finally, when the individual indicates that he is a homemaker, the reason for quitting work is evaluated.

The household income (HH) module also takes into account income of other household members so as to determine the level of total household income. Various sources of income are considered, for instance labour income, child benefits and poverty relief. Spending patterns are

evaluated in the consumption (CO) module. The financial transfers (FT) module evaluates financial or material gifts or support to others (e.g., parents, children, friends and neighbours) and received gifts. The housing (HO) module looks for example at housing ownership or renting behaviour of the respondent. Related to that the respondent is asked what mortgages or loans he has and what the associated regular repayment obligations or his required rent payments are. Possessions of the respondent are considered in the assets (AS) module. In this respect, one could think of bank or saving accounts, debts, bonds, life insurance and tangible assets like cars and companies.

7.1.2 The drop-off questionnaire

The drop-off questionnaire contains additional questions on issues like mental and physical health, health care and social networks. This questionnaire is answered only once by a specific individual, so it is not longitudinal. It contains questions about life satisfaction and quality of life in early old age based on the satisfaction of various needs (SHARE Guide to Release 2.3.0 Waves 1 & 2, 2009). Besides, it encloses a life orientation test that determines if the respondent has a pessimistic or optimistic attitude. In addition, the questionnaire includes a question on the feelings of the respondent, for instance whether he feels depressed or sad.

7.1.3 The vignette questionnaire

The vignette sample answers vignette questions instead of the drop-off questionnaire. Individuals are asked to rate the hypothetical vignette persons in both waves. A vignette sample is only available in a selection of all countries participating in the SHARE project. In wave 1 there is a vignette sample in Belgium, France, Germany, Greece, Italy, the Netherlands, Spain and Sweden. In wave 2 also some respondents from Denmark, Poland and the Czech Republic fill in a vignettes questionnaire. Each wave has two versions of the vignettes questionnaire: type A and B in wave 1 and type B and C in wave 2[21]. These versions differ in the order in which the vignettes are shown and the gender of the vignette persons. In wave 1 these different versions were assigned randomly to the individuals in the vignette sample. For wave 2 however, type B was assigned to respondents aged 64 or below and type C was assigned to respondents aged 65 or older.

To illustrate typical self-report and vignette questions, an example for the memory domain is provided here (www.share-project.org, 2010; vignette questionnaire wave 1, type A[22]).

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

Vignette 1: 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.

Vignette 2: 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.

[21] Vignettes of type A in wave 1 correspond to vignettes of type B in wave 2 and type B in wave 1 corresponds to type C in wave 2 (SHARE Guide to Release 2.3.0 Waves 1 & 2, 2009).

[22] The type B vignette questionnaire contains the same vignette descriptions although the order of the vignette evaluations and the name of the vignette persons differ compared to the type A questionnaire.

Vignette 3: 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 (Lisa/Sue/Eve) had with concentrating or remembering things?

For both the self-report and the vignette evaluations the possible response categories are "none", "mild", "moderate", "severe" or "extreme". The vignettes are ordered according to the severity of the problems in the domain they describe: vignette 1 represents the least problems, while vignette 3 shows the most severe problems with concentrating and remembering things.

In wave 1 the vignette sample consists of 4,544 individuals. Those individuals are asked 34 evaluation questions (self-reports and vignette evaluations). The first of these questions are self-reports on six health domains: pain, sleep, mobility, memory, breath and depress and on work disability. After that, for each of these domains vignettes are shown to the respondents. For each health domain three vignettes are provided; nine work disability vignettes are shown. Vignettes for the various domains are mixed to some extent in the wave 1 questionnaires: pain and sleep vignettes are interchanged, the same holds for memory and mobility and for breath and depress.

In wave 2 7,731 individuals are part of the vignette sample. However, nine of them have not answered any vignette evaluation in the second wave, so in fact vignette evaluations are available from 7,722 individuals. The vignette questionnaire used in wave 2 differs somewhat from the one used in wave 1. Still, vignette evaluations are included for the six health domains pain, sleep, mobility, memory, breath and depress. However, only one vignette evaluation on each domain is available. Besides, three vignette descriptions on work disability are included (www.share-project.org, 2010; vignette questionnaire wave 2, type B[23]). In wave 2, not only health vignettes are used; also, vignette descriptions about satisfaction with social contacts, income, daily activities, job and live in general and the influence on decisions at the municipality level are evaluated. These vignette evaluations will not be used in this thesis so we will not explain them in more detail here.

7.1.4 The generated variables modules

Finally, the SHARE dataset contains a number of generated variables. Only the most relevant ones will be discussed here. First, there is the ISCO-88 module (International Standard Classification of Occupations) that organizes the occupations of the respondents into international comparable groups. Furthermore, there is a module with generated education variables. Country-specific education levels, as obtained from the various CAPI modules, are translated into international comparable measures of education using the ISCED-97 coding (International Standard Classification of Education) (SHARE Guide to Release 2.3.0 Waves 1 & 2, 2009). This classification consists of 9 categories: ISCED code 1 to ISCED code 6, no education (coded "0"), "still in school" (coded "95") and "other" (coded "97").

[23] In contrast to the type B vignette questionnaire, the type C questionnaire does not contain questions on self-reported work limitations or vignette descriptions about work limitations (since it asked to individuals at or above the normal retirement age). Besides, the type C questionnaire contains more vignette evaluations regarding health care responsiveness (since the elderly are expected to make more use of health care facilities). The other questions and vignette descriptions are exactly the same as in the type B questionnaire, including the order in which they are asked and the names used for the vignette persons (www.share-project.org, 2010; vignette questionnaires wave 2, type B & type C).

In addition, a generated variables module on health has been included. This module contains generated variables related to cognitive function, mental health, physical health, behavioural risks, the grip strength test and the walking speed test. These generated variables are abstracted from the data gathered in the various CAPI modules. Examples of generated health variables are the extent of depression, the number of chronic diseases and the number of symptoms that a respondent has and the number of limitations with mobility, arm function and fine motor function. In addition, the number of limitations with activities of daily living (ADL) is addressed; this includes problems with dressing, walking across a room, bathing or showering, eating, getting in and out of bed and using the toilet. Related to that there is also a generated variable for the number of limitations with instrumental activities of daily living (IADL), including using a map, preparing a hot meal, shopping for groceries, making telephone calls, taking medications, doing work around the house or garden and managing money. For some of the measured tests generated variables are obtained; for instance for the walking speed test the average of the two measurements is taken (SHARE Guide to Release 2.3.0 Waves 1 & 2, 2009).

Finally, there is a module with imputed values for a number of variables. Imputed values have been obtained using five different methods. The variables for which imputed values have been obtained are a large number of income variables, like the annual gross income from employment last year, the annual public old age pension, the annual public disability insurance and the annual early retirement pension. Also, variables on the financial situation of the household like the annual gross income of other household members are included. Furthermore, out-of-pocket expenditures on several types of health care usage are part of the imputations module. Finally, exchange rates are included. These are necessary since all financial amounts are reported in the local currency. To work with these amounts and compare them across countries, one thus has to use the exchange rate to convert them into the same currency, i.e., Euros (SHARE Guide to Release 2.3.0 Waves 1 & 2, 2009).

7.2 Constructing a master dataset

The SHARE dataset is a very rich dataset containing a large amount of variables and respondents for a considerable number of countries. However, for the purpose of this study we only need a selection of all modules, variables and respondents. Section 7.2.1 discusses the construction of the sample that we will use in our testing procedure. The construction of variables and the selection of modules used will be discussed in section 7.2.2.

7.2.1 The sample

The overall SHARE sample consists of 31,115 individuals in wave 1 and 34,415 respondents in wave 2. Since we want to evaluate the quality of the vignettes method, we need data on the evaluation of vignettes, such that we can only use those respondents in the vignette sample. This restricts our sample to 4,544 respondents in wave 1 and 7,731 individuals in wave 2. In addition, we only consider those individuals aged between 50 and 85 years old. Although the age of respondents varies from 24 to 102 years old, we consider this selection of respondents since there are only very few observations on individuals younger than 50 years old and individuals older than 85 years old. Excluding individuals younger than 50 years old and individuals older than 85 years old reduces the wave 1 sample from 4,544 respondents to 4,308 individuals. The wave 2 vignette sample declines in size from 7,731 to 7,375 respondents. So, removing all observations on individuals younger than 50 years old and individuals older than 85 years old results in a total loss of 592 observations. Moreover,

we can only make use of those respondents for which information is available in both waves, i.e., the longitudinal part of the sample. More specifically, in our sample we select those individuals that have been in the vignette sample in wave 1 and that have been part of the wave 2 questionnaire[24]. Imposing these conditions further restricts our available sample; only 2,997 respondents[25] remain. We have data on these respondents for both the first and the second wave, so we have 5,994 observations in total. In the end, some more observations may be lost. This is due to the fact that there are some missing values for the individual characteristics, vignette evaluations and self-reports that we will use in estimating our model. Finally, some additional observations will be lost in the simulation procedure. In the end, we obtain corrected latent health levels for around 2,930 respondents; the exact number varies somewhat across health domains. More details on the exact number of observations can be found in the respective sections.

We thus lose a large number of observations because of our information requirements. Therefore, our sample may no longer be representative for the entire population. We only consider a selection of the whole age distribution which also affects the representativeness of our sample. Moreover, we do not incorporate the refreshment sample that is part of the complete SHARE dataset, since a typical individual in the refreshment sample has not been part of the questionnaire in the first wave. However, this is not that much of a concern for the present purpose. We want to test the validity of the vignettes method and then it does not directly matter whether our sample is representative for the population. If the vignettes method does a good job in correcting subjective self-assessments, this must also hold for our subsample of the entire SHARE sample. So therefore we could test the quality of the method for this subsample only and still get reliable results. We will however not be able to draw out-of-sample conclusions about the quality of the vignettes method in correcting self-reports, for instance for individuals in different age ranges. This is something that we have to keep in mind while discussing our test results and avenues for future research.

7.2.2 Data selection and construction of variables

Data from the various modules and the two waves are available in separate files. We combine the data from the two waves and the relevant modules into one dataset, referred to here as the master dataset. The modules that are included in the master dataset are the vignettes module, the physical health module, the mental health module, a selection of the employment and pensions module, the cognitive function module, the objective measured tests grip strength, walking speed, chair stand and peak flow, the household income module and a selection of the coverscreen and drop-off questionnaires. Besides, a number of generated variables modules have been included, namely the health and ISCED module. Finally, some of the imputed values have been added. The most important imputed value is income; next to that we included information on out-of-pocket expenses on prescribed medicines and various types of care. Income is subdivided into several income sources and for each of these sources separate imputed values have been produced.

In the raw SHARE datasets each respondent is identified by a string variable called *mergeid*[26]. This variable is unique for each respondent, but equal across waves. Therefore, merging the data

[24] Since we do not need to obtain corrected health measures for the second wave, we can also incorporate individuals in our sample that have been part of the wave 2 main sample instead of the vignette sample.
[25] These respondents are from Germany, Sweden, the Netherlands, Spain, Italy, France, Greece and Belgium.
[26] The *mergeid* variable has the form "CC-hhhhhh-rr" in which CC stands for the country code consisting of two letters, hhhhhh is a household identifier and rr identifies the respondent within a household (SHARE Guide to Release 2.3.0 of Waves 1 & 2, 2009).

from both waves leads to two rows of observations on one respondent with the same *mergeid* such that the *mergeid* variable is no longer unique. Hence, we constructed a new variable in which the *mergeid* is encoded into a numerical value and multiplied by 100. Thereafter either 1 or 2, depending on whether the data are from wave 1 or wave 2 respectively, is added to this encoded *mergeid*. This results in a unique numerical variable labelled *mergeid3*. To identify waves we generated a variable labelled *wavenumber*. This variable equals 1 if the data are from wave 1 and 2 if the data are from wave 2.

Furthermore, an important feature of the dataset is the way in which missing values are treated. It is possible for a respondent to answer a question with "don't know" or "refuse to answer". Then, the responses are coded as missing values in the dataset. Different codes are assigned for various types of missing values: "don't know" is coded "-1", while a "refusal" is coded "-2". For financial amounts a missing value is indicated by "-9999991" for "don't know" and "-9999992" for a "refusal". Besides, the missing value code "-5", meaning "not answered", is used in the vignette answers (SHARE Guide to Release 2.3.0 Waves 1 & 2, 2009). Since negative values will affect the calculation of descriptive statistics considerably, we converted these negative values into missing values in the dataset, such that they are not considered in calculating descriptive.

We created and manipulated a number of variables in the dataset. First of all, a variable containing the age of the respondent is produced. This is done using information in the SHARE dataset on the month and year of interview for both waves and the month and year of birth of the respondent. We can then calculate the age of an individual at the time of interview[27]. In addition, we constructed dummy variables for various five-year age groups. The first dummy is for individuals aged 50 to 55 years old. Then there is an age group for individuals aged 55 to 60, for those aged 60 to 65, etcetera. The final age group is for individuals aged between 80 and 85 years old.

We are also interested in the educational attainment of respondents. The demographic (DN) module asks individuals about their highest educational degree obtained. These education levels may differ considerably across countries. Therefore we use the international comparable ISCED-coded education levels discussed in section 7.1.4. From this we constructed dummy variables for low, middle and high levels of education, just as is done by Voňková & Hullegie (2010)[28].

Besides, we edited the marital status variable. In the demographic module a variable measuring the marital status of the respondent in the first wave has been included. If a respondent is part of the sample in wave 2 and his marital status has not changed in between the two waves, as indicated by another variable in the same module, the original marital status variable has a missing value in the second wave. This causes an enormous amount of missing values in the original marital status variable, while in fact the marital status of the respondent is known. Therefore we replaced these wave 2 missing values with the corresponding wave 1 values if the marital status has not changed. In addition, we generated a dummy indicating whether someone is married or not. This dummy variable equals "1" if the respondent is either married and living together with a spouse or married and living separated from the spouse or if the respondent has a registered partnership. On the other hand, the dummy variable equals "0" if the respondent indicates to be divorced, widowed or if he has never been married.

[27] The age of an individual is the year of interview minus the year of birth if the interview month is larger than the month of birth (i.e., when the interview month is later in the calendar year than the month of birth). However, if the interview month is smaller than the month of birth, the age of an individual equals the year of the interview minus the year of birth minus 1.

[28] ISCED codes 0 and 1 are labelled low levels of education, ISCED codes 2 and 3 are labelled middle levels of education and ISCED codes 4, 5 and 6 are labelled high levels of education.

Moreover, we created a dummy variable for the gender of respondents. This variable equals 1 if the respondent is female and zero otherwise. Ethnicity is described using the country of birth of respondents. A variable indicating whether the country of birth is the same as the country of interview is used to generate a dummy variable describing ethnicity. This dummy equals "1" if the respondent's country of birth is not equal to the country of interview and zero otherwise.

Furthermore, we constructed an income variable since a single measure for individual income is not directly available. The employment and pensions module asks the respondents an enormous amount of questions on the various income sources, like the period in which income has been received from these sources, the associated exact amounts received or an indication of the approximate amounts received in the previous year. The imputations module of the SHARE dataset translates these into several measures of monthly and annual income from various sources. Adding these amounts allows one to construct an overall measure of individual income. Since all financial amounts in the imputations file are in local currency, we converted them into Euro amounts using the yearly nominal exchange rates matching the year of interview. In constructing a total income amount we considered among others public and private (occupational) old age pension, public and private early or pre-retirement pension, public disability insurance and public unemployment benefits. Besides, annual gross income from employment and self-employment and life insurance payments received, private personal pension and regular payments from charities received have been included. The sum of all these sources then is the total amount of income. From this we also created a variable that equals the logarithm of total income.

7.3 Data description and descriptive statistics

In this section descriptive statistics and graphs will be provided and explained in order to get a good grasp of the dataset. The most important covariates and socio-economic variables that will be used in the empirical tests in section 8 will be discussed here. Besides, the vignette evaluations will be illustrated using descriptive statistics and graphs so as to get an idea of the potential problems with reporting behaviour.

7.3.1 A description of the vignette evaluations

To illustrate a typical vignette evaluation and the differences in reporting, Figure 3 and Figure 4 on pages 40 and 41 show several graphs of vignette evaluations and self-reports for the health domains pain and memory. Figure 5 on page 42 shows similar graphs for work disability. The percentage of respondents choosing a particular response is shown above the bars in the histograms. For wave 1 we can compare multiple vignettes on a health domain. This is no longer possible in wave 2 since only one vignette is included for each health domain then. A comparison is still possible for work disability since three vignettes have been included in wave 2. Figure 5 only shows the work disability vignettes that have been included in both waves. The graphs include all respondents in the vignette sample, i.e., 4,544 respondents in wave 1 and 7,731 in wave 2. There are some differences in the number of respondents for which self-reports and vignette evaluations are observed. On average in wave 1 around 4,517 individuals and in wave 2 7,672 respondents filled in the self-assessments for the two health domains. An outlier is observed for the wave 2 work disability self-assessment which has only been completed by 4,552 individuals. The response rate is a bit smaller for the vignettes. Regarding the two health domains that we consider here, on average 4,460 in wave 1 and 7,621

respondents in wave 2 evaluated the vignettes. Again responses for work disability form an outlier: on average only 4,535 respondents completed these wave 2 vignettes.

From the self-reports in Figure 3 one can see that a bit more difficulties in the pain domain are reported in the second wave. The fraction of respondents reporting no difficulties is smaller in wave 2 (28% instead of 32% in wave 1). In both waves 36% of the respondents reported mild difficulties. The fractions of respondents reporting moderate and severe difficulties have increased: from 23% in wave 1 to 25% in wave 2 and from 7% to 9% respectively. Finally, extreme difficulties were reported by approximately 2% of the respondents in both waves.

The three vignettes in wave 1 represent increasing degrees of (objective) difficulties in the pain domain. The first vignette represents the fewest difficulties, while the third vignette person has much more difficulties. This pattern can also be observed from the wave 1 vignette evaluations. For the first vignette a large fraction of the respondents reported only mild difficulties. For the second vignette however, there is a shift in the distribution towards moderate difficulties. Finally, the histogram of the third vignette shows an even greater shift to the right. For this third vignette the largest fraction of individuals reported severe difficulties. Also important to note is that the evaluation of the second, and to a larger extent, the third vignette is more ambiguous than the evaluation of the first vignette. Although the true health status of the vignette person is fixed, individuals differ considerably in their evaluation. For the first vignette a very large fraction of the respondents chose the same categorical response: almost 57% of the individuals reported mild difficulties, while only 16% reported no difficulties and 22% reported moderate difficulties. The distributions of the second and third vignette show a larger spread. For the second vignette 51% of the respondents reported moderate difficulties, 18% reported mild difficulties and another 26% reported severe difficulties. With respect to the third vignette 49% of the individuals reported severe difficulties, 26% reported moderate difficulties and another 19% reported extreme difficulties. This larger spread in the evaluation of the second and third vignette may point at the possibility that individuals find it easier to evaluate the health status of a vignette person whose health situation is closer to the health status of the respondent himself. Since most of the respondents in the wave 1 vignette sample report none or only mild difficulties, they may be more diverse in evaluating vignette persons with enormous amounts of difficulties.

The wave 2 pain vignette is the same as the first pain vignette in wave 1. Comparing these two vignette evaluations shows no large changes in the rating of the vignette across waves. In wave 1 around 57% evaluated the vignette person as having mild difficulties, 22% argued that the vignette person has moderate difficulties and another 16% reported no difficulties. In wave 2 61% of the respondents reported mild difficulties, 18% reported no difficulties and 18% reported moderate difficulties. So the distribution of the vignette evaluation is somewhat more concentrated in wave 2 compared to wave 1, but these differences are rather small. In addition, while in wave 1 no difficulties were reported more often than moderate difficulties, this is the other way around in wave 2 where more individuals reported none instead of moderate difficulties.

Figure 4 shows similar graphs for the cognitive (memory) domain. Self-reports for this domain have not changed much over the two waves. In the first wave, 44% of the respondents reported no difficulties. In addition, 35% reported mild difficulties and 16% said to have moderate difficulties. Severe difficulties were reported by approximately 4% of the respondents and only around 0.4% of the respondents reported extreme difficulties. In wave 2, around 41% of the respondents reported

themselves to have no difficulties at all. 37% of the respondents said to have mild difficulties. Around 16% of the individuals said to have moderate difficulties. Besides, 4% reported severe difficulties and only 0.8% said to have extreme cognitive difficulties. From the comparison of self-reports in wave 1 and 2 one can see a somewhat larger prevalence of no limitations in wave 1 although the differences are very small.

When evaluating the assessments of the wave 1 memory vignettes one can again observe a shift in the distribution of the vignette evaluations for the first, second and third vignette. Most respondents, i.e., 49%, evaluate the first vignette person to have only mild difficulties. The second vignette person has moderate difficulties according to the largest fraction, i.e., 44%, of the respondents. Finally, the third vignette person is evaluated to have severe difficulties by most respondents, i.e., 47%. Responses are the most dispersed for the second vignette. The memory vignette used in wave 2 is the same as the first vignette from wave 1. In both waves the vignette person is evaluated similarly. In wave 1 49% of the respondents argued that the vignette person has only mild difficulties; in wave 2 53% of the respondents reported mild difficulties. No difficulties were reported by 22% of the respondents in wave 1, in wave 2 the corresponding fraction is 27%. 23% of the respondents reported moderate difficulties for the vignette person in wave 1, in wave 2 17% of the respondents evaluated the vignette person to have moderate difficulties. Severe and extreme difficulties were reported by only 6% and 0.4% of the respondents in wave 1 respectively. The corresponding fractions in wave 2 are 3% and 0.2%.

At last, Figure 5 shows analogous figures for the work disability vignettes[29]. First, self-reports on work disability are rather similar across waves. Around 52% of the respondents in wave 1 against 53% of the wave 2 respondents said to have no difficulties that limit the kind or amount of work they can do. Mild limitations have been reported by 24% of the wave 1 respondents and 25% of the wave 2 respondents. In wave 1 15% of the respondents said to have moderate limitations; the corresponding amount in wave 2 is 13%. Finally, severe and extreme limitations have been reported by 7% and 2% of the wave 1 respondents respectively. In wave 2 the corresponding fractions are 6.5% and 2%. So only minor differences between the self-assessments in wave 1 and wave 2 exist.

Evaluating the three vignettes within wave 1, one could observe that the distribution of responses for the first vignette is somewhat more shifted to the right than the distribution of responses for the second vignette. So, on average the second vignette person is evaluated to be in better health than the first vignette person. The third vignette person is reported to be in even worse health than the first two vignette persons, although there is somewhat less consensus for this person since the categorical responses “moderate” and “severe” represent almost equal fractions of respondents, i.e., 37% and 38% respectively. Besides, 14% of the individuals argued that the vignette person has mild difficulties. No limitations at all and extreme limitations are reported by 2% and 8% of the respondents respectively. For the second wave the conclusions reached from comparing the first and second vignette are similar to those obtained in wave 1. The third vignette person is again evaluated to be in the worst health status. For the first work disability vignette the largest fraction of respondents, i.e., 44%, reported moderate limitations. Besides, a fraction of 39% argued that the vignette person has mild difficulties. 11% of the respondents reported severe limitations. Only 5% of

[29] For the work disability self-report the respondent is asked the following question: “Do you have any impairment or health problem that limits the kind or amount of work you can do?”. The question asked in case of vignette evaluations on work disability is “How much is […] limited in the kind or amount of work he/she can do?” (www.share-project.org, 2010; Type A Vignette Questionnaire Wave 1).

the respondents evaluated the vignette person to have no work limitations. Finally, extreme difficulties have been reported by only 0.4% of the respondents. The evaluation of the second vignette person is already less mixed: most of the respondents, i.e., 47%, evaluated the vignette person to have mild difficulties. The second largest fraction of the respondents, i.e., 33%, reported moderate limitations. Besides, 14% said that the vignette person has no difficulties limiting his work ability. Severe and extreme limitations have been reported by 6.5% and 0.2% respectively. Regarding the third vignette, the largest fraction of respondents, i.e., 45%, reported moderate limitations for the vignette person. In addition, 29% reported severe difficulties and another 21% evaluated the vignette person to have only mild work limitations. No difficulties were reported by only 3% of the respondents and the smallest fraction of respondents, i.e., 2%, evaluated the vignette person to face extreme limitations.

Comparing similar vignettes across waves shows some differences in the distribution of the vignette evaluations. This is especially noticeable for the third work disability vignette: while almost equal fractions of respondents reported moderate and severe limitations in wave 1, a considerable larger fraction of respondents reported moderate limitations in wave 2 (45% responded with “moderate” and only 29% with “severe” in wave 2). There is also a perceptible decline in the fraction of respondents reporting extreme limitations for this vignette: still 8% of the respondents in wave 1 but only 2% in wave 2. This is remarkable since the true health status of the vignette person is the same across waves.

Overall, looking at the graphs in Figure 3 to Figure 5 differences in reporting behaviour are visible. Although the true health status of the vignette person is the same for all respondents, there are considerable differences in reported health levels. For some vignettes there are even two categorical responses representing almost equal fractions of respondents. Systematic reporting differences may lead to serious problems when using uncorrected self-reported health measures.

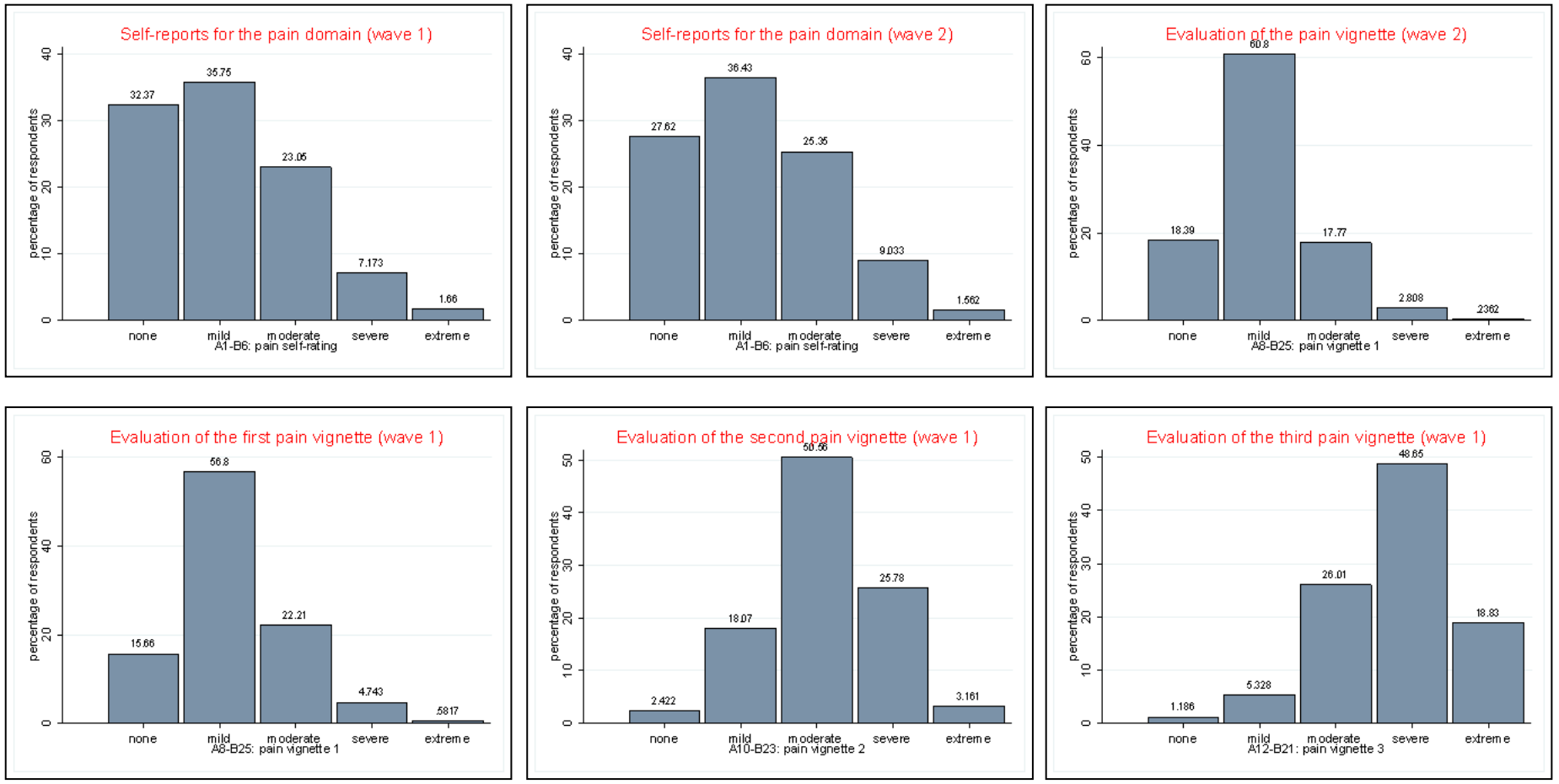

Figure 3: Histograms for self-reports and vignette evaluations on the pain domain.

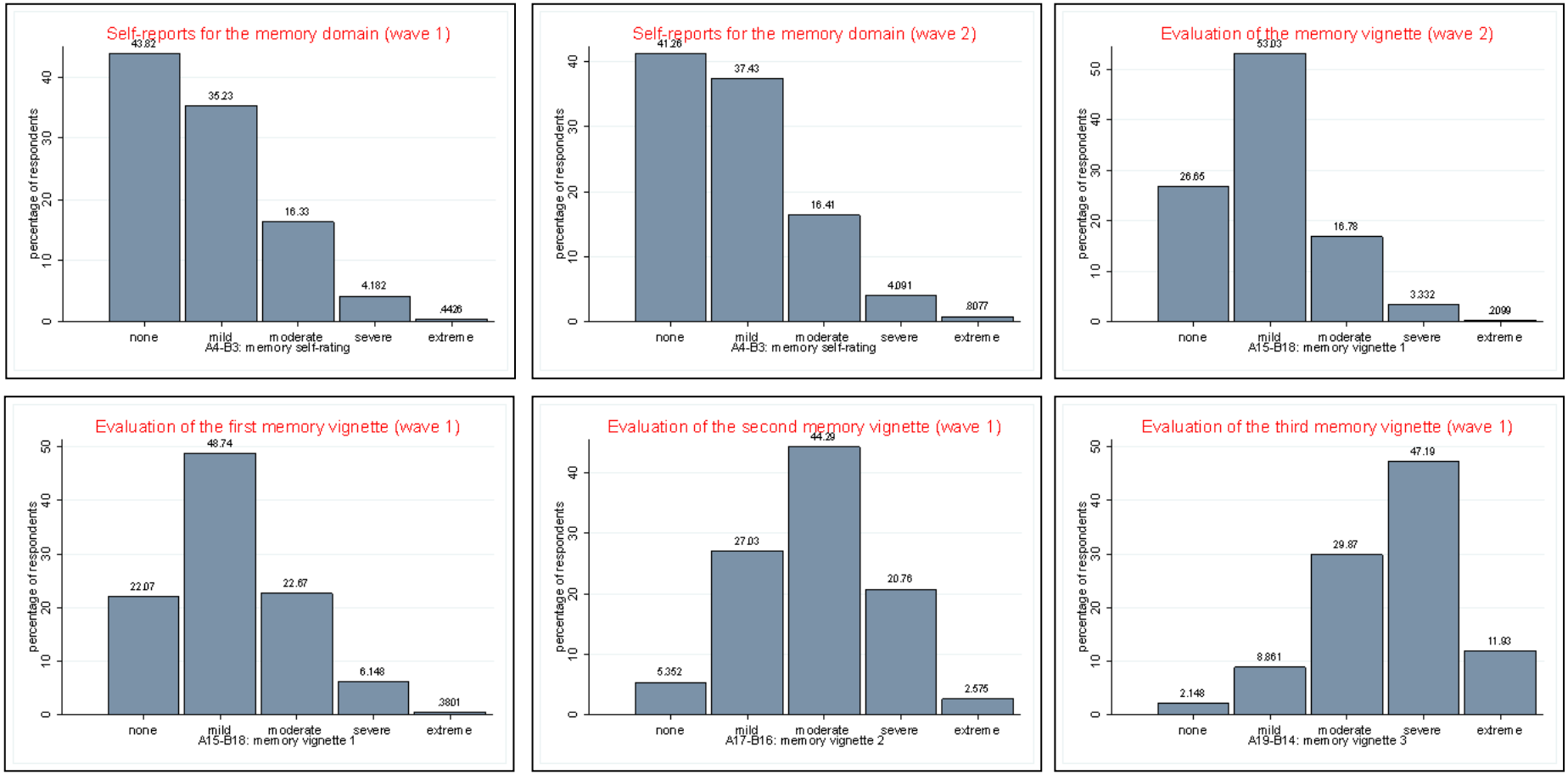

Figure 4: Histograms for self-reports and vignette evaluations on the memory domain.

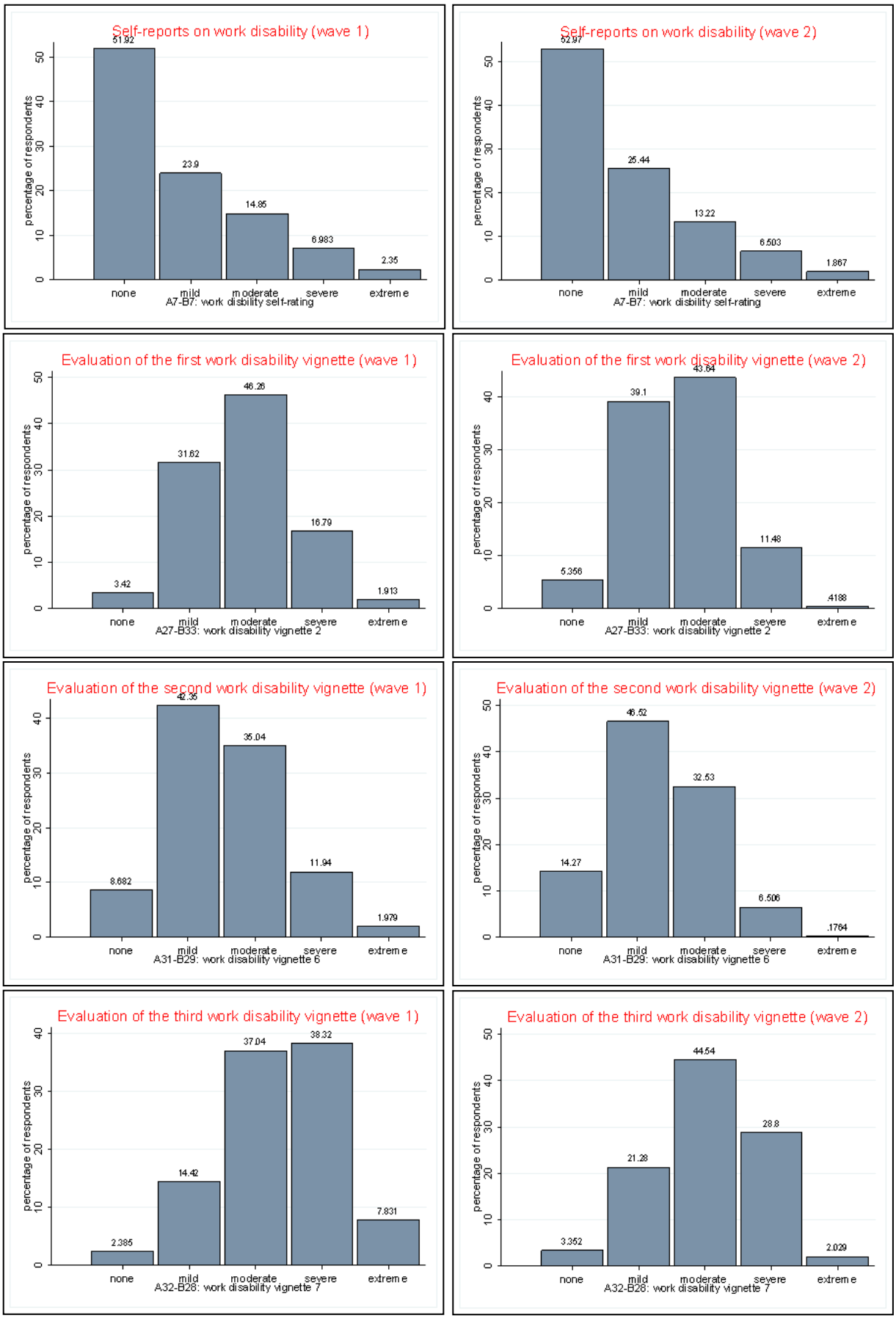

Figure 5: Histograms for self-reports and vignette evaluations on work disability.

It is also interesting to look at differences in the evaluation of vignettes across countries. Therefore, Table 1 on page 48 shows the number and percentage of respondents with a particular categorical response for the pain vignettes and self-reports in the various countries in wave 1[30]. The number of observations and the proportion of respondents reporting a particular category are shown. Table 2 provides such descriptive statistics for wave 2. The vignettes presented to the respondents are the same in all countries (although provided in the own language). Therefore significant cross-country differences in the fraction of individuals choosing a specific category for the vignettes may point at the existence of reporting heterogeneity.

The number of respondents to which vignettes and self-assessments are presented differs across countries. For instance, in France in wave 1 877 respondents answered the self-assessment on pain, while only 415 individuals did so in Sweden. The number of respondents has increased enormously for some countries in the second wave. For example in Germany only 507 respondents evaluated the self-assessment in wave 1, while this has risen to 1,156 respondents in the second wave. On the other hand, the number of respondents in France has decreased considerably, from 877 in wave 1 to only 376 in wave 2. For all countries there are some differences in the number of individuals responding to the self-assessments and the three vignettes within a wave due to missing responses. Overall, these differences are rather small; e.g., in Belgium 564 respondents answered the self-assessment, 560 individuals evaluated the first vignette and the second and third vignettes have been evaluated by 558 and 559 respondents respectively.

The results in the tables point at the presence of reporting heterogeneity. Although the health status of the vignette person is the same across countries, the evaluation of the vignette persons differs considerably across countries. For instance, in wave 1 most of the respondents in Germany, i.e., 54.38%, reported mild difficulties for the first pain vignette. Besides, 21.31% of the individuals evaluated the vignette person as having no difficulties and another 21.31% evaluated the vignette person to have moderate difficulties. Comparing this to the way in which this same vignette has been evaluated in Sweden yields some interesting observations. In Sweden only 3.22% of the respondents reported no difficulties. In addition, 35.40% of the respondents evaluated the vignette person to have mild difficulties, while the largest proportion of individuals, i.e., 39.36%, reported moderate difficulties. Severe difficulties have also been reported by a considerable proportion of the Swedish respondents, i.e., 20.30%. So although in Germany one categorical response clearly represents the largest fraction of the respondents, in Sweden there is much more heterogeneity in reporting: the categories of “mild” and “moderate” difficulties represent similar fractions of individuals. Furthermore, the distribution of responses is somewhat more shifted towards more severe difficulties in Sweden, while the distribution of the responses in Germany lies somewhat more around none, mild or moderate difficulties. So, German respondents are found to evaluate this particular vignette person as having fewer difficulties than Swedish respondents do, although the vignette person represents a fixed true health level. Similar conclusions can be reached for the second pain vignette. In Germany 51.29% of the respondents evaluated the second vignette person to have moderate difficulties, while in Sweden 59.85% of the respondents reported severe difficulties for this vignette. This also holds for the third vignette: 53.57% of the Germans evaluate the vignette person as having severe difficulties; on the other hand 72.52% of the Swedish think that the vignette person has extreme difficulties. This can be interpreted as the Swedish having lower cut-points, i.e.,

[30] These descriptive statistics incorporate data for the entire wave 1 vignette sample.

reporting more limitations than the Germans for a fixed level of health. Relating this to the differences in self-assessments between both countries gives some remarkable insights. In Sweden 52.05% of the respondents said to have no difficulties, 28.92% reported mild difficulties and only 11.33% of the respondents evaluated themselves to have moderate difficulties. In Germany, the distribution of the self-assessments in the pain domain is rather different: only 27.61% of the respondents reported no difficulties, 33.93% said to have mild difficulties and another 26.82% evaluated themselves to have moderate difficulties. So, based on the self-assessments, the Germans seem to have more difficulties in the pain domain than the Swedish respondents. However, this is not corrected for differences in reporting behaviour, such that this probably illustrates something else than true health differences.

Similar comparisons can be done for the wave 1 vignette evaluations and self-reports in other countries. Comparing Spain and Italy for instance, one can see that a larger proportion of the Italians (40.36%) reported mild difficulties than the Spanish respondents (25.92%) did. In Spain on the other hand, more respondents evaluated themselves as having no difficulties (36.72%), while only 25.85% of the Italians chose this category. The evaluation of the first vignette is however quite similar in both countries: in Spain 50.33% of the respondents evaluated the vignette person to have mild difficulties against 48.74% in Italy. Also for the other categorical responses, the proportions of individuals reporting them are comparable in Spain and Italy. However, larger differences between Italy and Spain appear in the other vignette evaluations. For the second and third vignette Italians reported fewer limitations than the Spanish respondents did.

Comparing the vignette evaluations in France and Greece also shows some differences. In France one particular categorical response is reported by a rather large fraction of the respondents for all three vignettes. For the first vignette this is the categorical response “mild” (57.96%), for the second it is “moderate” (57.39%) and for the third “severe” (61.29%). But, in Greece the categorical responses are somewhat more diverse: for the first pain vignette 50.49% evaluated the vignette person as having “mild” difficulties, while another 33.38% said that the vignette person has no difficulties at all. For the second vignette, 47.08% of the respondents reported moderate difficulties, but another 30.22% of the individuals evaluated the vignette person as having only mild difficulties. Besides, an additional 17.97% of the individuals responded with “severe difficulties”. Finally, for the third pain vignette 50.76% of the respondents in Greece reported severe difficulties. Moderate difficulties were reported by 23.92% of the respondents in Greece, against 25.06% of the French respondents. However, in Greece there are also relatively many individuals that reported either mild or extreme difficulties (10.71% and 13.91% respectively), while these categorical responses are much less prevalent among the French respondents (2.71% and 9.06% respectively). Such greater dispersion of responses in Greece is less visible for the self-assessment. 26.80% of the French and 35.47% of the respondents in Greece reported no difficulties. For mild difficulties these proportions are 34.21% and 34.91%. Moderate difficulties were reported by 32.50% of the respondents in France and 20.72% of the respondents in Greece. Finally, severe difficulties have been reported by 5.36% of the respondents in France and 7.09% of the respondents in Greece. The proportions for reporting extreme difficulties are comparable for the two countries. So although the Greek respondents report fewer difficulties, the spread in the distribution seems to be similar in both countries.

Something else that is conspicuous is the very large fraction of respondents reporting mild difficulties for the first pain vignette in Belgium and the Netherlands. In Belgium 71.43% of the respondents reported to have mild difficulties, in the Netherlands this proportion is even larger with 78.69%. The most of the proportions in the table are smaller than 60% such that these amounts of

concentration in one response category are quite remarkable. This is however no longer the case for the second and third pain vignette. The categorical responses even seem more widely dispersed compared to other countries in these cases.

A similar analysis for the wave 2 data is provided in Table 2 on page 49. Here more countries are observed, but only one vignette per health domain is available. First of all, significant differences can be found in the self-assessments on pain across countries. German respondents report somewhat more difficulties compared to Swedish and Dutch respondents. In Sweden 28.27% of the respondents reported no difficulties, against 26.56% of the German respondents and 32.18% of the Dutch respondents. With respect to mild difficulties these proportions are 31.66% of the Germans, 38.82% of the Swedish and 43.68% of the Dutch respondents. Moderate difficulties have been reported by 29.07% of the Germans, 21.31% of the Swedish and 17.43% of the Dutch. Besides, 10.90% of the German respondents, 9.92% of the Swedish respondents and only 5.36% of the Dutch respondents said to have severe difficulties. Finally, the proportions of respondents reporting extreme difficulties are quite comparable for these three countries. The distribution of categorical responses for the Germans is most shifted to the right, i.e., towards more extreme difficulties. The responses of the Swedish are already somewhat more concentrated around mild and no difficulties, but this concentration is even more apparent among the Dutch respondents.

Similar differences in the concentration in the distribution of responses can be observed for the vignette evaluation. First, in Germany 58.46% of the respondents reported the vignette person to have mild difficulties, in Sweden this is reported by 62.13% of the respondents and in the Netherlands an even larger proportion, i.e., 71.79% of the respondents, evaluated the vignette person as having mild difficulties. Although a large amount of the Swedish respondents reported mild difficulties, the proportion of Swedish reporting severe and extreme difficulties is also relatively large compared to the corresponding percentages of German and Dutch respondents. In that respect, Sweden is somewhat more comparable to Spain. In Spain 61.06% of the respondents reported mild difficulties for the vignette person, another 21.53% reported moderate difficulties. 11.35% of the Spanish respondents evaluated the vignette person as having no difficulties. Besides, an additional 5.87% of the Spanish respondents, against 5.32% of the Swedish respondents, said that the vignette person has severe difficulties. It must be added however that a smaller fraction of the Spanish respondents evaluated the vignette person as having no difficulties (11.35% against 15.32% of the Swedish respondents). So the Spanish seem to be inclined to report a bit more extreme difficulties for the vignette person than the Swedish respondents do. Regarding the self-assessment the opposite is observed: in Sweden there is considerably more concentration in the categorical responses on the self-assessment question than in Spain. Italy seems to be somewhat in between the case of Sweden and the case of the Netherlands with respect to the self-assessment. However, comparing the vignette evaluations in Sweden and the Netherlands to that of the Italians leads to a different conclusion. Then the Italians evaluate the vignette person to have much more difficulties. Italians are somewhat more like the Spanish in evaluating the vignette person.

When we contrast the self-assessment of pain in France to that in the aforementioned countries, we see that the French reported more extreme difficulties than the respondents in the other countries. Only 22.07% of the French, against proportions of 26.56% and over for the other countries, reported to have no difficulties at all. Another 40.16% of the French reported mild difficulties. With respect to reporting moderate difficulties, the French lead, since 32.71% of them reported to have such difficulties in the pain domain. However the smallest proportions of French

report to have severe difficulties compared to the other countries evaluated before: only 3.72% of the French chose “severe” against proportions of 5.36% and over for the other countries.

The self-assessment on the pain domain in Denmark is quite comparable to the one in the Netherlands, although the fraction of Dutch respondents reporting mild difficulties is somewhat larger than the corresponding fraction of Danish respondents (i.e., 43.68% against 38.57%). Besides the proportion of respondents in the Netherlands reporting moderate difficulties is somewhat smaller than the corresponding fraction in Denmark (i.e., 17.43% against 22.95%). The way in which the vignette has been evaluated in Denmark is comparable to the way it has been evaluated in Spain. Together with the Italian and Spanish respondents, the Polish respondents are most diverse in their evaluation of the vignette person. Besides, they are the most inclined to report more extreme difficulties: 53.14% of the Polish respondents reported the vignette person to have mild difficulties, 25.13% evaluated the vignette person as having moderate difficulties and 4.85% reported severe difficulties. Czech has a quite large fraction of respondents (25.19%) reporting no difficulties for the vignette person. The self-assessment of the Czech respondents is quite comparable to the one in Germany. Looking at the self-assessment on pain in Belgium, one can conclude that a rather small fraction of Belgians reported no difficulties (only 20.76%). The Belgians reported mild, moderate and severe difficulties more often than respondents in other countries: mild difficulties were reported by 43.10% of the respondents, moderate difficulties by 27.38% and another 6.85% reported severe difficulties. The Belgians are however less extreme in evaluating the vignette person: 67.54% of the respondents evaluated the vignette person as having mild difficulties, 18.91% reported no difficulties at all and another 11.62% reported moderate difficulties. Comparing this way of evaluating the vignette to that in other countries, Belgium looks most like the Netherlands. Finally, Greece is somewhat in between Spain and Denmark regarding the self-assessment on the pain domain. With respect to the vignette evaluation, the proportion of respondents reporting no difficulties for the vignette person is largest in Greece: 28.44% of the respondents evaluated the vignette person as having no difficulties at all. Another 51.93% of the Greek respondents reported mild difficulties. Besides, 16.15% of the respondents in Greece said that the vignette person has moderate difficulties.

For those countries that have been in the vignette sample in both waves, it is interesting to look at the differences and similarities in self-assessments and vignette evaluations across waves. In Germany, the self-assessment has moved somewhat more towards moderate and severe difficulties in wave 2, although the differences are only slight. In Sweden the differences between wave 1 and wave 2 self-assessments are a bit larger. While 52.05% of the Swedish respondents reported no difficulties at all in wave 1, only 28.27% of them did so in wave 2. Furthermore, there has been an increase in the proportion of Swedish respondents reporting mild and moderate difficulties. Regarding mild difficulties the proportion has increased from 28.92% in wave 1 to 38.82% in wave 2. For moderate difficulties this increase has been from 11.33% to 21.31%. There has also been an increase of three percentage points for reporting severe difficulties. So from the self-assessments one would be inclined to think that the Swedish have become less healthy with respect to the domain of pain. However, changes in reporting behaviour over time may also be responsible for such a tendency. This possibility is supported by the observed changes in the vignette evaluations over time. Although the vignette person is the same in both waves, differences in the evaluation can be observed across waves. 3.22% of the Swedish reported no difficulties for the vignette person in wave 1, 15.32% of them did so in wave 2. For mild difficulties, the proportion has increased from 35.40% in wave 1 to 62.13% in wave 2. However the fraction of Swedish respondents reporting moderate and

severe difficulties for the vignette person has declined between wave 1 and wave 2. In wave 1 39.36% reported moderate difficulties and 20.30% reported severe difficulties. The corresponding fractions in wave 2 are 16.81% and 5.32%. So although Swedish respondents reported more difficulties for themselves in wave 2, they also report fewer difficulties for the vignette person.

In the Netherlands the changes are smaller in magnitude. For the self-assessments, the fraction of respondents reporting no difficulties has decreased somewhat in wave 2 (from 37.10% in wave 1 to 32.18% in wave 2). On the other hand, there has been an increase in the proportion of respondents reporting moderate difficulties: from 13.75% in wave 1 to 17.43% in wave 2. Thus, the Dutch also reported more difficulties in wave 1 compared to wave 2. With respect to the vignettes, one can observe that the fraction of Dutch respondents evaluating the vignette person as having mild difficulties has decreased from 78.69% in wave 1 to 71.79% in wave 2. Besides, the fraction of respondents reporting no difficulties for the vignette person has increased from 11.59% in wave 1 to 18.62% in wave 2. So, just as in Sweden one can observe a tendency towards reporting more difficulties in the self-assessments and fewer difficulties for the vignettes.

For Spain the differences in the self-assessments across waves are rather small. As regards the vignette evaluation, there has been a considerable increase in the fraction of respondents reporting mild difficulties (from 50.33% in wave 1 to 61.06% in wave 2), while there has been a decrease in the fraction of respondents reporting moderate difficulties (from 28.63% in wave 1 to 21.53% in wave 2). Thus, the Spanish reported fewer difficulties for the vignette person in wave 2 compared to wave 1. A similar conclusion can be reached for the way in which the Italians evaluated the vignette person in waves 1 and 2. In France, the distribution of responses has become more concentrated in wave 2, since the fraction of respondents reporting none or severe difficulties has decreased, while the fraction reporting mild difficulties, and to a lesser extent moderate difficulties, has increased considerably. The same holds for the evaluation of the vignette person in waves 1 and 2 in France. Only minor changes between waves 1 and 2 exist for the self-assessment and vignette evaluation in Greece. Finally, for Belgium there has been a slight movement in the self-assessment towards reporting more difficulties in wave 2. For the vignette evaluation the fraction of respondents reporting no difficulties has increased from 11.07% in wave 1 to 18.91% in wave 2. On the other hand, the fractions of respondents reporting mild, moderate and severe difficulties have decreased considerably in wave 2 compared to wave 1. Regarding mild difficulties, the decrease has been from 71.43% in wave 1 to 67.54% in wave 2. The proportion of respondents reporting moderate difficulties has decreased from 15.36% in wave 1 towards 11.62% in wave 2. Finally, the fraction of respondents evaluating the vignette person as having severe difficulties has decreased from 2.14% in wave 1 to 1.94% in wave 2. So in Belgium there has been a trend towards reporting more difficulties in the self-assessment, but fewer difficulties for the vignette person. This is comparable to the tendencies observed for the Netherlands and Sweden.

Table 1: Evaluation of vignettes and self-assessments on pain per country in wave 1.

	Germany		Sweden		the Netherlands		Spain		Italy		France		Greece		Belgium	
	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.
Self-assessment																
None	140	27.61	216	52.05	197	37.10	170	36.72	114	25.85	235	26.80	255	35.47	135	23.94
Mild	172	33.93	120	28.92	231	43.50	120	25.92	178	40.36	300	34.21	251	34.91	243	43.09
Moderate	136	26.82	47	11.33	73	13.75	104	22.46	106	24.04	285	32.50	149	20.72	141	25.00
Severe	49	9.66	28	6.75	20	3.77	63	13.61	29	6.58	47	5.36	51	7.09	37	6.56
Extreme	10	1.97	4	0.96	10	1.88	6	1.30	14	3.17	10	1.14	13	1.81	8	1.42
Total	**507**		**415**		**531**		**463**		**441**		**877**		**719**		**564**	
Pain vignette 1																
None	107	21.31	13	3.22	62	11.59	58	12.58	43	9.89	115	13.47	240	33.38	62	11.07
Mild	273	54.38	143	35.40	421	78.69	232	50.33	212	48.74	495	57.96	363	50.49	400	71.43
Moderate	107	21.31	159	39.36	42	7.85	132	28.63	148	34.02	216	25.29	103	14.33	86	15.36
Severe	15	2.99	82	20.30	5	0.93	37	8.03	28	6.44	24	2.81	9	1.25	12	2.14
Extreme	0	0.00	7	1.73	5	0.93	2	0.43	4	0.92	4	0.47	4	0.56	0	0.00
Total	**502**		**404**		**535**		**461**		**435**		**854**		**719**		**560**	
Pain vignette 2																
None	4	0.80	4	1.00	9	1.70	9	1.95	16	3.66	32	3.76	22	3.06	12	2.15
Mild	62	12.33	16	3.99	74	13.96	72	15.62	96	21.97	151	17.72	217	30.22	118	21.15
Moderate	258	51.29	99	24.69	254	47.92	263	57.05	248	56.75	489	57.39	338	47.08	306	54.84
Severe	168	33.40	240	59.85	158	29.81	111	24.08	64	14.65	169	19.84	129	17.97	111	19.89
Extreme	11	2.19	42	10.47	35	6.60	6	1.30	13	2.97	11	1.29	12	1.67	11	1.97
Total	**503**		**401**		**530**		**461**		**437**		**852**		**718**		**558**	
Pain vignette 3																
None	3	0.60	1	0.25	7	1.32	2	0.43	13	2.97	16	1.88	5	0.70	6	1.07
Mild	26	5.16	8	1.98	17	3.20	21	4.54	42	9.61	23	2.71	77	10.71	24	4.29
Moderate	147	29.17	11	2.72	146	27.50	138	29.81	175	40.05	213	25.06	172	23.92	160	28.62
Severe	270	53.57	91	22.52	212	39.92	275	59.40	154	35.24	521	61.29	365	50.76	285	50.98
Extreme	58	11.51	293	72.52	149	28.06	27	5.83	53	12.13	77	9.06	100	13.91	84	15.03
Total	**504**		**404**		**531**		**463**		**437**		**850**		**719**		**559**	

Table 2: Assessment of vignette evaluations and self-reports on pain per country in wave 2.

	Germany		Sweden		the Netherlands		Spain		Italy		France		Denmark	
	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.
Self-assessment														
None	307	26.56	134	28.27	168	32.18	168	32.50	198	28.61	83	22.07	326	31.84
Mild	366	31.66	184	38.82	228	43.68	151	29.21	284	41.04	151	40.16	395	38.57
Moderate	336	29.07	101	21.31	91	17.43	116	22.44	147	21.24	123	32.71	235	22.95
Severe	126	10.90	47	9.92	28	5.36	71	13.73	52	7.51	14	3.72	60	5.86
Extreme	21	1.82	8	1.69	7	1.34	11	2.13	11	1.59	5	1.33	8	0.78
Total	**1,156**		**474**		**522**		**517**		**692**		**376**		**1,024**	
Pain vignette														
None	248	21.51	72	15.32	97	18.62	58	11.35	80	11.58	58	15.47	148	14.65
Mild	674	58.46	292	62.13	374	71.79	312	61.06	386	55.86	230	61.33	644	63.76
Moderate	193	16.74	79	16.81	42	8.06	110	21.53	195	28.22	81	21.60	204	20.20
Severe	38	3.30	25	5.32	7	1.34	30	5.87	27	3.91	4	1.07	11	1.09
Extreme	0	0.00	2	0.43	1	0.19	1	0.20	3	0.43	2	0.53	3	0.30
Total	**1,153**		**470**		**521**		**511**		**691**		**375**		**1,010**	

	Greece		Belgium		Czech		Poland	
	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.	freq.	% of resp.
Self-assessment								
None	183	33.58	185	20.76	235	25.49	135	23.94
Mild	187	34.31	384	43.10	340	36.88	129	22.87
Moderate	121	22.20	244	27.38	255	27.66	179	31.74
Severe	49	8.99	61	6.85	89	9.65	97	17.20
Extreme	5	0.92	17	1.91	3	0.33	24	4.26
Total	**545**		**891**		**922**		**564**	
Pain vignette								
None	155	28.44	166	18.91	229	25.19	90	16.16
Mild	283	51.93	593	67.54	549	60.40	296	53.14
Moderate	88	16.15	102	11.62	120	13.20	140	25.13
Severe	17	3.12	17	1.94	11	1.21	27	4.85
Extreme	2	0.37	0	0.00	0	0.00	4	0.72
Total	**545**		**878**		**909**		**557**	

7.3.2 Covariates and socio-economic variables

The SHARE dataset includes a large number of socio-economic variables that can be used as control variables in the testing procedure that we will develop in section 8. In this section basic descriptive statistics for these demographic and socio-economic covariates will be provided. The descriptives are provided for those individuals in our constructed sample as discussed in section 7.2.1, so in total 2,997 individuals are considered in each wave. Descriptive statistics for wave 1 are provided in Table 3, both per country as well as for our whole sample. Table 4 shows similar descriptive statistics for wave 2. The variables for which these statistics are tabulated are age, female, the logarithm of income, education level, marital status, household size, ethnicity and employment status. Some of these covariates are described by their mean and standard deviation, for others the proportion of respondents reporting a particular category is shown.

The mean age of respondents in our sample is 63.0 years old in wave 1 and 65.0 in wave 2[31]. There are some differences in the mean age between countries. The mean is lowest in the Netherlands in wave 1, with an average of 62.04 years old. In wave 2 the average age is lowest in the Greece, i.e., 64.37 years old. The highest average is observed in Spain in wave 1 (64.06 years old) and wave 2 (66.37 years old). The variation in the standard deviations is rather small between waves and between countries.

The female dummy equals 1 if a respondent is female and zero otherwise. Table 3 and Table 4 show that in all countries the fraction of males is smaller than the fraction of females. In both waves, the overall average fraction of males in the sample is 46.01%. There are some differences in the fractions of males and females in the sample across countries. In Italy and France a high fraction of 57.19% of the respondents is female. On the other hand, only 50.10% of the respondents is female in Greece.

The logarithm of income differs somewhat between countries: in wave 1 the average logarithm of total individual income ranges from 6.9562 in Spain to 8.5091 in Germany. The standard deviation of the logarithm of income is very low in Sweden, while it is quite high in Spain. In wave 2, comparable differences between countries can be observed. Overall, the average logarithm of income is lower in wave 2 compared to wave 1[32]. Only in France the average logarithm of income has increased in wave 2 compared to wave 1. The magnitude of the decrease in the other countries differs somewhat. The logarithm of income is on average lower in the Mediterranean countries compared to the Nordic countries.

The education level of respondents is described using the educational dummies for low, middle and high levels of education discussed in section 7.2.2. In wave 1, on average 35.60% of the respondents reported low education levels, 42.99% of them said to have a middle level of education and another 21.41% reported high levels of education. However, there are remarkable differences in education levels between countries. For instance, in Germany only 1.12% of the respondents reported to have a low education level, while this fraction equals 66.11% in Spain. On the other hand, 73.61% of the Germans said to have a middle level of education and 25.28% of them said to have a high education level. Thus, the distribution of education levels in Germany is skewed towards

[31] It is logical that the average is higher in the second wave, since we consider the same respondents two years later. However, the average is not exactly two years higher for all countries, the reason for this being that there are some differences in the interview month and year across countries and across individuals, such that some individuals may be interviewed within two years, while others have to wait a bit longer before the next questionnaire to be conducted.

[32] This may result (in part) from gross income being used in wave 1 and net income in wave 2.

high levels of education. Alternatively, in Sweden the distribution is much more equal: 32.04% has a low education, 34.51% a middle level of education and 33.45% has a high level of education in wave 1. Though the distribution of education levels is skewed to the right in Germany, it is more skewed towards lower education levels in Italy. Similar cross-country differences in education levels can be observed in wave 2. Still, low education levels are much more common in the Mediterranean countries, while especially Germany has almost no individuals with low education. The Nordic countries again show the largest proportions of individuals with a high education level. It is interesting to see such large cross-country differences in education levels, while an international comparable coding of education levels has been used. This points at enormous differences exist in the true educational attainment of individuals across European countries.

The household size variable shows the magnitude of the household, including the respondent himself. The household size ranges from 1 to 9 in both waves. The average size of a household in wave 1 is 2.23 individuals with a standard deviation of 1.00. In wave 2, the average household size equals 2.14 with a standard deviation of 0.92. The differences in household size between countries are not very large. In wave 1 one can see that the mean (1.95) and standard deviation (0.71) of the household size variable are rather small in Sweden and higher in Spain (2.80 and 1.27 respectively) and Italy (2.57 and 1.02 respectively). In wave 2 the average household size has decreased in each country.

The marital status variable has 6 possible values. A value of 1 means that the respondent is married and lives together with his/her spouse. When the marital status variable equals 2, the respondent has a registered partnership, 3 means married and living separated from his/her spouse. A value of 4 means “never married’, while a value of 5 means “divorced”. Finally, a value of 6 implies “widowed”. The fractions of respondents choosing a particular category are shown in Table 3 and Table 4. One can observe that most of the respondents, i.e., around 73% in both waves, are married and living together with a spouse. In addition, a considerable part of the respondents reported to be widowed: 12.7% in wave 1 and 13.4% in wave 2, which is to be expected considering the age distribution of the sample. There are substantial differences in the marital status across countries. For instance, in wave 1 80.79% of the Dutch respondents reported to be married and living together, while only 67.38% of the French were married and living together with a spouse. Comparable differences are observed in wave 2. Also for the other categorical responses considerable differences between countries can be observed. For example, 15.89% of the French were widowed in wave 1, but only 7.63% of the Dutch reported to be so. In addition, the registered partnership is absent in countries like Germany and Spain in wave 1, although 8.45% of the Swedish respondents said to have a registered partnership. These differences are similar in wave 2. Finally, divorce is much more common in some countries than in others. For instance in wave 1 only 1.26% of the Spanish reported to be divorced, while a considerable fraction of the French (9.93%) were divorced at that time. This can also be observed in wave 2: 1.67% of the Spanish respondents against 9.60% of the French reported to be divorced.

The ethnicity variable discussed in section 7.2.2 can have two different values: “yes” if the respondent was born in the country of interview and “no” if the interview country is not the country of birth of the respondent. Both in wave 1 as well as in wave 2 the fraction of respondents of which the country of birth equals the country of interview is 93.16%. Some differences in these fractions can be observed between countries. For instance, in wave 1 85.56% of the Germans said that the interview and birth country coincide, while 98.15% of the Greek respondents did so.

Finally, the employment status variable has 6 possible values: “1” for retired individuals, “2” for employed or self-employed individuals, “3” for unemployed individuals. A permanently sick or disabled respondent is assigned a value of “4”. A value of “5” means that the respondent is a homemaker and finally, a value of “97” means that the individual has some other employment status then those five listed before. Information on employment status is missing for 3 individuals in wave 1 and 17 individuals in wave 2. In both waves a significant part of the respondents reported to be retired: 45.39% in wave 1 and 51.85% in wave 2. This is to be expected since we only look at those individuals aged between 50 and 85 years old. In addition, in wave 2 we evaluate the same respondents two years later, such that retirement rates will be higher then. Considerable differences in the retirement rates can be observed in the various countries: only 34.31% of the Spanish respondents said to be retired in the first wave, while 53.75% of the Italians reported retirement as the current employment status[33]. The retirement rates are higher in wave 2 in all countries. There are however some differences in the magnitude of the increase across countries. In Germany the retirement rate increases with almost 11 percentage points. On the other hand, the increase is only 2.26 percentage points in Greece. This may be due to differences in the age distribution of individuals. Still, large differences in retirement rates across countries exist in the second wave. For instance, only 40.04% of the Greek respondents is retired, while 60.31% of the Italian respondents reported to be retired in the second wave. Overall, the second largest part of the respondents reported to be employed or self-employed: 30.13% in wave 1 and 25.77% in wave 2. Again, serious differences between countries can be found in the proportions of employed individuals. Also being a homemaker is reported quite often: overall, 17.84% of the respondents reported to be a homemaker in wave 1 against 16.01% in wave 2. There are again enormous differences between countries in the fraction of respondents being a homemaker: in wave 1 only 0.70% of the Swedish respondents chose this category, while 30.96% of the Spanish respondents said to be a homemaker. Also in Greece, Italy and the Netherlands the proportion of respondents reporting to be a homemaker is quite large. In wave 2 such large differences between countries can still be observed.

[33] These differences in fractions of retired individuals may result to some extent because of cross-country differences in the normal retirement age.

Table 3: Descriptive statistics of socio-economic covariates (wave 1).

	Age			Female			Log(income)			Education level				Household size		
	# obs.	mean	st.dev	# obs.	mean	st.dev	# obs.	mean	st.dev	# obs.	low (%)	middle (%)	high (%)	# obs.	mean	st.dev
All	2,997	62.9960	8.8668	2,997	0.5399	0.4985	2,997	8.0202	3.3757	2,980	35.60%	42.99%	21.41%	2,997	2.2332	0.9950
DE	270	62.7630	8.2896	270	0.5556	0.4978	270	8.5091	3.1633	269	1.12%	73.61%	25.28%	270	2.0815	0.7522
SE	284	63.0211	8.2067	284	0.5141	0.5007	284	7.6394	1.1101	284	32.04%	34.51%	33.45%	284	1.9507	0.7116
NL	354	62.0367	8.2843	354	0.5169	0.5004	354	8.3339	3.5602	350	14.57%	60.86%	24.57%	354	2.1497	0.8332
ES	239	64.0586	9.2098	239	0.5356	0.4998	239	6.9562	3.8864	239	66.11%	23.43%	10.46%	239	2.7992	1.2741
IT	320	63.1000	8.4347	320	0.5719	0.4956	320	7.4993	3.7169	320	57.50%	35.31%	7.19%	320	2.5719	1.0238
FR	605	63.7091	9.5172	605	0.5719	0.4952	605	8.7967	3.0008	596	40.10%	37.75%	22.15%	605	2.0942	1.0145
GR	487	62.3142	9.0074	487	0.5010	0.5005	487	7.5360	3.7573	486	46.91%	33.74%	19.34%	487	2.3388	1.1084
BE	438	63.0160	9.0264	438	0.5434	0.4987	438	8.1393	3.4960	436	24.54%	49.08%	26.38%	438	2.0959	0.8503

	Ethnicity			Marital status							Employment status						
	# obs.	yes (%)	no (%)	# obs.	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)	# obs.	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	97 (%)
All	2,996	93.16%	6.84%	2,995	73.19%	1.40%	1.14%	5.118%	6.44%	12.65%	2,994	45.39%	30.13%	3.87%	2.40%	17.84%	0.37%
DE	270	85.56%	14.44%	270	80.00%	0.00%	1.48%	4.81%	4.81%	8.89%	270	48.52%	30.37%	5.56%	2.96%	12.59%	0.00%
SE	284	93.31%	6.69%	284	70.07%	8.45%	0.00%	4.58%	8.45%	8.45%	284	46.83%	44.37%	3.87%	3.87%	0.70%	0.35%
NL	354	96.33%	3.67%	354	80.79%	2.54%	0.85%	1.98%	6.21%	7.63%	354	37.01%	31.36%	1.41%	3.67%	25.14%	1.41%
ES	239	96.23%	3.77%	239	78.66%	0.00%	2.09%	5.86%	1.26%	12.13%	239	34.31%	25.52%	5.02%	3.35%	30.96%	0.84%
IT	320	98.75%	1.25%	319	74.92%	0.31%	0.63%	6.58%	2.51%	15.05%	320	53.75%	17.50%	1.56%	1.56%	25.63%	0.00%
FR	604	87.09%	12.91%	604	67.38%	0.00%	1.16%	5.63%	9.93%	15.89%	602	52.99%	30.73%	3.99%	1.50%	10.80%	0.00%
GR	487	98.15%	1.85%	487	69.82%	0.62%	1.44%	8.01%	4.72%	15.40%	487	37.78%	34.29%	2.26%	1.03%	24.02%	0.62%
BE	438	92.24%	7.76%	438	72.37%	1.14%	1.37%	3.20%	9.13%	12.79%	438	47.26%	26.03%	7.53%	2.97%	16.21%	0.00%

Note: the following country abbreviations are used: All countries (All), Germany (DE), Sweden (SE), the Netherlands (NL), Spain (ES), Italy (IT), France (FR), Greece (GR) and Belgium (BE).

Table 4: Descriptive statistics of socio-economic covariates (wave 2).

	Age			Female			Log(income)			Education level				Household size		
	# obs.	mean	st.dev	# obs.	mean	st.dev	# obs.	mean	st.dev	# obs.	low (%)	middle (%)	high (%)	# obs.	mean	st.dev
All	2,997	65.0557	8.8495	2,997	0.5399	0.4985	2,997	7.7593	3.5032	2,981	35.52%	43.01%	21.47%	2,997	2.1408	0.9234
DE	270	65.2815	8.3007	270	0.5556	0.4978	270	8.2196	3.1536	269	1.12%	73.23%	25.65%	270	1.9815	0.6592
SE	284	65.1831	8.2115	284	0.5141	0.5007	284	7.0807	1.5156	284	32.04%	34.51%	33.45%	284	1.9190	0.6153
NL	354	64.6582	8.2643	354	0.5169	0.5004	354	8.3965	3.2373	350	14.57%	60.86%	24.57%	354	2.0452	0.7284
ES	239	66.3724	9.2035	239	0.5356	0.4998	239	6.1253	4.3647	239	66.11%	23.01%	10.88%	239	2.5732	1.1161
IT	320	65.4969	8.3935	320	0.5719	0.4956	320	7.4176	3.7306	320	57.50%	35.31%	7.19%	320	2.5125	1.0232
FR	605	65.1141	9.5256	605	0.5719	0.4952	605	8.8297	2.8600	597	39.87%	37.86%	22.28%	605	1.9769	0.9522
GR	487	64.3696	8.9942	487	0.5010	0.5005	487	6.8227	4.1671	486	46.91%	33.74%	19.34%	487	2.3039	1.0953
BE	438	64.7968	8.9918	438	0.5434	0.4987	438	8.1049	3.4530	436	24.31%	49.54%	26.15%	438	1.9977	0.7270

	Ethnicity			Marital status							Employment status						
	# obs.	yes (%)	no (%)	# obs.	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)	# obs.	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	97 (%)
All	2,997	93.16%	6.84%	2,993	72.30%	1.40%	1.17%	5.11%	6.58%	13.43%	2,980	51.85%	25.77%	2.21%	2.89%	16.01%	1.28%
DE	270	85.56%	14.44%	270	78.15%	0.00%	1.48%	4.81%	4.44%	11.11%	269	59.48%	22.30%	4.46%	2.60%	9.29%	1.86%
SE	284	93.31%	6.69%	282	69.86%	8.87%	0.00%	4.61%	9.22%	7.45%	281	57.30%	37.72%	1.07%	2.85%	0.36%	0.71%
NL	354	96.33%	3.67%	354	79.66%	1.98%	1.13%	1.98%	6.50%	8.76%	352	45.17%	26.14%	0.28%	4.83%	22.73%	0.85%
ES	239	96.23%	3.77%	239	76.57%	0.00%	1.67%	5.86%	1.67%	14.23%	239	40.59%	19.67%	2.09%	5.86%	29.29%	2.51%
IT	320	98.75%	1.25%	319	74.29%	0.31%	0.63%	6.58%	2.82%	15.36%	320	60.31%	13.44%	0.63%	2.81%	22.81%	0.00%
FR	605	87.11%	12.89%	604	66.56%	0.17%	1.32%	5.30%	9.60%	17.05%	594	59.09%	27.27%	2.69%	1.18%	7.91%	1.85%
GR	487	98.15%	1.85%	487	68.99%	0.62%	1.44%	8.01%	5.13%	15.81%	487	40.04%	33.26%	0.41%	1.44%	24.44%	0.41%
BE	438	92.24%	7.76%	438	72.15%	1.14%	1.37%	3.20%	9.13%	13.01%	438	52.28%	21.92%	5.71%	3.88%	14.16%	2.05%

Note: the following country abbreviations are used: All countries (All), Germany (DE), Sweden (SE), the Netherlands (NL), Spain (ES), Italy (IT), France (FR), Greece (GR) and Belgium (BE).

7.3.3 Objective health measures and health limitations

A couple of objective health measures are included in the SHARE dataset. The objective tests that have been conducted are the walking speed test, the grip strength test, the chair stand test, the peak flow test and a couple of tests on cognitive functioning. The latter consists of several parts: verbal fluency, numeracy (mathematical performance), immediate recall and delayed recall of a list of words. The results of these cognitive tests are summarized in Table A1 - 1 in Appendix 1 on page 88. The verbal fluency score ranges from 0 to 100. A higher value means a higher degree of cognitive function. The score on immediate and delayed recall lies between 0 and 10. Mathematical performance is measured using the numeracy score which ranges from 1 ("bad") to 5 ("good"). Some cross-country differences in cognitive functioning can be observed. For instance, the verbal fluency score is noticeably lower in the Mediterranean countries of Italy, Spain and Greece, while it is especially high in Sweden. For immediate and delayed recall the lower scores are observed in Spain and Italy, while the highest scores appear in Germany, the Netherlands and Sweden. Finally, for the numeracy score similar patterns are observed: low scores in Spain and Italy and higher scores in Sweden, the Netherlands and Germany.

Table A1 - 2 on page 88 shows the results of four other objective tests. The range of the grip strength measure is from 3 up to 73. The walking speed measure has a range from approximately 0.0980 to 4.1322 meters per second. Outcomes for the peak flow test range from 32 to 870[34]. For the peak flow test two measurements have been recorded, in Table A1 - 2 the averages of these two measurements are used. Finally, results for the chair stand test range from 0.08 to 97[35].

The number of observations for the walking speed measure is very low, i.e., 264 in wave 1 and 290 in wave 2, which makes it difficult to draw firm conclusions based on these descriptive statistics. The reason for this small amount of observations is that the walking speed test has only been conducted among those individuals aged 76 and older (SHARE Guide to Release2.3.0 Waves 1 & 2, 2009). The walking speed is measured by dividing the sum of the distances of the two measurements by the time needed to walk these distances. Only those individuals who needed between 0.54 and 30 seconds are included in the variable *wspeed* (SHARE Guide to Release 2.3.0 Waves 1 & 2, 2009). A high walking speed indicates a higher mobility level of the respondent. In wave 1 higher levels are observed in the Netherlands, Germany and Sweden. Lower scores are observed in Spain, Greece and Italy. In wave 2, the relative positions of countries have changed a bit. Low scores are now observed in Italy and Greece and high scores in Sweden, Spain and France. For the grip strength test Germany and Sweden score high, while for the Mediterranean countries lower scores can be observed. The chair stand test and the peak flow test have only been conducted in the second wave. The chair stand test, described in section 7.1.1, is measured as the number of seconds it took the respondent to perform the test five times. A lower score on this test indicates higher ability. High scores are observed in for instance Greece, France and Spain, while respondents in Germany, and Italy do a better job in general, i.e., have lower scores. For the peak flow test high scores can be observed in Germany and Italy, while low scores appear in France and Belgium.

[34] The raw measurements range up to 999. However, 999 is recorded if the respondent chose not to perform the test; 993 is recorded if the respondent was unable to do the test and 890 is recorded for very high measurements. Moreover, 30 is measurements below 60. We changed these four different values into missing values. A couple of observations below 60 though unequal to 30 have been recorded, we kept these observations.

[35] The code 99 was used when it took the respondent more than one minute. Three observations larger than 60 though unequal to 99 have been recorded. We changed a value of 99 into a missing value.

Finally, one could look at the health status of the respondents in our sample. Therefore, Table A1 - 3 and Table A1 - 4 on page 89 provide descriptive statistics on the prevalence of several health limitations or health problems. Difficulties with activities of daily living and instrumental activities of daily living, which have been described in section 7.1.4, are considered. Also the number of mobility limitations is illustrated in this table. In addition, the prevalence of chronic diseases[36] and long-term illness is shown. Moreover, the occurrence of depression according to the so-called EURO-D scale[37] is illustrated in both waves.

Regarding ADL limitations, in wave 1 92.55% of the respondents reported zero limitations. In wave 2 this proportion decreased slightly to 91.54%. There are slight differences in the average number of limitations and the associated standard deviations per country. However, in all countries and in both waves by far the largest part of the sample indicated to have no limitations at all. Difficulties with instrumental activities of daily living have been reported a bit more in both waves, but still the largest fraction of respondents reported zero limitations. In wave 1, the number of ADL limitations is somewhat larger in Belgium, Spain and Italy. The average number of ADL limitations has increased in most countries in wave 2, except for Sweden. The average number of IADL limitations is quite high in Spain and Italy. The average number of reported mobility limitations is larger than the average numbers of IADL and ADL limitations and it is highest in Spain and Italy. Regarding mobility and IADL limitations, the average number of reported limitations has increased for all countries in wave 2 compared to the first wave. The reported number of chronic diseases in wave 1 is highest in Spain and Italy and lowest in the Netherlands and Germany. Regarding the fraction of respondents with a long-term illness, in wave 1 the highest rates can be found in Germany (58.52%) and Sweden (55.28%). Also in wave 2 these countries have the largest fractions of individuals reporting long-term illnesses. The lowest rates can be observed in Greece and the Netherlands with 28.75% and 37.25% of the respondents reporting long-term illnesses respectively. Changes in between waves are quite modest. Finally, depression scores can range from 0 (not depressed) to 12 (depressed). The average score differs somewhat between countries. It is rather low in the Netherlands and Sweden in wave 1, while it is quite high in Italy, France and Spain. In wave 2, for some countries the average score has decreased, while in other countries it has increased. Low scores are reported in the Netherlands, Sweden and Greece and high scores especially in Spain and Italy.

[36] The prevalence of chronic diseases is only illustrated for wave 1, since the assessment of chronic diseases has not been part of the questionnaire in wave 2.

[37] More details on this scale can be found in the SHARE Guide to Release 2.3.0 Waves 1 & 2 (2009). For now it suffices to mention that a high score on this scale illustrates the prevalence of depression.

8. Testing the appropriateness of the vignettes approach

Now that the SHARE data have been explained and descriptive statistics of the data have been provided, it is time to discuss the way in which the value of the vignettes method can be assessed. The theory discussed in sections 3 to 6 illustrated the usefulness as well as the potential shortcomings of the vignettes approach. Although using vignettes to correct self-reported health measures is thought to be a valuable correction method, the underlying assumptions of response consistency and vignette equivalence may not hold under certain circumstances. These potential problems with the vignettes approach may result in corrected self-reported health measures that are no better than uncorrected health measures in approaching the true underlying health status of respondents. In the worst case scenario the corrected self-reported health measures may even be inferior to the uncorrected self-reports.

As has been discussed in section 6, several attempts have already been undertaken to determine the quality of the vignettes method as a correction procedure. However, no firm conclusions on this can be drawn yet. The quality of the vignettes method in correcting self-reported health measures seems to depend for instance on the particular vignette used and the type of health domain considered (e.g., Voňková & Hullegie, 2010). To get a better idea of the quality of the correction method, we here conduct a different type of test then those already conducted in other studies. Instead of testing the specific underlying assumptions of vignette equivalence and response consistency we compare the explanatory power of the corrected self-reported health measures to that of the uncorrected ones in a simple regression framework. In this section we will first discuss the model that we estimate and the results of this estimation procedure will be discussed. Thereafter the details of the procedure adopted to test the appropriateness of the vignettes approach will be explained in detail. Besides, the results of this testing procedure will be illustrated and discussed.

8.1 Generating corrected health measures

To compare the quality of the corrected self-reported health measures to that of the uncorrected self-assessments by contrasting the explanatory power of the corrected and uncorrected health measures, we first need to obtain those corrected measures. The HOPIT model discussed in section 5.2 is used to correct self-reported health measures using anchoring vignettes. We estimated a modified version of the HOPIT model since we allow for unobserved heterogeneity across respondents and a potential correlation between the error terms of the model. Including unobserved heterogeneity in the HOPIT model has also been done by Kapteyn et al. (2007). However they incorporate the unobserved heterogeneity term in the cut-point equations, while we use two error terms in the latent health equations. We do not include objective health measures in our model, an extension that has been discussed in section 5.2, since we are only interested in testing the quality of the anchoring vignettes method and not in the effect of including objective health measures.

Briefly summarizing the discussion of the model from section 5.2, the vignette component of the HOPIT model yields us the individual-specific cut-points. These can then be used in the self-assessment component in which health is modelled as a function of individual characteristics. The two components can be estimated for all health domains and the work disability domain separately. After estimating the model one can obtain predicted levels of corrected latent health for each individual. We should not translate these into corrected self-reports measured on a scale equivalent

to that of the uncorrected self-reports using the estimated individual-specific thresholds, since this will cause part of the correction for reporting heterogeneity to be lost. Subjectivity of the self-assessment was due to the existence of differences in reporting behaviour and thus differences in the thresholds. When we use these varying thresholds to obtain corrected self-reports we plug some heterogeneity in reporting behaviour in these corrected health measures again, such that we would not obtain truly corrected health measures. It is then better to take a reference individual and impose his cut-points on the other individuals. These uniform cut-points can then be used to determine what a typical respondent would have reported were he to have these cut-points instead of his own cut-points. Moreover, we could calculate probabilities of reporting in a particular category using these corrected self-reports. For the present purpose we do not go into the details of creating such corrected self-reports and choosing reference individuals. We will only use the corrected latent health variable in our testing procedures. The fact that this health measure is not measured on a comparable, interpretable scale does not affect our testing procedures and results.

In this section we will first describe our version of the HOPIT model, which is a slightly adapted one. Thereafter we turn to the estimation of the model in section 8.1.2. When the model has been estimated, we run a simulation which is explained in detail in section 8.1.3. Finally, before turning to the testing procedure in section 8.2, we will give some descriptive statistics and explanations for the obtained corrected latent health variable in section 8.1.4.

8.1.1 A modified version of the HOPIT model

In the standard HOPIT model two components can be distinguished: the vignette component and the self-assessment component. The vignette component identifies the cut-points as functions of a number of individual characteristics, while the self-assessment component uses the estimated cut-point equations to obtain corrected latent health levels for each individual. In the vignette component the latent health level of the vignette person is modelled as an index function of a vignette-specific constant, to take into account the differences in the health levels of the various vignette persons.

We use a bivariate probit model to estimate cut-point equations and the latent health equation. Because of that, we are only able to include one vignette per health domain into our model. This is not a problem, since, as has already been indicated in section 5.2, theoretically one vignette is enough to do the correction procedure (King et al., 2004). We have chosen to include the first vignette for each health domain. We use the bivariate probit model because we allow for correlation between the error terms of the two latent health index functions. When we incorporate such a correlation, we should take into account that the vignette evaluation and the self-report are no long independent. Therefore we need to consider joint probabilities in our log-likelihood function. Mathematically, using the several model specifications that have been discussed in section 5.2, the model to be estimated can be described by Equation 16 to Equation 21[38]. The vignette component consists of Equations 16 to 18 and the self-assessment component is described by Equations 19 to 21.

$$Y_{vi}^{*} = v_i \;\; where \; v_i \sim N(0,1) \tag{16}$$

$$Y_{vi} = m \quad \Leftrightarrow \quad \tau_i^{k-1} \leq Y_{vi}^{*} < \tau_i^{k} \tag{17}$$

[38] The notation used in Equation 16 to Equation 21 differs slightly from that used in section 5.1 and section 5.2.

$$\tau_i^k = \gamma^k X_i \quad (18)$$

$$Y_{ri}^* = X_i'\hat{\beta} + u_i \; where \; u_i \sim N(0, \sigma^2) \quad (19)$$

$$Y_{ri} = k \quad \Leftrightarrow \quad \tau_i^{k-1} \leq Y_{ri}^* < \tau_i^k \quad (20)$$

$$\tau_i^k = \gamma^k X_i \quad (21)$$

where $m, k = 1, \ldots, 5$ and $\tau^0 = -\infty, \tau^5 = \infty$. Equations 16 to 18 do no longer contain the subscript j since we only use one vignette. Besides the function for the latent health level of the vignette person does no longer contain a constant term, but only an error term. This is done since we normalize the true latent health level of the vignette person to zero. Equation 18 and Equation 21 show the cut-point equations. The coefficients γ^k in these cut-point equations are the same in the vignette and self-assessment component. There are four cut-points that have to be estimated, since there are five categorical responses possible. The cut-points are a function of the individual characteristics X_i of a particular respondent. After estimating the cut-point equations predictions can be obtained that illustrate the threshold used by a particular respondent in evaluating the vignette person (and by assumption his own health). Predictions can be obtained when all individual characteristics are observed for the particular individual. Equation 19 defines the latent health index function as a function of the individual characteristics. This function describes the true health level of the respondent. In Equation 19 the error term u_i is assumed to be normally distributed with mean 0 and variance σ^2. After estimating the model we can predict the corrected latent health level, i.e., $\hat{Y}_{ri}^*$, of the respondent given his individual characteristics. Equation 20 describes the responses to the self-assessment question: depending on in between which thresholds the latent true health level of the respondent lies, he will give a specific response.

According to e.g., Van Soest et al. (2007) the error term in Equation 19 can be interpreted as unobserved heterogeneity in the self-reports. Regarding the error terms of the model, Bago d'Uva et al. (2008) and Kapteyn et al. (2007) explicitly assume that the error term in the latent health equation for the respondent and the error term in the health index function for the vignette person (i.e., u_i and v_i) are independent of each other. This independence is not explicitly assumed by Bago d'Uva et al. (2009). Van Soest et al. (2007), Kapteyn et al. (2009), Kapteyn et al. (2007) and Voňková & Hullegie (2010) incorporate unobserved heterogeneity affecting reporting behaviour of individuals in the threshold equations[39]. As Kapteyn et al. (2007) indicate this unobserved heterogeneity in the thresholds implies that the vignette evaluations are correlated with the self-assessments. Van Soest et al. (2007) and Voňková & Hullegie (2010) use an extension of the model that also includes objective measures. They assume that the error term of the objective part of the model is independent of the error term of the vignette component. At the same time they allow for correlation between the error term of the objective part and the error term in the health index function of the self-assessment part. More specifically, they assume a bivariate normal distribution of these two error terms. Correlation between these two error terms is allowed for since "*both will be affected by a common unobserved factor driving drinking behaviour*" (Van Soest et al., 2007, p.15).

In our model we relax the assumption that the error terms in the health index functions of the vignette component and the self-assessment component are independent. We assume that the error terms u_i and v_i may be correlated and that they follow a bivariate normal distribution where the correlation is captured by the correlation coefficient ρ. This is in the end comparable to incorporating heterogeneity in the threshold equations. So, by assuming correlation between the

[39] The way in which the thresholds are modelled when an unobserved heterogeneity term is included has already been illustrated in Equation 10 of section 5.2.

two error terms, we thus allow for unobservables affecting both the vignette evaluation as well as the self-assessment.

Estimation of the model requires the specification of a log-likelihood function. To specify the log-likelihood function that should be maximized we need the probabilities for the outcome variable to take on a specific value, as has been indicated in section 5.1 discussing the basics of ordered response models. We now have two observed outcome variables, namely the reported limitations for the vignette person and the self-reported limitations. Since we allow for correlation between the two outcome variables we need to look at all possible combinations of outcomes for the vignette evaluation and self-reported health. Thus, we need to consider 25 probabilities in our log-likelihood function. For instance, we need to consider the probability that the vignette evaluation equals 1 when the self-assessment also equals 1, the probability that the vignette evaluation equals 1 when the self-assessment equals 2, etcetera. The probability of a particular combination of the vignette evaluation and the self-assessment appearing is indicated by p_{ikm}, where i indicates that the probability is individual-specific, k indicates the value of the self-report ($k = 1, 2, \dots, 5$) and m indicates the value of the vignette evaluation ($m = 1, 2, \dots, 5$). Depending on the combination of outcomes, the probability of this combination appearing may be the sum of a couple of probabilities resulting from the (bivariate) normal cumulative distribution. Using these 25 probabilities the log-likelihood function to be maximized can be specified as in Equation 22.

$$\log\big(L(\beta, \gamma^1, \gamma^2, \gamma^3, \gamma^4, \sigma^2, \rho)\big) = \sum_{i=1}^{n} \sum_{k=1}^{5} \sum_{m=1}^{5} Y_{ri}^{k} \cdot Y_{vi}^{m} \cdot \log(p_{ikm}) \qquad (22)$$

where Y_{ri}^{k} equals 1 if $Y_{ri} = k$ and 0 if $Y_{ri} \neq k$ and similarly Y_{vi}^{m} equals 1 if $Y_{vi} = m$ and 0 if $Y_{vi} \neq m$. Maximum likelihood estimation will then result in estimated coefficients for the latent health index function of the respondent, i.e., $\hat{\beta}$'s, and estimated coefficients for the four cut-point equations, i.e., $\widehat{\gamma^k},\ k = 1, 2, 3, 4$. In addition we obtain estimates for the standard deviation σ and the correlation coefficient ρ. The next section goes into the details of estimating this bivariate probit model.

8.1.2 Estimating the modified version of the HOPIT model

Estimation of the model described in section 8.1.1 can be done per health domain; we will consider the health domains pain, mobility, sleep, breath and depress[40]. In addition, we will look at the work disability domain. Important to note is that we estimate the model using data from the first wave only. This is done because we want to obtain corrected self-reports in wave 1 in order to explain a particular outcome measure in wave 2. Therefore we only need to have corrected self-reports in the first wave instead of for both waves. We use the longitudinal subsample of the SHARE dataset consisting of 2,997 respondents in both waves, the construction of which has been discussed in section 7.2.1. A couple of observations drop out when estimating the model since for some respondents the individual characteristics contain missing values. Estimates are obtained using observations for 2,980 respondents in the first wave.

Both the cut-points as well as the latent health index function for a respondent are functions of a vector of individual characteristics X_i'. Numerous individual characteristics can be included in the model. Previous studies have included various individual characteristics. For instance Kapteyn et al.

[40] We also tried to estimate the model for the memory domain. This resulted in problems with obtaining numerical derivatives. We used several optimization methods, but none of them resulted in estimates for the model. Therefore we will not consider the memory domain in our testing procedure.

(2009) include age dummies, years of education, a dummy for being female, country dummies and a dummy for being married. Besides they incorporate a number of variables indicating the health status of a respondent. Bago d'Uva et al. (2009) include age dummies, a dummy for ethnicity, the logarithm of wealth, the education level and a dummy indicating whether individuals are younger than 65 and not working. Moreover, Voňková & Hullegie (2010) include country dummies, dummies for age groups, gender, education level dummies, a dummy for living alone, a dummy for suffering from a long-term illness and an indicator for the amount of physical activity the respondent is engaged in. Van Soest et al. (2007) include age dummies, a dummy for being female, a dummy for being married and a dummy for going out, a dummy for ethnicity and a couple of measures for the education level and alcohol consumption of the parents of the respondent since they specifically focus on drinking behaviour. Finally, Bago d'Uva et al. (2008) include a female dummy, age group dummies, education level dummies, the logarithm of income and a dummy for living in an urban area as socio-demographic variables.

We considered several specifications of the model by including different variables as individual characteristics in a stepwise procedure. We started with a basic specification of the model estimating all equations as a function of a constant only. Thereafter we introduced more and more individual characteristics and estimated the cut-point equations and the latent health equation as functions of these individual characteristics. We considered the Wald test statistic to evaluate the joint significance of a particular individual characteristic in all cut-point equations[41]. In addition, we took into account the significance of the individual characteristics in the latent health equation by evaluating the Z-statistics for the various coefficients. Our final model specification includes the following individual characteristics: six dummies for several five-year age groups as specified in section 7.2.2 with those individuals aged 50 to 55 being the reference group, a dummy for being female, dummies for low and middle levels of education (compared to a high level of education) and seven country dummies with Germany being the reference group[42].

We simultaneously estimate the two steps of the HOPIT model, i.e., the vignette and the self-assessment component, using maximum likelihood estimation and the log-likelihood function specified in Equation 22. Estimating this model yields us the coefficients for the latent health index function, i.e., the $\hat{\beta}$'s, the coefficients for the four cut-point equations, i.e., the $\widehat{\gamma^k}$'s for $k = 1, 2, 3, 4$. In addition we get estimates for the standard deviation of the error term u_i and the correlation between the two error terms, i.e., u_i and v_i. As soon as the estimated coefficients are known we can obtain predicted values for the latent health level of the respondent and the individual-specific cut-points. The estimates and predicted values will be used in the simulation exercise that will be discussed in section 8.1.3.

The estimation results for each health domain can be found in Table A2 - 1 to Table A2 - 6 in Appendix 2. Coefficient estimates, standard errors and the Z-values are provided for the health index

[41] We considered the null hypothesis that the coefficients of a specific individual characteristic equal zero in all cut-point equations, i.e., we tested the null hypothesis $H_0: \gamma_1^1 = 0 \cap \gamma_1^2 = 0 \cap \gamma_1^3 = 0 \cap \gamma_1^4 = 0$ versus the alternative hypothesis of $H_1: \gamma_1^1 = 0 \cup \gamma_1^2 = 0 \cup \gamma_1^3 = 0 \cup \gamma_1^4 = 0$. Such a testing procedure is also used by Jones et al. (2007).

[42] While looking for the optimal specification of the HOPIT model we also estimated specifications including the logarithm of income, a dummy for being married, a dummy for ethnicity and dummies for the size of the respondent's household. However, these individual characteristics appeared to be far from significant in the cut-point equations according to the Wald test statistic such that we excluded them in the end. In some instances some of the age dummies and/or one of the education dummies were insignificant in the four cut-point equations; we have not excluded these however.

function, the cut-point functions, sigma and the correlation ρ. In addition the log-likelihood and the number of observations used in the estimation procedure are shown.

In a probit model the magnitude of the coefficients is difficult to interpret directly. However, the sign and the significance of the coefficients can be interpreted (Stock & Watson, 2007). The estimated coefficients for the pain domain presented in Table A2 - 1 show us that compared to the Germans the respondents in the other countries have lower levels of corrected latent health on average, as indicated by the negative coefficients on the country dummies. This means that on average German respondents have more limitations than respondents in other countries, given the other individual characteristics. The country dummies are significant at the 1% significance level, except for Greece and Belgium. Looking at the coefficients on the age dummies in the latent health equation, with respondents aged 50 to 55 being the reference group, one can see that only for the upper four age groups the corrected latent health level differs significantly. The coefficients on these upper four age groups are positive and highly significant at the 1% significance level. So, compared to individuals aged 50 to 55, these older individuals on average have a higher corrected latent health level, i.e., more limitations. This is to be expected since one would anticipate more health problems for older individuals. The coefficient on the female dummy is positive and highly significant meaning that females on average have a higher corrected latent health level and thus more health limitations in the pain domain. In addition, the dummies for low and middle levels of education, with a high level of education being the reference group, are positive, although the middle education dummy is insignificant. This implies that holding other individual characteristics equal, those individuals with high levels of education have a lower corrected latent health level on average and thus fewer health limitations in the pain domain.

Moreover, we can evaluate the estimation results for the cut-point equations. In the first cut-point equation, only the coefficient on the age dummy for individuals aged 55 to 60 is significantly different from zero; coefficients on all other age dummies are insignificant at the 5% level. This coefficient is positive, so individuals aged 55 to 60 on average have a higher value of the first cut-point than individuals aged 50 to 55. This implies that those aged 55 to 60 need a higher level of latent health, i.e., more limitations, in order to report mild instead of no limitations in the pain domain. The coefficients on female and a middle level of education are insignificant. On the other hand, the coefficient on low education is borderline significant at the 1% level. This coefficient is positive, meaning that compared to individuals with a high level of education, low educated respondents have a higher value of the first cut-point. So, the latter need to have more actual limitations than the high educated individuals in order to report mild instead of no limitations, i.e., concerning the first two response categories low educated respondents are more optimistic about their health status. The coefficients on all country dummies are significant, most of them at the 1% significance level. Except for Greece all coefficients are negative. This implies that compared to Germany respondents in all countries, except for Greece, on average have a lower level of the first cut-point. Thus, respondents in these countries need to have fewer actual limitations than German respondents so as to report mild instead of no limitations. For Greek respondents the reverse holds.

The estimates of the upper three cut-point equations show that both the significance as well as the sign of the coefficients differs across cut-points. In the second cut-point equation none of the age dummies is significant. For the third cut-point only the dummy for individuals aged 65 to 70 is significant and positive, meaning that these individuals use a higher value of the third cut-point than respondents aged 50 to 55. In the fourth cut-point equation only the dummy for individuals aged 75

to 80 is significant and positive. The female dummy is significant only for cut-points three and four. In both of these cut-point equations the coefficient is positive, which implies that females use a higher value of these cut-points and thus seem to be less inclined to report extreme or severe limitations. The dummy for a low education level is only significant in the second cut-point equation, while the dummy for a middle education level is highly insignificant in all cut-point equations. The coefficient on low education is negative in the second cut-point equation, while it was positive in the first one. This implies that compared to a respondent with a high level of education, a low educated respondent uses a lower value of the second cut-point on average, i.e., the latter will report moderate limitations earlier. Besides, since the low educated respondents used a higher value of the first cut-point, there is less mass in between the first and the second cut-point, meaning that only for a rather small range of latent health levels low educated individuals will report to have mild limitations. Finally, we can consider the estimated coefficients for the country dummies in the equations for the upper three cut-points. It is interesting to see that there are differences in the signs of the coefficients on the country dummies across the cut-point equations. For instance, in the second cut-point equation the estimated coefficients are significant except for France. For those countries for which the coefficient is significant, a negative sign is observed for Sweden, Spain and Italy, meaning that respondents in these countries use a lower level for the second cut-point compared to the Germans. In addition, the estimated coefficients in the third cut-point equation show that respondents in those countries also have significantly lower levels of the third cut-point compared to German individuals. In the second cut-point equation positive signs are observed for the Netherlands, Greece and Belgium. This points at higher second cut-point values being used in those countries compared to Germany. In the third cut-point equation the estimated coefficients for these three countries are insignificant. Finally, evaluating the estimates of the fourth cut-point equation one can see that significantly lower values of the fourth cut-point are used in Sweden, the Netherlands and Italy. So in these countries, respondents need fewer true limitations in order to report extreme limitations compared to German respondents. Coefficients are insignificant for Spain, France, Greece and Belgium.

A similar analysis of the sign and significance of the coefficients in the cut-point equations and the latent health equation can be done for the other health domains. When we look at the mobility domain we can see that the dummy for a middle level of education is significant in the latent health equation. In addition, the sign of the coefficient on this dummy variable is positive, just as the sign of the dummy for low education, meaning that compared to individuals with a high level of education those individuals with a middle level of education have a higher level of corrected latent health on average and thus more limitations. Looking at the first cut-point equation we can see that only the dummy for individuals aged 80 to 85 is significant and positive. So, compared to individuals aged 50 to 55 those aged 80 to 85 have a higher corrected latent health level on average and thus more limitations. The coefficient on the female dummy is also significant in the first cut-point equation and its sign is negative. This implies that on average females have a lower value of the first cut-point than males. The coefficients on both education dummies are highly significant in the first cut-point equation and both have positive signs. This means that individuals with middle or low levels of on average have a higher value of the first cut-point than high educated respondents. So, the former will report no limitations for a broader range of corrected latent health levels. Most of the country dummies are insignificant in the first cut-point equation, except for Sweden and Italy. The coefficient on the dummy for Sweden is negative and that for Italy is positive. This points at lower and higher

values of the first cut-point in Sweden and Italy respectively, compared to Germany. For the second cut-point equation all age dummies are insignificant. The two education dummies are significant and positive, meaning that those with a low or middle level of education use a higher value of the second cut-point than those with a high level of education. As regards the country dummies again only the coefficients on Sweden and Italy are significant; their signs are the same as in the first cut-point equation. In the third cut-point equation only the age dummy for individuals aged 75 to 80 is significant and positive, meaning that those individuals use a higher third cut-point value on average than those aged 50 to 55. The gender and education dummies are insignificant in the third cut-point equation. For the country dummies the same applies as in the first and second cut-point equation. The story becomes a bit different for the uppermost cut-point equation. Here the dummies for the four oldest age groups are significant and the signs of the coefficients are positive. This implies that compared to individuals aged 50 to 55, the older respondents use higher values for the fourth cut-point and will thus report extreme limitations for a smaller range of corrected latent health levels. Only the dummy for low education is significant and negative, meaning that individuals with a low level of education use a lower value of the fourth cut-point than those with a high education level. Finally, most of the country dummies are significant, except for Spain, Italy and Belgium. The coefficients for Sweden, the Netherlands and Greece are negative, such that respondents in these countries on average use a lower value of the fourth cut-point than German respondents do. The coefficient on France is positive, meaning that the French use a higher fourth cut-point value than the Germans; so they report extreme limitations for a smaller range of corrected latent health levels.

In the sleep domain only the dummies for individuals aged 75 to 80 and those aged 80 to 85 are significant in the latent health equation. Their signs are positive, meaning that these older individuals have higher values of corrected latent health on average and thus more limitations. The dummies for middle and low levels of education are both significant and positive. This implies that higher educated individuals have lower levels of corrected latent health and thus fewer limitations on average. The dummies for Sweden, the Netherlands, Spain and Greece are significant and negative. Thus, individuals in those countries on average have lower levels of latent health and thus fewer limitations than the Germans. Regarding the cut-point equations we can see that individuals aged 80 to 85 use a significantly higher value of the first cut-point than those aged 50 to 55; so these older individuals report no limitations for a larger range of corrected latent health levels. In addition, these older individuals also use a significantly higher value of the third and fourth cut-point. Most of the other age dummies are also significant and positive in the third cut-point equation. Furthermore, females use a significantly lower value of the first cut-point than males; the female dummy is however insignificant in the other cut-point equations. Those individuals with a low level of education use a higher value of all cut-points compared to higher educated individuals. Respondents with a middle level of education only have a significantly higher level of the second and third cut-point. Regarding the country dummies lower values for the cut-points are used in all countries compared to Germany, but the coefficient on Italy is insignificant. In the second cut-point equation only the coefficients on Sweden, the Netherlands, Spain and France are significant at the 5% level. Coefficients are negative for these countries meaning that individuals also use a lower value of the second cut-point compared to Germany. In the third cut-point equation only Swedish and French respondents use a significantly different value of the cut-point compared to Germans; Swedish respondents use a lower value while the French use a higher value of the third cut-point. Finally, the coefficients on Sweden, the Netherlands, France and Greece are significant in the fourth cut-point. Of these four country dummies only the coefficient on France is positive, meaning that respondents

in France use a higher value of the fourth cut-point than the Germans. On the other hand, individuals in the other three countries use a significantly lower value of this cut-point than German respondents do.

For the breath domain only the upper four age group dummies are significant and positive. This implies that these older individuals have more limitations on average than individuals aged 50 to 55. In addition, the female dummy is significant and positive. So females have higher levels of limitations than males. Regarding the country dummies, we can see that the coefficients on Sweden, France and Belgium are positive and significant. The coefficient on Spain is significant but negative. So compared to individuals in Germany respondent in Sweden, France and Belgium have higher levels of latent health, so more limitations, while individuals in Spain have fewer limitations. With respect to the cut-point equations we observe significantly lower first cut-point values for individuals aged 70 to 75 and individuals aged 80 to 85, compared to individuals aged 50 to 55. In addition, on average a higher first cut-point value is observed for females compared to males. The education dummies are insignificant in the first cut-point equation. Furthermore, most of the country dummies are significant, except for the Netherlands. The signs are mostly negative, except for Italy, France and Belgium. This implies that in these three countries higher values for the first cut-point are used than in Germany, while individuals in the other countries use lower values for the first cut-point on average. In the second cut-point equation all age dummies are insignificant. Females have a significantly higher level of the second cut-point than males. Regarding the country dummies, the coefficients on the Netherlands, Italy, France, Greece and Belgium are positive and significant. On the other hand, Spanish individuals use a significantly lower level of the second cut-point compared to German respondents. From the third cut-point equation we can see that individuals aged 75 to 80 use a significantly higher value of the third cut-point compared to individuals aged 50 to 55. Besides, respondents with a middle level of education use a significantly higher value of the third cut-point compared to the high educated. Also females use a higher cut-point value than males. Finally, the coefficients on the country dummies are only significant for Spain, France and Belgium. Spanish respondents use a significantly lower value of the third cut-point, while French and Belgian respondents use a significantly higher value of that same cut-point compared to German respondents. With respect to the fourth cut-point we can see that individuals aged 60 to 65 and individuals aged 65 to 70 use a significantly higher value of the fourth cut-point than individuals aged 50 to 55. So these older individuals are less inclined to report extreme limitations. For the rest only the coefficients on the Netherlands and France are significantly different from zero. The Dutch use a lower value of the fourth cut-point, while the French use a higher value of the fourth cut-point compared to German respondents.

When we look at the work disability domain we can see that individuals aged 55 to 60 use a significantly lower level of the first and second cut-point than individuals aged 50 to 55; so the former group will report work limitations earlier. Individuals aged 70 to 75 use a lower value of the first cut-point. Individuals with a low level of education use higher values of the first three cut-points than high educated individuals. All country dummies except for Belgium are insignificant in the first cut-point equation. Respondents in Belgium use a lower value of the first cut-point compared to German respondents. In the second cut-point equation the coefficients on Sweden and Spain are significant and negative, meaning that individuals in these countries use a lower level of the second cut-point than individuals in Germany. Regarding the third cut-point equation we can see that individuals aged 60 to 65 and individuals aged 65 to 70 use a significantly higher value of the third cut-point than those individuals aged 50 to 55. The coefficients on the country dummies are

significant except for France and Italy; for all countries the coefficients are negative, meaning that compared to Germany individuals in these countries use lower values of the third cut-point. Finally, we can see that all age dummies are significant and positive in the fourth cut-point equation. So, all age groups use higher levels of the uppermost cut-point on average. From this one can see that the individuals aged 50 to 55 will report extreme work limitations earlier. Furthermore, the country dummies are significant except for France. Besides the coefficients are negative, meaning that on average the Germans use the highest values of the fourth cut-point.

Finally, the estimates for the depress domain show that individuals with low and middle levels of education have higher levels of latent health and thus more limitations on average than individuals with a high level of education. Also most country dummies are significant and negative, except for Italy, pointing at lower levels of corrected latent health in these countries compared to Germany. Regarding the cut-point equations we can observe that individuals aged 75 to 80 use a significantly lower value of the first cut-point while individuals aged 80 to 85 use a significantly higher value of the first cut-point compared to those aged 50 to 55. In addition, females use a lower first cut-point value than males and individuals with a low education use a higher value of the first cut-point than high educated individuals. The country dummies are significant in the first and second cut-point equation, except for Italy. The coefficients on the country dummies are negative, so respondents in Germany on average use the highest value of the first and second cut-point. Regarding the second and third cut-point one can see that only the dummy for individuals aged 80 to 85 is significant and positive, meaning that those individuals use a higher value of the second cut-point than individuals aged 50 to 55. Furthermore, we can see that females use a significantly lower level of the second cut-point than males and besides low educated respondents use a significantly higher level of the second cut-point compared to high educated respondents. For the third cut-point individuals with low and middle levels of education use a significantly higher level of the cut-point than high educated individuals. For the country dummies still all coefficients are negative, while Italy and France are not significant in the third cut-point equation. Finally, all age dummies except for individuals aged 55 to 60 are significant and positive. This implies that on average respondents aged 50 to 55 use the lowest level of the fourth cut-point in the depress domain. Thus, these individuals report extreme limitations regarding depression for a broader range of corrected latent health levels. In addition, females use a higher value of the fourth cut-point than males. The country dummies are significant except for Sweden and Belgium. For those countries of which the coefficient is significantly different from zero, most signs are negative, except for France. So only the French use a significantly higher value of the fourth cut-point than the German respondents.

8.1.3 A simulation exercise

After estimating the bivariate probit model we run a simulation. The vignette evaluation is indicative for the unobserved heterogeneity term that we included in our bivariate probit model. We want to obtain the health level of an individual conditional on his vignette evaluation. Therefore, we want to take into account the unobserved term, which is conditional on the vignette answer, in the latent health level. However, then no closed form solution is available such that we have to use a simulation exercise. During this simulation we draw random numbers from a bivariate normal distribution for the two error terms of the model for each respondent. We do this 1,000 times[43]. The

[43] We have also investigated the results from performing the simulation 2,000 and 10,000 times. The former did not affect the number of observations available for the predicted corrected latent health variable. The latter only resulted in 2 more

exact procedure followed in this simulation exercise is described in detail in Appendix 3. We use a bivariate normal distribution with mean zero and the covariance matrix in Equation 23.

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma \cdot \rho \\ \sigma \cdot \rho & 1 \end{bmatrix} \quad (23)$$

By assumption the variance of the error term in the latent health index function of the respondent, i.e., the variance of u, equals σ^2; the variance of the error term of the latent health index function of the vignette person, i.e., the variance of v, equals 1. In addition, the covariance between the two error terms[44] equals the standard deviation of u multiplied by the correlation coefficient ρ. Using the draws for the vignette error term v we can determine whether the respondent has evaluated the vignette person correctly. This can be done by comparing the draw for v to the predicted values of the cut-points. The vignette person is evaluated correctly if for instance the respondent reported no limitations for the vignette person and the draw for the error term v is smaller than the predicted value of the first cut-point. If the vignette evaluation variable equals 2, the draw for the error term v must be smaller than the predicted value of the second cut-point, but larger than or equal to the predicted value of the first cut-point. Something similar applies when the vignette person is evaluated as having moderate, severe or extreme limitations. If the draw for the error term v satisfies this, the true latent health level of the vignette person was such that, according to the individual-specific cut-point values, the respondent should have evaluated the vignette person as he did in reality. We generate a new variable that equals 1 if the vignette person is evaluated correctly and zero otherwise.

Moreover, we add the draw for the error term of the own latent health, i.e., u_i, to the predicted value of corrected latent health. We generate a new variable that equals the sum of the predicted corrected latent health level, resulting from the estimated latent health equation for the respondent, and the draw for the error term of the own latent health equation[45]. To obtain a corrected latent health variable that we could use in our testing procedure we only want to consider those draws for the error terms for which the observed vignette evaluation is possible. Thus, for all 1,000 simulations that we perform we determine whether the vignette person is evaluated correctly and count the number of times this is the case. For those draws for which the vignette person is evaluated correctly we add the h's, i.e., the latent health levels plus the draws for the error terms, and divide this by the number of times the vignette person has been evaluated correctly. This then results in an average predicted corrected latent health level of the respondent taking into account the random draws for the error terms. This variable is labelled $hpred$; this is the latent health variable corrected for reporting heterogeneity that we can use in our testing procedure. This corrected health measure will be discussed in more detail in the next section.

8.1.4 Corrected latent health

We have now obtained corrected latent health levels, captured by the variable $hpred$, for the various domains for which the model has been estimated. We have obtained such corrected latent health levels for most of the respondents in the longitudinal sample, although some missing values

observations for the predicted corrected latent health level. We therefore decided to use 1,000 iterations for the simulation procedure.

[44] The correlation between two random variables x and y is defined as $\rho_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}$ (Heij et al., 2004). This can then be rewritten to obtain the covariance between x and y: $cov(X,Y) = \rho_{xy} \cdot \sigma_x \cdot \sigma_y$.

[45] This variable is labelled h and is defined as $h_i = x_i' \hat{\beta} + u_i$, where u_i is the drawn random variable.

appear. The number of individuals for which corrected latent health is known differs somewhat across health domains; the exact figures are provided in Table 5. Since the corrected latent health variable is corrected for reporting heterogeneity, it should approximate the true health level of the respondent better than the self-reported health measure does if the vignettes method is a valid correction method.

To get more insight into the corrected latent health variable we provide descriptive statistics and graphs for this variable in Table 5 and Figure 6 respectively. The table shows the average, standard deviation and range of the corrected latent health variable per health domain and per country. The corrected latent health level is measured on another scale than the self-report. The former is measured on a continuous scale ranging from $-\infty$ to ∞, whereas the latter is measured on a discrete scale ranging from 1 to 5. Although we could therefore not compare these two health measures directly, they can be used in our estimation procedure as corrected and uncorrected health measures respectively. The scale of the latent health variable cannot be interpreted directly but means, standard deviations and ranges can be compared across countries. Evaluating the pain domain, one can see for instance that the differences in mean latent health level are rather small between Greece and Germany: the average latent health level in Germany is 0.24 and in Greece it is 0.13. The range of the latent health level differs somewhat more for these two countries, which is reflected in the standard deviation. Larger differences can be observed between Germany and Sweden. The average latent health level in Sweden is much lower than that in Germany, i.e., -1.91 compared to 0.24. From this one would conclude that the Swedish respondents have fewer limitations in the pain domain than the German respondents. There are only very small differences between Spain and the Netherlands; for these two countries there is an enormous overlap in the ranges and in addition the averages lie very close to each other, i.e. the mean is -0.46 in the Netherlands and -0.43 in Spain. The relative position of countries seems to differ slightly across health domains. While the Germans seem to have the most limitations in the pain domain, they score relatively low in the sleep domain and breathing domain. In the latter two domains Italy, France and Belgium have a higher average corrected latent health level and thus more limitations. However, in the domains of mobility, depression and work disability, German respondents have quite a lot of limitations; only Italian respondents have a higher average latent health level. Dutch respondents have quite low values of latent health in all domains and even the smallest average in the depression domain. There are however considerable differences in the relative position of Spain: the average latent health level in Spain is very low compared to other countries in the breath domain, but quite high in the mobility and work disability domains[46].

The histograms in Figure 6 show the distribution of corrected latent health levels across individuals per health domain. The magnitude of the latent health variable cannot be interpreted directly, however relative comparisons are possible. A higher latent health level implies that there are more limitations for a respondent, i.e., that the respondent is in worse health. Comparing the distributions of latent health levels across health domains shows that for one health domain there is somewhat more concentration than for another. For instance, in the pain domain the range is from around -4.3 to 2.4 while in the depress domain latent health levels only range in between -3.9 and -0.9. The fraction of respondents in a particular class of latent health levels also differs across health domains. In the pain domain, and to a lesser extent in the sleep domain, only a small number of classes account for a large fraction of the respondents. For instance, in the pain domain around 55%

[46] In the mobility and work disability domains higher averages are only observed in Germany and Italy.

of the respondents is within the tree highest bars of the histogram. On the other hand, we see a larger number of bars with similar height in the figures for the mobility and breath domains. Also interesting to see in the sleep domain is that there is a small concentration of respondents for latent health levels in between -6 and -4 and a larger concentration of respondents for health levels in between -3 and -1. So, overall one can observe some differences in the distributions of corrected latent health levels across health domains.

Table 5: Descriptive statistics for predicted corrected latent health per health domain and per country (wave 1).

	Pain domain				
	# obs.	mean	st.dev	min	max
All	2,938	-0.2570	0.7530	-4.3360	2.4311
DE	265	0.2434	0.4154	-0.6676	1.3039
SE	278	-1.9096	0.4566	-3.4716	-0.6624
NL	349	-0.4585	0.4084	-1.4301	0.5642
ES	238	-0.4281	0.4958	-1.4631	0.9465
IT	313	-0.3603	0.4561	-1.5256	0.8106
FR	582	-0.0593	0.4958	-1.1229	2.4311
GR	484	0.1348	0.5299	-4.3360	1.7366
BE	429	0.1289	0.4617	-0.9687	1.7368

	Mobility domain				
	# obs.	mean	st.dev	min	max
All	2,933	-2.8931	0.7730	-4.9672	-0.4426
DE	265	-2.4738	0.5108	-3.5702	-1.1427
SE	279	-3.2735	0.6447	-4.4295	-1.5826
NL	347	-3.0393	0.5964	-4.0940	-0.9887
ES	237	-2.5042	0.6412	-3.9093	-1.0769
IT	312	-2.1201	0.5843	-3.4755	-0.4426
FR	574	-2.9791	0.7011	-4.3269	-1.3121
GR	485	-3.5708	0.6907	-4.9672	-1.9921
BE	434	-2.6848	0.6195	-3.8200	-1.0116

	Sleep domain				
	# obs.	mean	st.dev	min	max
All	2,937	-2.2813	0.9713	-5.9585	-0.2626
DE	265	-2.1713	0.3819	-4.5465	-1.0938
SE	278	-4.6905	0.4611	-5.9585	-3.5423
NL	346	-2.3896	0.4016	-3.3275	-1.3515
ES	238	-2.3197	0.4929	-3.5757	-1.0100
IT	315	-1.5015	0.4568	-2.5642	-0.2626
FR	583	-1.7367	0.5204	-3.7724	-0.5409
GR	484	-2.5072	0.5351	-3.6944	-1.2194
BE	427	-1.8734	0.4724	-4.6685	-0.4500

	Breath domain				
	# obs.	mean	st.dev	min	max
All	2,933	-2.1074	0.8544	-4.0430	0.0007
DE	268	-2.4831	0.2531	-3.1227	-1.7943
SE	279	-1.7017	0.2858	-2.3493	-0.7210
NL	349	-2.8252	0.2812	-3.4496	-2.0162
ES	237	-3.2788	0.3206	-4.0430	-2.5042
IT	309	-2.4310	0.2937	-3.2162	-1.5543
FR	577	-0.8713	0.3096	-1.4964	0.0007
GR	484	-2.8112	0.3029	-3.3988	-1.7259
BE	430	-1.5419	0.3150	-2.5984	-0.0874

	Depress domain				
	# obs.	mean	st.dev	min	max
All	2,933	-2.5249	0.5178	-3.9164	-0.8865
DE	265	-2.1713	0.3819	-4.5465	-1.0938
SE	280	-2.5565	0.4095	-3.4285	-1.3875
NL	346	-3.1888	0.3372	-3.9164	-1.9598
ES	237	-2.7083	0.4045	-3.7569	-1.8375
IT	309	-1.8581	0.3873	-2.6535	-0.8865
FR	576	-2.6256	0.4331	-3.4133	-1.4326
GR	485	-2.3278	0.4383	-3.2425	-1.2512
BE	433	-2.5745	0.3740	-3.2848	-1.3323

	Work disability domain				
	# obs.	mean	st.dev	min	max
All	2,924	-2.5308	0.6687	-4.5638	-0.6320
DE	265	-2.1713	0.3819	-4.5465	-1.0938
SE	279	-2.5319	0.5620	-3.9400	-0.9167
NL	342	-2.9083	0.4718	-3.9742	-1.1332
ES	238	-2.1952	0.5965	-3.5518	-0.7589
IT	310	-2.0098	0.5462	-3.1443	-0.6320
FR	578	-2.4183	0.6288	-4.0183	-0.7650
GR	483	-3.1446	0.6320	-4.5638	-1.3985
BE	429	-2.4742	0.5358	-3.5997	-0.7173

Note: the following country abbreviations are used: All countries (All), Germany (DE), Sweden (SE), the Netherlands (NL), Spain (ES), Italy (IT), France (FR), Greece (GR) and Belgium (BE).

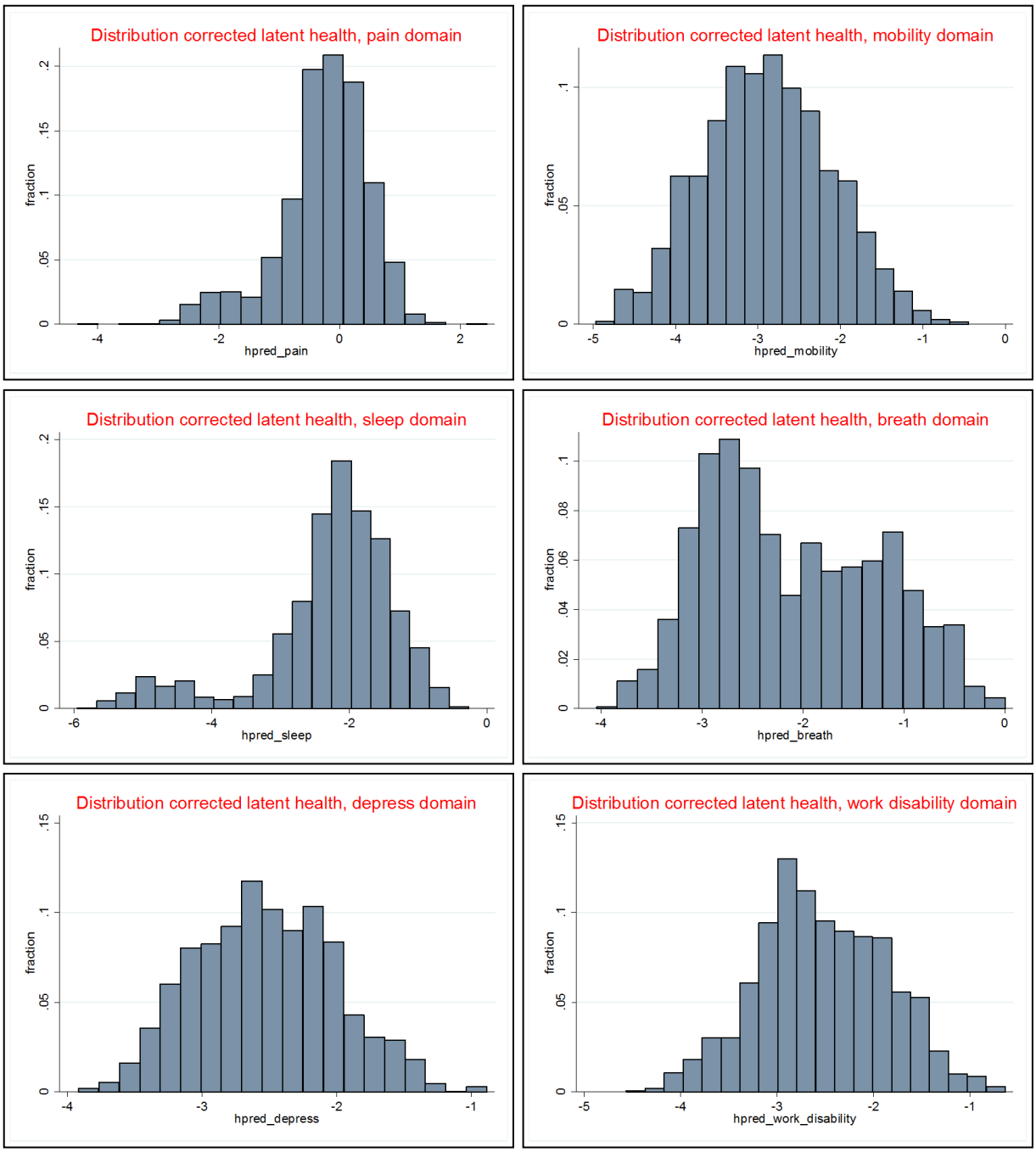

Figure 6: Distributions of predicted corrected latent health per health domain (wave 1).

To evaluate the relationship between the corrected and uncorrected health measures, we look at the correlation between the two for each health domain. These correlations indicate the extent to which corrected and uncorrected health move together. A perfect negative correlation, i.e., $\rho = -1$ would imply that a high corrected latent health level is associated with a low value for the uncorrected self-report. In case of a perfect positive correlation a high corrected latent health level is associated with a high value for the self-report. Finally, when the correlation equals zero, no clear co-movement of the corrected and uncorrected health levels is observable. If the correction with anchoring vignettes does not have a large influence, we would expect an almost perfect correlation, since in that case the uncorrected and corrected health measures would move very closely together. On the other hand, when health measures are significantly altered because of the correction procedure, we may observe

a correlation close to zero, since health levels may then shift considerably as a result of the correction. Low correlation coefficients do not necessarily validate the vignettes method as a correction tool, since the fact that health measures are altered by the correction does not bring them closer to the true health level per se, which is what we would like the vignettes method to do in the end. The correlation coefficients for the raw self-reported health measures and the corrected latent health levels are provided per health domain and per country in Table 6. The significance of the correlation coefficients is also indicated. From the table one can see that overall the correlation between corrected and uncorrected health is significantly different from zero for each health domain, even at the 1% significance level. In addition, all correlations are positive meaning that a higher value of the corrected health measure is associated with a higher value of the uncorrected health measure. Overall, the correlation coefficients are quite small, i.e., closer to zero than to one. Moreover, some differences in the magnitude of the correlations can be observed both across health domains and across countries. The overall correlations range from 0.1529 in the breath domain to 0.2846 in the work disability domain. This may indicate that the correction with anchoring vignettes has a larger effect on health in one domain than in the other. However, based on these correlation coefficients we cannot draw any conclusions yet on the quality of the vignettes method and potential differences therein across health domains. We can only observe that the corrected and uncorrected health measures move closer together in one health domain than in another.

The differences in correlations are somewhat larger when we look across countries. In the pain domain, the lowest correlation is observed for Germany, i.e., 0.1198, while the highest correlation of 0.3231 is observed for Spain. In the sleep domain the correlation is insignificant in Sweden; this is also the case in the breath domain. In the latter domain insignificant correlation is also observed for Spain. In the depress domain the correlation coefficient is insignificant for Germany and the Netherlands. Correlations are less apparent in the breath domain: some of the correlations are insignificant and others are only significant at the 10% significance level. Only for three countries we observe correlations that are significant at the 1% level in the breath domain. Comparing the relative positions of countries according to the magnitude of the correlations illustrates that in general low correlations appear in Germany, Sweden, Belgium and the Netherlands although some exceptions exist. Higher correlations are observed in the Mediterranean countries of Italy, Spain, Greece and France. This may indicate that the vignettes method does less in those countries compared to other countries, although we should not interpret this as differences in the quality of the correction procedure across countries.

Table 6: Correlation coefficients (wave 1).

	Corrected versus uncorrected health – correlation coefficients					
	Pain	*Mobility*	*Sleep*	*Breath*	*Depress*	*Work disability*
All countries	0.2473***	0.2787***	0.2584***	0.1529***	0.1744***	0.2846***
Germany	0.1198*	0.1829***	0.1593***	0.1181*	0.0132	0.2236***
Sweden	0.1520**	0.1258**	0.0309	0.0665	0.1723***	0.1658***
Netherlands	0.2146***	0.1094**	0.1740***	0.2509***	0.0668	0.1908***
Spain	0.3231***	0.3073***	0.2713***	0.0748	0.2206***	0.2225***
Italy	0.2846***	0.3447***	0.2430***	0.2612***	0.1644***	0.3153***
France	0.2870***	0.3353***	0.1650***	0.1517***	0.1389***	0.3268***
Greece	0.1679***	0.3282***	0.2612***	0.0858*	0.1199***	0.2481***
Belgium	0.1924***	0.2472***	0.1433***	0.1095**	0.1267***	0.1645***

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

In addition, we could look at the correlation between the two error terms of the model, which have been estimated together with the cut-point equations and the latent health equation. The correlation coefficients between the error terms u_i and v_i together with the associated standard deviations and Z-statistics are provided per health domain in Table 7. If the correlation between the two error terms is perfect, this implies that the self-report and vignette evaluation of an individual are also affected by unobservable characteristics. On the other hand, when the correlation coefficient is not significantly different from zero, then the report of an individual is determined by observable characteristics only. So a correlation coefficient that is significantly different from zero indicates that unobserved heterogeneity of individuals plays a role in reporting behaviour. Thus, in that case there is correlation between the evaluation of the vignette person and the self-report. From Table 7 one can see that significant correlations between the error terms only exist for a selection of the health domains considered. Only for the health domains pain, sleep and depress are the correlation coefficients significantly different from zero even at the 1% significance level. For the other health domains however, the correlation coefficients are highly insignificant with p-values of 0.4 and over. So unobservables seem to affect reporting behaviour only in a selection of the health domains considered.

Table 7: Correlations between the error terms of the model per health domain.

	Correlation coefficients between the error terms			
	coef. (constant)	*st.dev*	*Z-value*	*p-value*
Pain	0.1261	0.0222	5.69	0.000
Mobility	0.0120	0.0239	0.50	0.615
Sleep	0.0778	0.0224	3.48	0.001
Breath	0.0205	0.0244	0.84	0.402
Depress	0.0778	0.0232	3.35	0.001
Work disability	0.0116	0.0240	0.48	0.629

To be able to evaluate the validity of the anchoring vignettes method it is however not enough to look at correlations. Therefore we test for the appropriateness of the vignettes method in the next sections. First, we explain the testing procedure in section 8.2; the test results will be discussed in section 8.3.

8.2 The testing procedure

Now that we have obtained the corrected health measure it is time to test for the appropriateness of the vignettes method as a correction procedure. If anchoring vignettes are an appropriate tool for correcting self-reports corrected health measures should give a better approximation of the true objective health status of a respondent than the uncorrected self-reported health measures. Outcome variables like various types of health care usage, labour market status and the results of the objective measured tests should depend on the true health status of a respondent in the end. So when the corrected latent health levels are closer to the true health status of a respondent than the uncorrected ones, one would expect that the corrected health measure explains a larger part of the variation in the outcome measure than do the uncorrected self-reports. This in particular is what we will utilize in our testing procedure for the validity of the vignettes method. When we incorporate both corrected and uncorrected health measures as explanatory variables in a regression framework and the vignettes method does a good job in approaching the true health status, the coefficients of the corrected health measure should take up the effect of health on the outcome variable under

consideration. On the other hand, the effect of the uncorrected health measures on the outcome variable should almost vanish with regard to the effect of the corrected measures. A simple comparison of the significance of the coefficients on the corrected and uncorrected health measures will indicate which of these two measures does a better job.

This idea for testing the quality of using anchoring vignettes as a correction tool can be applied to a broad range of outcome variables. Detailed measures of outcomes like health care usage, employment status and the results for objective measured tests like the walking speed test and the chair stand test are all available in the SHARE dataset. Therefore we can execute a range of regressions and look at whether the results all point at one direction as regards the value of the vignettes method. Before evaluating the relative explanatory power of corrected and uncorrected health measures, we estimate a regression equation with only the uncorrected self-report included. This is done in order to determine whether the (uncorrected) health measure has an influence on the outcome variable. If the uncorrected health measure has no significant effect on the outcome variable it is not useful to consider the effect of the corrected health measure. I.e., if the self-report is insignificant, we could expect an insignificant role of health in the outcome variable and therefore it would not be useful to compare the explanatory power of corrected and uncorrected health measures in that case.

When the self-report is found to have an effect on the outcome measure, we can evaluate the relative explanatory power of the corrected and uncorrected health measures. The general form of the regression equations that we will be estimating for this purpose is presented in Equation 24 and Equation 25. If the outcome variable is a dummy variable we use a probit regression as described in Equation 24. If the outcome variable is a continuous variable we use an OLS regression instead. This OLS regression equation is provided in Equation 25.

$$P\left(Y_{i,t} = 1 | Y_{i,t-1}, Y_{ri,t-1}, hpred_{i,t-1}, X_{i,t-1}\right) = \Phi\left(\beta_0 + \beta_1 \cdot Y_{i,t-1} + \beta_2 \cdot Y_{ri,t-1} + \beta_3 \cdot hpred_{i,t-1} + \beta_4 \cdot X_{i,t-1}\right) \quad (24)$$

$$Y_{i,t} = \beta_0 + \beta_1 \cdot Y_{i,t-1} + \beta_2 \cdot Y_{ri,t-1} + \beta_3 \cdot hpred_{i,t-1} + \beta_4 \cdot X_{i,t-1} + u_{i,t} \quad (25)$$

In the regression equations the dependent variable is a specific outcome measure in wave 2 ($Y_{i,t}$). This might be the usage of a specific type of health care, a dummy for the employment status of a respondent or the result of an objective test. The right-hand side of the equations consists of a constant, the lagged value of the outcome measure, the wave 1 uncorrected self-reported health measure for a particular health domain ($Y_{ri,t-1}$), the wave 1 corrected health measure ($hpred_{i,t-1}$) for a particular health domain, corresponding to the health domain used for the self-reported health measure, and a vector of individual characteristics in wave 1 ($X_{i,t-1}$). The lagged value of the outcome measure is included since for some outcomes there may be considerable persistence over time. Excluding the lagged value of the outcome measure may then result in significant omitted variable bias. We have tried to include several individual characteristics[47]; in the final specification we have only included those that were significant. Therefore, the composition of the vector of individual characteristics differs somewhat depending on the outcome measure under consideration. However, for a particular outcome measure we kept the composition of this vector equal across health

[47] We considered the following individual characteristics: the several age dummies, a dummy for being female, a dummy for being married, the education dummies, the country dummies, a dummy describing the ethnicity of respondents, the household size and the logarithm of income.

domains. So, the specification of the regression equations differs somewhat across outcome variables, but for a specific outcome measure it is the same across health domains.

The test regression equations can be estimated for each health domain for which we have obtained corrected health measures. There may be differences in the explanatory power of the corrected and uncorrected self-reports across domains. This may be due to two things: firstly, it may be possible that one domain is more relevant to the outcome variable under consideration than another. Secondly, it may be that the vignettes method does a better job in correcting self-reports in one domain than in another. Voňková & Hullegie (2010) have investigated this latter point and found that the vignettes method is sensitive to the health domain for which vignettes are used.

We can assess the explanatory power of both the corrected and uncorrected health measures by looking at their t-statistics and significance levels in the estimated regression equations. In general we expect one of the following four situations to appear in our test regressions.

Firstly, it is possible that neither the corrected health measure nor the uncorrected self-report is significant in the test regression. In that case, it might be possible that health does not have a significant effect on the specific outcome variable under consideration. Apparently, when the true health of a respondent does not influence the outcome variable, we would also expect that the corrected and uncorrected health measures are unrelated to the outcome variable since these proxy true health to some extent. If the self-report was significant before including the corrected health measure, observing insignificant coefficients on both types of health measures in our test regressions prevents us from drawing conclusions about the quality of the vignettes method.

Secondly, it may be possible that we observe a significant coefficient on the corrected health measure, though an insignificant coefficient on the self-report. This could be interpreted as corrected latent health being closer to the true health status than uncorrected self-reported health measures. This would then be evidence in favour of the vignettes method as a valid procedure for correcting self-reported health measures.

Thirdly, we could find an insignificant coefficient on the corrected latent health variable but a significant coefficient on the uncorrected self-report. This might then point at the uncorrected self-reported health measure being closer to the true health status than the corrected health measure. We could interpret this as evidence against the quality of the vignettes method as a tool to correct for reporting heterogeneity.

Finally, we could observe a significant coefficient for both the corrected as well as the uncorrected health measure. In that case we cannot draw firm conclusions on the appropriateness of the vignettes method for correcting observed self-reported health measures. However, we may deduce something about the relative explanatory power of the corrected and uncorrected self-report if there are differences in the levels at which the coefficients are significant. For instance, if the self-reported health measure is significant at the 1% significance level, while the corrected health measure is only significant at the 10% significance level, we could still interpret this as weak evidence against the validity of the vignettes method. On the other hand, if the significance of the corrected health measure is more obvious, this can be interpreted as weak evidence in favour of the vignettes method.

8.3 Test results

In the previous section we discussed the way in which we test for the appropriateness of the vignettes method as a correction tool. In this section we will provide and discuss the results of this testing procedure. We ran a large number of test regressions for various outcome measures and different health domains. The outcome measures that we considered will be discussed in section 8.3.1. The results of the various test regression will be discussed in section 8.3.2. Conclusions on the performance of the anchoring vignettes method are provided in section 8.3.3.

8.3.1 The outcome variables in the test regressions

One of the outcome variables that we considered is the labour market status of respondents. We generated dummies for several possible labour market states. More specifically, based on the employment status variable discussed in section 7.3.2, we generated a dummy for working, a dummy for being retired and a dummy for being disabled[48]. Then, we explain a particular labour market state of a respondent in wave 2 using the corrected and uncorrected health measures and a number of individual characteristics in wave 1 as explanatory variables. In addition, we include the labour market status of the respondents in the first wave. Per labour market status we estimated the test regression for each health domain and for the work disability domain.

Secondly, we can explain several types of health care usage from the corrected and uncorrected self-reported health measures. Important to note here is that one has to consider carefully whether the specific outcome variable will be influenced by the perceived health level of the respondent or his objective true health status. We would only like to think about outcome variables that are affected by the true health status of the respondent, since we want to evaluate whether the vignettes method does a good job in correcting the self-reports, i.e., in bringing them closer to the objective true health status of the respondent. If the outcome measure depends for instance on the perceived health status, the self-report may do a better job in explaining the outcome measure by definition. This may occur when we consider the number of contacts with a general practitioner in the last twelve months as the outcome variable. Subjectivity may play a role in visiting the general practitioner: when we have two individuals with the same true health condition, it may be that one of them thinks that it will disappear automatically such that he will decide not to go to the general practitioner. On the other hand, the other individual may be much more worried about the condition and will go to the general practitioner. This is what also may be taken up in the self-reported health measure since the individual who thinks that the condition will disappear automatically, will probably not report serious health limitations, while the individual who is more worried about his condition may be inclined to report serious problems. Therefore we cannot assess the quality of the vignettes method in approaching the true health status using a regression with the number of visits to the general practitioner as the dependent variable.

However, various types of health care usage remain that we expect to be determined by true health in the end. These can be used as outcome measures in our test regressions. For instance we generated a variable measuring the number of times that an individual has been in hospital in the last twelve months. In addition, we constructed a variable counting the number of times that a respondent has had outpatient surgery in the last twelve months. Furthermore, we considered out-

[48] We have not evaluated unemployed individuals since, in estimating the probit equation with unemployment as the outcome variable and the several age dummies as control variables, a considerable amount of observations drop out because no unemployed individuals appear for most of the age groups.

of-pocket outpatient care expenditures and out-of-pocket inpatient care expenditures as dependent variables in our test regressions. Finally, we also incorporated out-of-pocket expenditures for prescribed medicines and out-of-pocket expenditures for nursing home care, day-care and home care as outcome variables. Typically, these outcome variables are expected to depend on the true health level of the respondent, i.e., the true health care needs of individuals, such that they can be used in our testing procedure.

Finally, we can explain the results of the measured tests discussed in sections 7.1.1 and 7.3.3 using the corrected and uncorrected self-reports. For instance, the walking speed test result can be used as a dependent variable in our test regressions. It is to be expected that the results from the walking speed test depend on the true ability or health status of a respondent. So in that case our testing procedure applies. We used the walking speed variable measuring the time in seconds it took the respondent to cover a particular distance divided by the distance, i.e., the walking speed in meters per second. We can do the same for the other measured tests, for instance the immediate and delayed recall tests, the verbal fluency test and the numeracy test for the cognition domain and the grip strength test, chair stand test and peak flow test. One remark must be added to the usage of the cognitive measured tests as dependent variables in the testing procedure: we have not estimated the bivariate probit model for the memory domain, such that we can only use corrected and uncorrected health measures in the other domains to explain the results of the cognitive tests. These test regressions may be less powerful, since the memory self-report is the most likely to be significant in the first place.

8.3.2 Estimated regression equations[49]

In estimating the test regression equations not all observations can be used because of missing values for the outcome variables, the self-assessments, the individual characteristics and/or the corrected latent health variables. First, we estimate the regression model with the self-report only; thereafter, we add the corrected latent health level to the regression model. Estimation results for the test regressions using the labour market status dummies as outcome variables are provided per health domain in Table A4 - 1 to Table A4 - 6 in Appendix 4 on page 98.

When we look at the results of the regression equations with a dummy for a particular labour market status as the outcome variable, we can see that most of the variation in the labour market status in wave 2 is explained from the labour market status in wave 1, i.e., there is a very high level of persistence in labour market status. In addition, we observe that most of the times the age dummy variables are significant. For the labour market status “working” the logarithm of income and the education dummies have significant coefficients. The age dummies for the oldest individuals, i.e.,

[49] For some of the outcome variables mentioned in the previous section, estimation results are not provided here. This concerns out-of-pocket expenditures on outpatient care, out-of-pocket expenditures on inpatient care and out-of-pocket expenditures for nursing home care. For the former outcome variable the self-reports were insignificant in all health domains except for the work disability domain. Adding the corrected health measure in the latter domain gives significant coefficients on both health measures, but we cannot determine which health measure is better, although the corrected health measure at least explains part of the variation in the outcome measure. Regarding the second outcome variable, we observe insignificant coefficients on the self-report in most of the regressions. Only in the mobility and the work disability domain the self-report is significant. The corrected health measure is significant only at the 10% level in these health domains, while the self-report is significant at the 5% level. Finally, as regards the latter outcome variable we observe insignificant self-reports in all health domains except for the mobility domain. In the mobility domain the self-report remains significant after including the corrected health measure; at the same time the corrected health measure is insignificant, which can be interpreted as evidence against the vignettes method.

those aged 75 to 80 and individuals aged 80 to 85, dropped out when estimating the probit model since none of the respondents in these age groups were still working. In the regression equations with the dummy for retired as the dependent variable we included the following covariates: aged dummies, country dummies, a dummy for being female, a dummy for being married, the logarithm of income and the education dummies. Finally, for explaining the state of disability we included a dummy for being married, age dummies, education dummies and country dummies.

The regression equations with a dummy for working as the outcome variable show an insignificant coefficient on the self-report in each health domain. Therefore we could not compare the explanatory power of corrected and self-reported limitations in the various health domain using this outcome variable. When we evaluate the estimates of the regression equations using dummies for retired and disabled as outcome variables, more positive results are found. In the first column with estimates for the disablement equation, i.e., the column in which only the self-report has been included, the self-report is highly significant in most health domains, except for the breath domain. The second column then shows that adding the corrected health measure does not alter the significance of the self-report. The self-report remains significant at the 1% significance level. This is the case for all health domains in which the self-report was significant in the first specification, i.e., all health domains except for the breath domain. When we look at the significance of the corrected health measure, results differ somewhat across health domains. For the pain domain, the depress domain and the work disability domain the corrected health measure is highly insignificant (p-values of 0.472, 0.355 and 0.602 respectively). However, the estimates for the mobility domain and the sleep domain show significant coefficients on the corrected health measure at the 5% significance level. Interesting to see is that the coefficients on the self-reported health measures are positive in all health domains, while the coefficients on the corrected health measures are negative in all health domains. So although a higher value of the self-report, i.e., more limitations in the specific health domain under consideration, increases the probability of being disabled, a higher value of the corrected latent health variable decreases the probability of being disabled given the other individual characteristics. This is a somewhat odd observation, since a higher value of corrected latent health also implies more limitations and thus worse health such that one would expect a positive coefficient on this corrected health variable too. Evaluating the significance of the corrected and uncorrected health measures one can thus see that the self-report explains a larger amount of the variation in the prevalence or reporting of disability than the corrected health measure does. Since we would expect that the labour market status of disability is largely determined by the true objective health status of a respondent, these regressions give us some evidence against the vignettes method. The evidence is however not very strong, since in two out of five health domains, both the corrected and uncorrected health measures are significant. In order to be able to draw firm conclusions, we would like to observe significance of either the corrected health measure or the uncorrected self-report instead of both being significant.

The retirement equations show that the self-report is only significant at the 10% significance level in the mobility domain. In the other health domains and the work disability domain the coefficient on the self-reported health measure is insignificant. Adding the corrected health measure to the regression equation does not alter the significance of the mobility self-report. The corrected health measure itself is highly insignificant with a p-value as large as 0.222. This shows that the explanatory power of the self-reported mobility limitation variable is larger than that of the corrected mobility limitations of an individual. Coefficients on both corrected and uncorrected health

are negative, meaning that an increase in the health variable, i.e., more limitations and thus worse health, are associated with a lower probability of being retired.

Summarizing the test results obtained using the labour market status of a respondent as the outcome variable we observe some weak evidence against the quality of the vignettes method as a correction procedure. However, most of the test regressions are inconclusive, since the self-report does not have a significant effect on the outcome variable in most cases.

We can perform a similar analysis using various types of health care usage as outcome variables. The results of these regressions for the several health domains are provided in Table A4 - 7 to Table A4 - 12 in Appendix 4 on page 104. Evaluating the regressions with the dummy for whether someone has been in hospital in the last twelve months as the outcome variable, we can see that the self-report is highly significant (at the 1% level) in all health domains. In addition, we observe a positive sign on this dummy variable which implies that a higher value for the self-report, i.e., worse health, is associated with a higher probability that the individual has been in hospital in the last twelve months. Including the corrected health measure does not affect the significance of the self-report. In most of the health domains the corrected health measure is insignificant. Only in the depress domain the corrected health variable is significant at the 10% level. But to conclude that the corrected health measure really affects the outcome variable we would like to observe significance at the 1% or 5% level. Since the self-report explains the largest part of the variation in the outcome variable compared to the corrected health measure in every health domain, we could conclude that the quality of the vignettes method in correcting health and bringing it closer to the true health status seems to be questionable.

Looking at the dummy variable for whether an individual has had outpatient surgery in the last twelve months shows that the self-report is significant only in the pain domain and the mobility domain. Moreover, we observe a positive sign, which is to be expected since logically worse health should be associated with a higher probability for an individual to have had surgery. Including the corrected health measures in these regression equations does not affect the significance of the self-report in a negative way. In fact, the self-report is now even significant at the 5% level in the mobility domain, while it was only significant at the 10% level before including the corrected health measure. In addition, the corrected health measure is insignificant in the pain domain and the mobility domain. So again, this indicates that the corrected health measure does not do a good job in explaining an objective health outcome, in this case surgery.

Finally, we can look at the regression with out-of-pocket expenditures on prescribed medicines as the outcome variable. For this outcome variable we observe highly significant coefficients on the self-reports in the pain domain and the breath domain. The self-report is significant at the 10% level in the mobility domain. Again, the estimated coefficients are positive, meaning that an increase in the self-report value, i.e., worse health, is associated with an increase in the out-of-pocket expenditures on prescribed medicines. Including the corrected health measure in the regression equation for the pain domain does not alter the significance of the self-report. Furthermore, the corrected health measure is insignificant. For the mobility domain, the self-report is no longer significant after including the corrected health measure. The corrected health measure itself is however also insignificant. Therefore we cannot draw any conclusions on the validity of the vignettes method based on these estimates. Finally, for the breath domain the significance of the self-report differs somewhat in both specifications of the regression equation. When the corrected health measure is included the self-report is significant at the 5% level, while before including the corrected measure significance at the 1% level could be observed. However, the coefficient on the

corrected health measure is insignificant. The coefficient on the corrected health measure is much larger in magnitude, but this cannot be interpreted directly since the scales of the corrected and uncorrected health measures are incomparable. Thus, in the pain and breath domain the included corrected health measure is insignificant, while the self-report is significant. This points at a larger explanatory power of the self-report. So, using the out-of-pocket expenditures on prescribed medicines as the outcome variable, there is also some evidence for the uncorrected health measures having a larger explanatory power than the corrected ones.

We also estimated such test regressions using the results from the objective measured tests, i.e., the walking speed test, the grip strength test, the chair stand test and the peak flow test, as outcome measures. Results for these test regressions for the several health domains can be found in Table A4 - 13 to Table A4 - 18 in Appendix 4 on page 110. Important to note is that most of the age dummies have been dropped while estimating the regression equation using walking speed as the outcome variable. This is caused by the fact that only individuals aged 76 and over were asked to do the walking speed test. This also results in a very small number of observations being used in the estimation procedure. We must bear this in mind when drawing conclusions. In the regressions explaining the results of the chair stand test and the peak flow test we were not able to include the lagged value of the outcome variable, since these tests have only been conducted in the second wave. Finally, we have no observations in our sample on test results for respondents from the Netherlands and Sweden, such that these country dummies are dropped when estimating the regression equations.

When we look at the results of these test regressions we can see that the pain, depress and sleep self-reports do not have a significant influence on the walking speed test result. So for these three domains we cannot evaluate the relative power of the corrected and uncorrected health measures. For the mobility domain on the other hand, we observe a highly significant coefficient on the self-report. The coefficient on the self-reported health measure is negative, meaning that a higher value of the self-report, i.e., worse health, is associated with a lower walking speed. Such a sign of the coefficient is to be expected. When we include the corrected health measure for mobility, we see that the significance of the self-reported health measure remains unchanged. Moreover, the corrected health measure has an insignificant effect on the walking speed test result. This implies that the self-report on health explains a larger part of the variation in the outcome measure than the corrected self-report. This could be interpreted as evidence against the quality of the vignettes method. However, only very few observations have been used in estimating the regression equation. In the test regression for the breath domain the coefficient on the self-reported health measure is also significant, but only at the 10% level. The same holds for the work disability domain. If the corrected health measures in the respective domains are included, the significance of the self-report in the breath domain is not affected. Furthermore, the corrected health measure in this domain is insignificant, which is again small evidence against the vignettes method. On the other hand, adding the corrected health measure to the regression equation in the work disability equation causes the self-reported health to become insignificant. However, the corrected health measure is also insignificant such that we cannot draw any conclusions on the validity of the vignettes method based on this result.

For the grip strength test we observe significant coefficients on the self-reported health measures in each health domain. In addition, the signs are negative which implies a smaller grip strength if the respondent reports more limitations. Adding the corrected health measures does not

alter the significance of the uncorrected self-report in most cases. For the pain domain and the depress domain the corrected health measure is insignificant, while the uncorrected one remains significant at the 1% level. In the mobility and sleep domain the significance of the self-report remains at the 1% level; in addition, the corrected health measure is significant though only at the 10% level. The coefficients on the corrected health measures are negative, which is to be expected. For the work disability domain the self-reported health measure was significant at the 5% level before adding the corrected health measure. Adding the latter measure causes the self-report to become significant only at the 10% level. However, the corrected self-report is insignificant in this domain. Therefore the uncorrected self-report still seems to explain the larger part of the variation in the grip strength test results, which seems to be negative as regards the validity of the vignettes method. The most positive results in favour of the vignettes method are observed in the breath domain. In this domain the self-reported health measure is significant at the 5% level in both regression equations. However, the corrected self-report is also significant at the 5% level. This may point at considerable explanatory power of the corrected self-report though ideally we would like to observe an insignificant coefficient on the self-report. However it can be interpreted as some weak evidence in favour of the vignettes method.

Regarding the peak flow test and the chair stand test the self-report is insignificant in most of the health domains. With respect to the peak flow test, a significant coefficient on the self-report is only observed in the breath domain. Adding the corrected health measure does not alter the significance of the self-reported health measure, though the corrected latent health variable is also significant at the 5% significance level. So, both the corrected health measure and the uncorrected self-report are significant at the 5% level, which points at some explanatory power of the corrected health measure, though we cannot tell which of the two health measures is better. For the chair stand test, the self-report is only significant at the 5% level in the mobility domain. Adding the corrected health measure here does not alter the significance of the self-report; the corrected health measure is also significant but only at the 10% level. Therefore, the self-report explains the largest part of the variation in the outcome measure. Thus, the corrected health measure does not seem to be closer to the true health status, which is a negative conclusion as regards the quality of the vignettes method.

8.3.3 Conclusions

Estimation results for 60 test regressions have been provided: for five health domains and the work disability domain we considered three regressions using labour market states as outcome variables, three regressions using various types of health care usage indicators as outcome variables and four regressions using the objective measured tests as outcome variables. A large part of these test regressions are inconclusive. In a considerable fraction of the test results[50] we observe insignificant coefficients on the self-reported health measure before adding the corrected health measure such that we cannot evaluate the relative explanatory power of the corrected health measure and the uncorrected self-report. The regressions in which the coefficient on the self-report is insignificant in the first column, i.e., in the specification without the corrected health measure included, cannot be used to evaluate the validity of the vignettes method.

However, also test regressions have been provided in which the self-report is (highly) significant. In these cases, the corrected health measure can be included in the regression equation

[50] We observe insignificant coefficients on the self-reported health measures in 32 of the 60 test regressions.

so as to observe the relative explanatory power of the corrected and uncorrected health measure. We have not found clear positive results on the validity of the vignettes method. A positive result would imply that the corrected health measure is highly significant, while the uncorrected self-report is insignificant at the same time. This is observed in none of the test regressions. The results that we found are either inconclusive or somewhat negative about the quality of the vignettes method as a correction procedure. We observe a considerable number of test regressions in which the self-reports remain significant after including the corrected health measure, while at the same time the corrected health measure is (highly) insignificant in explaining the variation in the outcome measure[51]. This is typically what one would expect to observe if the corrected health measure is not closer to the true health status, assuming that the outcome measure depends on the true health status of a respondent. When the corrected health measure is not closer to the objective true health status, then the validity of the vignettes method as a correction tool may be questionable, since the method has been developed in order to better approach the true health status of an individual. Furthermore, some test regressions are neither positive nor negative about the validity of the vignettes method. This is the case when both corrected and uncorrected health measures are insignificant, while the uncorrected self-report was significant before including the corrected health measure[52]. It is also the case when both corrected and uncorrected health measures are significant[53]. If differences in the significance levels exist, we can say something about the relative explanatory power of the health measures, but no firm conclusions can be reached.

Overall, evaluating the results of all test regressions would induce one to be at least suspicious in using the vignettes method as a correction tool. Although, our results are not indisputable in rejecting the vignettes method, we do observe some negative results on the quality of the method. Moreover, it is interesting to see that in neither case we observe positive results for the validity of the vignettes method. So in fact, based on our results, we would conclude that the potential problems discussed in section 6 may indeed affect the quality of the vignettes method.

[51] This is observed in 18 of the 60 test regressions.
[52] We observe insignificant coefficients on both the self-report and the corrected health measure in 2 test regressions.
[53] This is observed in 8 of the 60 test regressions.

9. Conclusion

In this thesis we evaluated the appropriateness of the anchoring vignettes method for correcting self-reported health measures. The usage of (self-reported) health measures in economics is widespread. Self-reports are intended to be an adequate proxy for the true health status, which is itself unobserved. However, a major problem with self-reported health measures is reporting heterogeneity. If reporting heterogeneity exists, individuals may translate some given fixed health level into different categorical responses. Although true latent health of these individuals is the same, we will observe health differences for these individuals when using the self-report as a proxy for true health. Several solutions have been proposed to alleviate or circumvent these difficulties. For instance, it has been suggested to include medical examinations and objective measured tests instead of or in addition to the self-reports. But, these various solutions were also found to be imperfect. Hence, the anchoring vignettes method has been developed as a way to correct for reporting heterogeneity and to obtain cleansed self-reported health measures. The basic idea utilized in this method is that the way in which individuals evaluate the health status of a hypothetical person can be used to anchor the reporting behaviour of an individual. Then, differences in reporting behaviour across individuals can be corrected for. However, the validity of this method rests on two assumptions: *vignette equivalence* and *response consistency*. If these assumptions fail the quality of the vignettes method may deteriorate significantly.

Previous studies have shown mixed results on the quality of the vignettes method. We conducted a different type of test in this thesis. Instead of testing whether the assumptions hold directly, we test for the relative explanatory power of self-reported health measures and health measures corrected for reporting heterogeneity in a regression framework. We chose several dependent variables that we would expect to be determined by true health in the end. These are then regressed on the lagged value of this dependent variable, the self-report, the corrected health measure and a vector of covariates. The corrected health measure will explain a larger part of the variation in the outcome variable than the self-report if the vignettes method brings the corrected measure closer to the true health status, which is a proxy for the method being appropriate.

Although in a large part of our regressions the self-report turned out to be insignificant in the first place, also a considerable part of the regressions allowed us to interpret the quality of the corrected health measure. The results that we found are negative on the quality of the vignettes method. We observed insignificant coefficients on the corrected health measure in the largest part of the regressions, though the significance of the self-reports was unaltered by including the corrected health measures. Of those regressions in which the corrected health measure turned out to be significant, it involved significance only at the 10% significance level in most of the cases, while we would like to observe significance at the 1% or 5% level. The most positive result that we found is a regression in which both the self-report as well as the corrected health measure are significant at the 5% level. But in general, the corrected health measure does not seem to do much.

Our negative findings may not be that strange to find, since previous studies have already observed problems with the satisfaction of the assumptions of *vignette equivalence* and *response consistency*. However, we have to bear in mind that our sample only consists of individuals aged 50 to 85, such that we cannot say much about the quality of the method for other age groups. We could expect to observe differences in the quality though, since the magnitude of the bias present in self-

reports may also vary across age groups. Moreover, it is interesting to evaluate the quality of the vignettes method when different vignettes are used. We have incorporated only the first vignette, i.e., the one representing fewest limitations. But, since Voňková & Hullegie (2010) found that the method is sensitive to the choice of the vignette, we may also want to evaluate its quality using different vignettes.

However, our results do seem to call for improvements of the vignettes method. Such improvements can look at the way in which the assumptions can be made more likely to hold. Vignette equivalence requires a clear and objective description of the vignette person's health status. This calls for accurate descriptions and there is probably some more room for improvements here, as has also been indicated by Bago d'Uva et al. (2009). The recent improvement suggested by Hopkins & King (2010) to change the order in which self-reports and vignettes are provided in surveys can also be tested. Finally, one may also continue to search for other methods to correct self-reported health measures, especially if future research finds additional negative results on the quality of the vignettes method.

References

Anderson, K.H. & Burkhauser, R.V. (1985). The Retirement-Health Nexus: A New Measure of an Old Puzzle. *Journal of Human Resources*, 20(3), 315-330

Bago d'Uva, T. & Doorslaer, E. van & Lindeboom, M. & O'Donnell, O. (2008). Does Reporting Heterogeneity Bias the Measurement of Health Disparities. *Health Economics*, 17(3), 351-375

Bago d'Uva, T. & O'Donnel, O. & Doorslaer, E. van (2008). Differential Health Reporting by Education Level and Its Impact on the Measurement of Health Inequalities Among Older Europeans. *International Journal of Epidemiology*, 37(6), 1375-1383

Bago d'Uva, T. & Lindeboom, M. & O'Donnell, O. & Doorslaer, E. van (2009). Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity. Tinbergen Institute Discussion Paper, No. 2009-091/3

Banks, J. & Emmerson, C. & Tetlow, G. (2007). Healthy Retirement or Unhealthy Inactivity: How Important Are Financial Incentives in Explaining Retirement? Conference Paper Institute of Fiscal Studies, April 2007

Berg, G.J. van den & Lindeboom, M. (2007). Birth is the Messenger of Death, But Policy May Help to Postpone the Bad News. Netspar Panel Paper 4, September 2007. Consulted on June 21, 2010 via <http://www.netspar.nl/events/panel/2007/oct/papervandenberg.pdf>

Bound, J. (1991). Self-Reported versus Objective Measures of Health in Retirement Models. *Journal of Human Resources*, 26(1), 106-138

Bound, J. & Schoenbaum, M. & Stinebrickner, T.R. & Waidmann, T. (1999). The Dynamic Effects of Health on the Labour Force Transitions of Older Workers. *Labour Economics*, 6(2), 179-202

Bound, J. & Stinebrickner, T. & Waidmann, T. (2010). Health, Economic Resources and the Work Decisions of Older Men. *Journal of Econometrics*, 156(1), 106-129

Butler, J.S. & Burkhauser, R.V. & Mitchell, J.M. & Pincus, Th.P. (1987). Measurement Error in Self-Reported Health Variables. *Review of Economics and Statistics*, 69(4), 644-650

Chirikos, T.N. & Nestel, G. (1984). Economic Determinants and Consequences of Self-Reported Work Disability. *Journal of Health Economics*, 3(2), 117-136

Crossley, T.F. & Kennedy, S. (2002). The Reliability of Self-Assessed Health Status. *Journal of Health Economics*, 21(4), 643-658

Dwyer, D.S. & Mitchell, O.S. (1999). Health Problems as Determinants of Retirement: Are Self-Rated Measures Endogenous? *Journal of Health Economics*, 18(2), 173-193

Garcia-Gomez, P. & Gaudecker, H.M. von & Lindeboom, M. (2010). Health, Disability and Work: Patterns for the Working Age Population. Netspar Panel Paper 17, January 2010. <http://www.netspar.nl/events/2009/annual/paperlindeboom.pdf>

Gruber, J. & Wise, D.A. (1998). Social Security and Retirement: An International Comparison. *American Economic Review*, 88(2), 158-163

Gupta, N.D. & Kristensen, N. & Pozzoli, D. (2010). External Validation of the Use of Vignettes in Cross-Country Health Studies. *Economic Modelling*, 27(4), 854-865

Heij, C. & Boer, P. De & Franses, P.F. & Kloek, T. & Dijk, H.K. van (2004). *Econometric Methods with Applications in Business and Economics*. New York: Oxford University Press

Hopkins, D.J. & King, G. (2010). Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability. Consulted on May 7, 2010 via <http://people.iq.harvard.edu/~dhopkins/implement.pdf>

Iburg, K.M. & Salomon, J.A. & Tandon, A. & Murray, C.J.L. (2002). Cross-population Comparability of Physician-Assessed and Self-Reported Measures of Health. In: *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*, edited by Murray, C.J.L. & Salomon, J.A. & Mathers, C.D. & Lopez, A.D. (2002). World Health Organization

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

Jones, A.M. & Rice, N. & Bago d'Uva, T. & Balia, S. (2007). Code for programming the HOPIT model. Consulted on May 29, 2010 via <http://www.york.ac.uk/res/herc/research/hedg/software.htm>

Jürges, H. (2007). True Health vs Response Styles: Exploring Cross-Country Differences in Self-Reported Health. *Health Economics*, 16(2), 163-178

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

Kapteyn, A. & Smith, J.P. & Soest, A. van (2009). Work Disability, Work and Justification Bias in Europe and the U.S.. RAND Center for the Study of Aging, Working Paper WR-696.
Consulted on 02-04-2010 via <http://www.rand.org/pubs/working_papers/WR696/>

Kapteyn, A. & Smith, J.P. & Van Soest, A. (2010). Work Disability, Work and Justification Bias in Europe and the US. Presented by Arie Kapteyn during the Netspar Mini Theme Conference on Anchoring Vignettes. Rotterdam, April 28, 2010. Slides can be consulted via <http://www.netspar.nl/events/2010/apr28/program/> (last consulted on May 7, 2010)

Kerkhofs, M. & Lindeboom, M. (1995). Subjective Health Measures and State Dependent Reporting Errors. *Health Economics*, 4(3), 221-235

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(1), 567-583

Lindeboom, M. & Doorslaer, E. van (2004). Cut-point Shift and Index Shift in Self-Reported Health. *Journal of Health Economics*, 23(6), 1083-1099

Lindeboom, M. (2005). Health and Work of Older Workers. Consulted on June 21, 2010 via <http://www2.eur.nl/bmg/ecuity/public_papers/ECuity3wp29Lindeboom.pdf>

Lindeboom, M. & Kerkhofs, M. (2009). Health and Work of the Elderly: Subjective Health Measures, Reporting Errors and Endogeneity in the Relationship Between Health and Work. Journal of Applied Health Economics, 24(6), 1024-1046

Murray, C.J.L. & Tandon, A. & Salomon, J.A. & Mathers, C.D. (2000). Enhancing Cross-Population Comparability of Survey Results. Global Programme on Evidence for Health Policy, Discussion Paper no. 35. World Health Organization, Geneva, Switzerland

Murray, C.J.L. & Tandon, A. & Salomon, J.A. & Mathers, C.D. & Sadana, R. (2002). Cross-population Comparability of Evidence for Health Policy. Global Programme on Evidence for Health Policy, Discussion Paper no. 46. World Health Organization, Geneva, Switzerland

Sadana, R. & Mathers, C.D. & Lopez, A.D. & Murray, C.J.L. & Iburg, K.M. (2002). Comparative Analyses Of More Than 50 Household Surveys On Health Status. In: *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*, edited by Murray, Ch.J.L. & Salomon, J.A. & Mathers, C.D. & Lopez, A.D. (2002). World Health Organization

Salomon, J.A. & Tandon, A. & Murray, C.J.L. (2001). Using Vignettes to Improve Cross-Population Comparability of Health Surveys: Concepts, Design, and Evaluation Techniques. Global Programme on Evidence for Health Policy, Discussion Paper no. 41. World Health Organization, Geneva, Switzerland

Salomon, J.A. & Tandon, A. & Murray, C.J.L. (2004). Comparability of Self Rated Health: Cross Sectional Multi-Country Survey Using Anchoring Vignettes. BMJ (British Medical Journal), 328(7434), 258. Consulted on April 3, 2010 via <http://www.bmj.com/cgi/content/full/bmj;328/7434/258>

SHARE (Survey of Health, Ageing and Retirement in Europe). Guide to Release 2.3.0 Waves 1 & 2. Version: November 13, 2009, updated December 10, 2009. Last consulted on June 21, 2010 via <www.share-project.org>

SHARE (Survey of Health, Ageing and Retirement in Europe). Questionnaires Wave 1 and Questionnaires Wave 2. Consulted on May 5, 2010 via <www.share-project.org>

SHARE (Survey of Health, Ageing and Retirement in Europe). Sample. Consulted on May 24, 2010 via <www.share-project.org>

SHARE. Tackling the Demographic Challenge: The Survey of Health, Ageing and Retirement in Europe. Consulted on May 24, 2010 via <http://www.share-project.org/t3/share/fileadmin/SHARE_Brochure/share_broschuere_web_final.pdf>

Shmueli, A. (2003). Socio-Economic and Demographic Variation in Health and Its Measures: The Issue of Reporting Heterogeneity. *Social Science & Medicine*, 57(1), 125-134

Soest, A. van & Delaney, L. & Harmon, C. & Kapteyn, A. & Smith, J.P. (2007). Validating the Use of Vignettes for Subjective Threshold Scales. IZA Discussion Paper, no. 2680

Soest, A. van & Kapteyn, A. & Smith, J.P. & Voňková, H. (2010). Anchoring Vignettes and Response Consistency. Preliminary version presented by Arthur van Soest during the Netspar Mini Theme Conference on Anchoring Vignettes. Rotterdam, April 28, 2010. Slides can be consulted via <http://www.netspar.nl/events/2010/apr28/program/> (last consulted on June 21, 2010)

Stern, S. (1989). Measuring the Effect of Disability on Labor Force Participation. *Journal of Human Resources*, 24(3), 361-395

Stock, J.H. & Watson, M.W. (2007). *Introduction to Econometrics.* Boston: Pearson Education Inc.

Tandon, A. & Murray, C.J.L. & Salomon, J.A. & King, G. (2002). Statistical Models for Enhancing Cross-Population Comparability. Global Programme on Evidence for Health Policy, Discussion Paper no. 42. World Health Organization, Geneva, Switzerland

Voňková, H. & Hullegie, P. (2010). Is the Anchoring Vignettes Method Sensitive to the Domain and Choice of the Vignette? Netspar Discussion Paper 01/2010-004

Wagstaff, A. & Doorslaer, E. Van (2000). Measuring and Testing for Inequity in the Delivery of Health Care. *Journal of Human Resources*, 35(4), 716-733

Appendices

Appendix 1 Objective health measures and health limitations

Table A1 - 1: Descriptive statistics for objective measures: cognitive tests.

	Verbal fluency			Immediate recall			Delayed recall			Numeracy		
	# obs.	mean	st.dev	# obs.	mean	st.dev	# obs.	mean	st.dev	# obs.	mean	st.dev
Wave 1												
All countries	2,968	18.9832	7.2107	2,978	4.9040	1.7620	2,980	3.4406	1.9918	2,987	3.3432	1.1115
Germany	266	20.4587	6.5824	268	5.6194	1.5784	269	3.8848	1.8238	267	3.6891	0.9985
Sweden	282	24.4575	6.8911	282	5.5319	1.5467	282	4.1702	1.7122	282	3.8050	0.9624
Netherlands	351	20.8803	5.8615	351	5.3789	1.7375	351	4.0484	2.0731	354	3.7062	1.0530
Spain	238	15.0042	5.7936	239	3.6067	1.6257	239	2.2469	1.7495	239	2.5021	1.0527
Italy	315	14.4857	5.8649	318	4.1289	1.6715	318	2.5094	1.6901	320	2.8781	1.0297
France	591	20.9459	7.8469	595	4.6672	1.7658	596	3.2970	1.8373	600	3.2567	1.1371
Greece	487	14.7023	4.6507	487	5.0903	1.6038	487	3.6222	2.1404	487	3.3901	1.0942
Belgium	438	20.5502	6.6126	438	5.0662	1.6907	438	3.5320	2.0006	438	3.4064	0.9916
Wave 2												
All countries	2,943	18.8118	7.2306	2,952	4.9634	1.7543	2,952	3.5335	1.9795	2,978	3.3486	1.1372
Germany	263	21.2738	7.4466	264	5.6212	1.5986	264	4.1705	1.7992	269	3.6766	1.1010
Sweden	281	23.0925	6.9549	282	5.4965	1.7025	282	4.3404	1.8896	281	3.6762	0.9701
Netherlands	352	20.9574	6.1820	353	5.5779	1.7239	353	4.2635	2.1824	353	3.7365	1.0613
Spain	232	14.0474	5.3495	232	3.8190	1.7061	232	2.4871	1.7182	239	2.5649	1.0861
Italy	316	14.9747	6.0421	317	4.2397	1.7042	317	2.7855	1.8686	320	2.8844	1.0604
France	582	20.3952	7.5917	584	4.8664	1.7608	584	3.4469	1.8963	592	3.2956	1.1831
Greece	482	14.5747	4.9248	484	4.7789	1.5386	484	3.2541	1.7959	487	3.4251	1.1009
Belgium	435	20.7264	6.6299	436	5.1927	1.6371	436	3.5619	1.9413	437	3.3776	1.0456

Table A1 - 2: Descriptive statistics for objective measures: grip strength, walking speed, chair stand and peak flow test.

	Grip strength			Walking speed			Chair stand			Peak flow		
	# obs.	mean	st.dev	# obs.	mean	st.dev	# obs.	mean	st.dev	# obs.	mean	st.dev
Wave 1												
All countries	2,855	34.5103	12.0889	264	0.6804	0.2884	-	-	-	-	-	-
Germany	263	37.0875	11.2710	13	0.7077	0.1413	-	-	-	-	-	-
Sweden	274	36.8613	12.6115	22	0.8094	0.2899	-	-	-	-	-	-
Netherlands	345	37.6638	10.9753	18	0.9069	0.3002	-	-	-	-	-	-
Spain	229	30.1703	11.0922	28	0.5841	0.2883	-	-	-	-	-	-
Italy	300	31.3500	10.9636	28	0.6175	0.2229	-	-	-	-	-	-
France	570	33.8614	12.0774	80	0.6381	0.2181	-	-	-	-	-	-
Greece	445	33.0202	13.0189	33	0.6367	0.2498	-	-	-	-	-	-
Belgium	429	35.8275	11.9116	42	0.7287	0.4138	-	-	-	-	-	-
Wave 2												
All countries	2,775	34.5968	11.6143	290	0.7380	0.4079	1,512	11.6531	7.3240	2,040	327.6456	138.1825
Germany	252	36.0476	10.8693	16	0.7592	0.2861	192	10.5679	6.9424	242	347.7335	147.2576
Sweden	279	37.0753	11.9003	38	0.8046	0.3150	-	-	-	-	-	-
Netherlands	343	35.6443	10.9259	31	0.7490	0.2993	-	-	-	-	-	-
Spain	211	30.1185	10.6064	23	0.8441	0.8161	161	12.2322	10.2142	207	324.0024	116.3513
Italy	286	32.3217	11.5002	24	0.5335	0.2061	205	10.2866	4.2744	293	365.9608	144.7651
France	544	33.6820	11.9012	74	0.7916	0.4714	347	11.5538	5.6590	469	306.5330	131.4105
Greece	434	34.9747	11.3346	28	0.5359	0.3131	335	12.9486	9.2779	431	329.7030	141.2726
Belgium	426	35.8005	11.9823	56	0.7548	0.2409	272	11.6377	6.1848	398	311.7701	135.5452

Table A1 - 3: Descriptive statistics for several aspects of the respondents' health status (wave 1).

	# ADL limitations			# IADL limitations			# mobility limitations		
	# obs.	mean	st.dev	# obs.	mean	st.dev	# obs.	mean	st.dev
All	2,994	0.1313	0.5638	2,994	0.1937	0.6765	2,994	1.2488	1.8911
DE	270	0.1111	0.4590	270	0.1296	0.4500	270	1.1704	1.4937
SE	284	0.1127	0.5263	284	0.1408	0.5775	284	0.9120	1.6070
NL	354	0.0734	0.4392	354	0.1412	0.5452	354	0.8588	1.6533
ES	239	0.1590	0.6735	239	0.3138	0.9426	239	2.0502	2.5842
IT	320	0.1781	0.7571	320	0.2719	0.9655	320	1.5188	2.1305
FR	602	0.1445	0.5102	602	0.1860	0.6229	602	1.2724	1.8093
GR	487	0.1109	0.5915	487	0.1910	0.6499	487	1.1684	1.8304
BE	438	0.1575	0.5415	438	0.2009	0.6057	438	1.2534	1.8569

	# chronic diseases			Long-term illness			Depression		
	# obs.	mean	st.dev	# obs.	yes (%)	no (%)	# obs.	mean	st.dev
All	2,994	1.4796	1.4084	2,994	45.62%	54.38%	2,971	2.2885	2.1947
DE	270	1.2556	1.2663	270	58.52%	41.48%	269	1.9368	1.7212
SE	284	1.3908	1.2119	284	55.28%	44.72%	284	1.6690	1.6823
NL	354	1.1582	1.2033	354	37.29%	62.71%	352	1.6420	1.8420
ES	239	1.8033	1.6111	239	56.90%	43.10%	236	2.8644	2.8237
IT	320	1.8094	1.7504	320	47.19%	52.81%	320	2.8906	2.6268
FR	602	1.4900	1.2902	602	47.34%	52.66%	592	2.6875	2.1748
GR	487	1.3819	1.4003	487	28.75%	71.25%	483	2.1656	2.1676
BE	438	1.6119	1.4399	438	47.26%	52.74%	435	2.2713	2.0365

Note: the following country abbreviations are used: All countries (All), Germany (DE), Sweden (SE), the Netherlands (NL), Spain (ES), Italy (IT), France (FR), Greece (GR) and Belgium (BE).

Table A1 - 4: Descriptive statistics for several aspects of the respondents' health status (wave 2).

	# ADL limitations			# IADL limitations			# mobility limitations		
	# obs.	mean	st.dev	# obs.	mean	st.dev	# obs.	mean	st.dev
All	2,990	0.1719	0.7130	2,990	0.2896	0.9343	2,990	1.4592	2.1492
DE	270	0.1815	0.7217	270	0.2111	0.7736	270	1.3667	1.9763
SE	284	0.1021	0.4370	284	0.1655	0.6381	284	1.0317	1.6863
NL	354	0.0763	0.4729	354	0.1921	0.7277	354	0.9040	1.7318
ES	239	0.4100	1.2732	239	0.5983	1.5949	239	2.1883	2.8318
IT	320	0.2438	0.8509	320	0.4188	1.1363	320	1.8406	2.4881
FR	598	0.1572	0.6304	598	0.2492	0.8306	598	1.3562	2.0345
GR	487	0.1376	0.6876	487	0.3060	0.9356	487	1.7166	2.2134
BE	438	0.1644	0.5704	438	0.2717	0.7604	438	1.4201	2.0186

	# chronic diseases			Long-term illness			Depression		
	# obs.	mean	st.dev	# obs.	yes (%)	no (%)	# obs.	mean	st.dev
All	-	-	-	2,989	45.77%	54.23%	2,941	2.2098	2.1995
DE	-	-	-	270	57.78%	42.22%	266	2.0639	1.9052
SE	-	-	-	283	54.42%	45.58%	279	1.6918	1.6310
NL	-	-	-	354	39.27%	60.73%	351	1.6068	1.7661
ES	-	-	-	239	59.00%	41.00%	232	2.9181	2.6996
IT	-	-	-	320	45.94%	54.06%	317	2.9117	2.5739
FR	-	-	-	598	48.83%	51.17%	579	2.5423	2.1708
GR	-	-	-	487	29.36%	70.64%	482	1.7178	2.0972
BE	-	-	-	438	44.75%	55.25%	435	2.3310	2.2238

Note: the following country abbreviations are used: All countries (All), Germany (DE), Sweden (SE), the Netherlands (NL), Spain (ES), Italy (IT), France (FR), Greece (GR) and Belgium (BE).

Appendix 2 HOPIT estimation results

Table A2 - 1: HOPIT estimation results for the pain domain.

	Latent own health				Cut-point 1		
	Coef.	*Std.err.*	*Z-value*		*Coef.*	*Std.err.*	*Z-value*
Age55to60	0.1771	0.1110	1.60	Age55to60	0.2023	0.0799	2.53
Age60to65	0.1875	0.1169	1.60	Age60to65	0.0937	0.0856	1.09
Age65to70	0.3475	0.1243	2.80	Age65to70	0.1690	0.0892	1.89
Age70to75	0.4682	0.1318	3.55	Age70to75	0.1473	0.0953	1.55
Age75to80	0.6594	0.1530	4.31	Age75to80	0.0991	0.1121	0.88
Age80to85	0.6707	0.1845	3.64	Age80to85	-0.0246	0.1341	-0.18
Female	0.6279	0.0730	8.60	Female	-0.0117	0.0520	-0.23
Low_educ	0.3475	0.1073	3.24	Low_educ	0.1997	0.0772	2.59
Middle_educ	0.1714	0.0963	1.78	Middle_educ	0.0227	0.0699	0.32
Sweden	-2.1730	0.1819	-11.94	Sweden	-1.0741	0.1361	-7.89
Netherlands	-0.6957	0.1610	-4.32	Netherlands	-0.3834	0.1106	-3.47
Spain	-0.8410	0.1807	-4.66	Spain	-0.5396	0.1280	-4.22
Italy	-0.7539	0.1670	-4.52	Italy	-0.7655	0.1219	-6.28
France	-0.4058	0.1466	-2.77	France	-0.4466	0.1011	-4.42
Greece	-0.1581	0.1531	-1.03	Greece	0.2554	0.1012	2.52
Belgium	-0.1573	0.1528	-1.03	Belgium	-0.4945	0.1072	-4.62
Constant	-0.4885	0.1604	-3.05	Constant	-0.8320	0.1117	-7.45

	Cut-point 2				Cut-point 3		
	Coef.	*Std.err.*	*Z-value*		*Coef.*	*Std.err.*	*Z-value*
Age55to60	0.0542	0.0739	0.73	Age55to60	0.0179	0.1028	0.17
Age60to65	0.0385	0.0771	0.50	Age60to65	0.1487	0.1093	1.36
Age65to70	0.0441	0.0825	0.53	Age65to70	0.5718	0.1365	4.19
Age70to75	0.0205	0.0873	0.23	Age70to75	0.2133	0.1265	1.69
Age75to80	0.0454	0.1013	0.45	Age75to80	0.1514	0.1411	1.07
Age80to85	0.0920	0.1238	0.74	Age80to85	0.2077	0.1818	1.14
Female	0.0573	0.0479	1.20	Female	0.2623	0.0698	3.76
Low_educ	-0.1358	0.0708	-1.92	Low_educ	-0.1260	0.1045	-1.21
Middle_educ	0.0044	0.0640	0.07	Middle_educ	0.0106	0.0948	0.11
Sweden	-0.9335	0.1091	-8.55	Sweden	-1.1824	0.1506	-7.85
Netherlands	0.4935	0.1119	4.41	Netherlands	0.0897	0.1684	0.53
Spain	-0.3433	0.1156	-2.97	Spain	-0.6378	0.1665	-3.83
Italy	-0.4136	0.1082	-3.82	Italy	-0.3914	0.1623	-2.41
France	-0.0829	0.0966	-0.86	France	0.2708	0.1579	1.71
Greece	0.3860	0.1034	3.73	Greece	0.2689	0.1617	1.66
Belgium	0.2194	0.1026	2.14	Belgium	0.2502	0.1638	1.53
Constant	0.6122	0.1054	5.81	Constant	1.6800	0.1594	10.54

	Cut-point 4				Sigma		
	Coef.	*Std.err.*	*Z-value*		*Coef.*	*Std.err.*	*Z-value*
Age55to60	-0.0207	0.1813	-0.11	Constant	1.5465	0.0390	39.68
Age60to65	0.3942	0.2235	1.76				
Age65to70	0.3135	0.2157	1.45		Rho		
Age70to75	0.3875	0.2421	1.60		*Coef.*	*Std.err.*	*Z-value*
Age75to80	0.6833	0.3070	2.23	Constant	0.1261	0.0222	5.69
Age80to85	0.2985	0.3092	0.97				
Female	0.2957	0.1314	2.25				
Low_educ	0.0114	0.1888	0.06	Number of observations =	2,980		
Middle_educ	0.1549	0.1693	0.91	Log-likelihood =	-6,712.2039		
Sweden	-1.2571	0.2928	-4.29				
Netherlands	-0.6844	0.2857	-2.40				
Spain	-0.1892	0.3913	-0.48				
Italy	-0.8737	0.3032	-2.88				
France	-0.1679	0.2982	-0.56				
Greece	-0.0534	0.3043	-0.18				
Belgium	0.1263	0.3326	0.38				
Constant	2.8255	0.2952	9.57				

Table A2 - 2: HOPIT estimation results for the mobility domain.

	Latent own health		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.0286	0.1563	0.18
Age60to65	0.1326	0.1597	0.83
Age65to70	0.5314	0.1627	3.27
Age70to75	0.7446	0.1691	4.40
Age75to80	1.1324	0.1869	6.06
Age80to85	1.3712	0.2154	6.37
Female	0.2313	0.0924	2.50
Low_educ	1.0010	0.1542	6.49
Middle_educ	0.5639	0.1413	3.99
Sweden	-0.8779	0.2239	-3.92
Netherlands	-0.6166	0.2011	-3.07
Spain	-0.4883	0.2238	-2.18
Italy	-0.0313	0.2012	-0.16
France	-0.7681	0.1874	-4.10
Greece	-1.3376	0.2070	-6.46
Belgium	-0.3508	0.1867	-1.88
Constant	-3.3482	0.2254	-14.85

	Cut-point 1		
	Coef.	*Std.err.*	*Z-value*
Age55to60	-0.0706	0.1392	-0.51
Age60to65	-0.1339	0.1431	-0.94
Age65to70	0.1321	0.1418	0.93
Age70to75	-0.1116	0.1535	-0.73
Age75to80	0.1340	0.1651	0.81
Age80to85	0.3560	0.1827	1.95
Female	-0.1710	0.0830	-2.06
Low_educ	0.6776	0.1411	4.80
Middle_educ	0.4361	0.1330	3.28
Sweden	-1.2381	0.2260	-5.48
Netherlands	-0.2016	0.1844	-1.09
Spain	-0.1632	0.2048	-0.80
Italy	0.4832	0.1743	2.77
France	0.1654	0.1657	1.00
Greece	-0.2407	0.1842	-1.31
Belgium	-0.1196	0.1723	-0.69
Constant	-2.6794	0.2016	-13.29

	Cut-point 2		
	Coef.	*Std.err.*	*Z-value*
Age55to60	-0.1534	0.1021	-1.50
Age60to65	-0.1523	0.1056	-1.44
Age65to70	0.0113	0.1067	0.11
Age70to75	-0.1166	0.1146	-1.02
Age75to80	-0.0764	0.1299	-0.59
Age80to85	0.1681	0.1441	1.17
Female	-0.0392	0.0638	-0.61
Low_educ	0.3601	0.1034	3.48
Middle_educ	0.2231	0.0962	2.32
Sweden	-0.6341	0.1703	-3.72
Netherlands	-0.2091	0.1457	-1.44
Spain	-0.2113	0.1608	-1.31
Italy	0.3577	0.1372	2.61
France	0.0434	0.1276	0.34
Greece	-0.1197	0.1361	-0.88
Belgium	-0.0116	0.1323	-0.09
Constant	-1.5416	0.1464	-10.53

	Cut-point 3		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.0181	0.0736	0.25
Age60to65	-0.0576	0.0781	-0.74
Age65to70	0.1087	0.0810	1.34
Age70to75	0.0998	0.0858	1.16
Age75to80	0.2113	0.0994	2.12
Age80to85	0.1767	0.1186	1.49
Female	0.0666	0.0475	1.40
Low_educ	0.0609	0.0713	0.85
Middle_educ	0.0799	0.0636	1.26
Sweden	-0.4760	0.1188	-4.01
Netherlands	-0.0610	0.1044	-0.58
Spain	-0.0517	0.1196	-0.43
Italy	0.0791	0.1106	0.72
France	0.3273	0.0969	3.38
Greece	-0.1746	0.1023	-1.71
Belgium	0.1591	0.0995	1.60
Constant	-0.5493	0.1065	-5.16

	Cut-point 4		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.1059	0.0839	1.26
Age60to65	0.0427	0.0872	0.49
Age65to70	0.2752	0.0970	2.84
Age70to75	0.2044	0.1016	2.01
Age75to80	0.2744	0.1210	2.27
Age80to85	0.4687	0.1562	3.00
Female	0.0411	0.0558	0.74
Low_educ	-0.1685	0.0840	-2.01
Middle_educ	-0.1003	0.0746	-1.34
Sweden	-0.5678	0.1290	-4.40
Netherlands	-0.6468	0.1217	-5.32
Spain	0.1205	0.1520	0.79
Italy	0.0759	0.1383	0.55
France	0.4795	0.1354	3.54
Greece	-0.7037	0.1194	-5.90
Belgium	-0.2050	0.1224	-1.67
Constant	1.1231	0.1270	8.85

	Sigma		
	Coef.	*Std.err.*	*Z-value*
Constant	1.4572	0.0564	25.85

	Rho		
	Coef.	*Std.err.*	*Z-value*
Constant	0.0120	0.0239	0.50

Number of observations =	2,980
Log-likelihood =	-6,429.9267

Table A2 - 3: HOPIT estimation results for the sleep domain.

	Latent own health		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.0035	0.1241	0.03
Age60to65	0.1297	0.1289	1.01
Age65to70	0.1385	0.1346	1.03
Age70to75	0.1731	0.1426	1.21
Age75to80	0.4526	0.1616	2.80
Age80to85	0.4357	0.1884	2.31
Female	0.4732	0.0793	5.97
Low_educ	0.8577	0.1236	6.94
Middle_educ	0.3451	0.1130	3.05
Sweden	-2.8586	0.2908	-9.83
Netherlands	-0.4870	0.1717	-2.84
Spain	-0.7782	0.1984	-3.92
Italy	0.0789	0.1710	0.46
France	-0.0236	0.1532	-0.15
Greece	-0.7947	0.1673	-4.75
Belgium	-0.0342	0.1586	-0.22
Constant	-2.5818	0.1815	-14.22

	Cut-point 1		
	Coef.	*Std.err.*	*Z-value*
Age55to60	-0.0467	0.1243	-0.38
Age60to65	0.1526	0.1265	1.21
Age65to70	0.1385	0.1316	1.05
Age70to75	0.1021	0.1400	0.73
Age75to80	0.1148	0.1609	0.71
Age80to85	0.4518	0.1717	2.63
Female	-0.1652	0.0767	-2.15
Low_educ	0.7010	0.1219	5.75
Middle_educ	0.1543	0.1132	1.36
Sweden	-2.2533	0.2875	-7.84
Netherlands	-0.6184	0.1652	-3.74
Spain	-0.8109	0.1928	-4.21
Italy	-0.1974	0.1602	-1.23
France	-0.3761	0.1459	-2.58
Greece	-0.6165	0.1589	-3.88
Belgium	-0.5990	0.1555	-3.85
Constant	-2.1908	0.1711	-12.80

	Cut-point 2		
	Coef.	*Std.err.*	*Z-value*
Age55to60	-0.0579	0.0963	-0.60
Age60to65	0.1412	0.0975	1.45
Age65to70	0.0952	0.1027	0.93
Age70to75	0.0165	0.1093	0.15
Age75to80	0.1342	0.1252	1.07
Age80to85	0.1306	0.1440	0.91
Female	-0.0753	0.0602	-1.25
Low_educ	0.5410	0.0959	5.64
Middle_educ	0.2193	0.0898	2.44
Sweden	-1.7944	0.2560	-7.01
Netherlands	-0.3038	0.1327	-2.29
Spain	-0.5508	0.1531	-3.60
Italy	0.0332	0.1283	0.26
France	-0.2597	0.1179	-2.20
Greece	-0.0549	0.1209	-0.45
Belgium	-0.2233	0.1227	-1.82
Constant	-1.4573	0.1358	-10.73

	Cut-point 3		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.0164	0.0744	0.22
Age60to65	0.1937	0.0778	2.49
Age65to70	0.1401	0.0814	1.72
Age70to75	0.2173	0.0858	2.53
Age75to80	0.2525	0.1010	2.50
Age80to85	0.4027	0.1187	3.39
Female	-0.0033	0.0478	-0.07
Low_educ	0.4233	0.0724	5.84
Middle_educ	0.1756	0.0657	2.67
Sweden	-1.8101	0.2006	-9.02
Netherlands	0.0138	0.1031	0.13
Spain	-0.0981	0.1172	-0.84
Italy	-0.0351	0.1083	-0.32
France	0.3329	0.0948	3.51
Greece	-0.1495	0.1006	-1.49
Belgium	0.1538	0.0985	1.56
Constant	-0.6772	0.1061	-6.38

	Cut-point 4		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.0654	0.0797	0.82
Age60to65	0.0645	0.0841	0.77
Age65to70	0.1626	0.0906	1.79
Age70to75	0.1152	0.0973	1.18
Age75to80	0.3070	0.1191	2.58
Age80to85	0.5352	0.1605	3.33
Female	-0.0038	0.0537	-0.07
Low_educ	0.2772	0.0799	3.47
Middle_educ	0.0447	0.0685	0.65
Sweden	-1.5312	0.1205	-12.71
Netherlands	-0.5455	0.1106	-4.93
Spain	0.1524	0.1425	1.07
Italy	-0.0381	0.1231	-0.31
France	0.4820	0.1180	4.08
Greece	-0.4250	0.1095	-3.88
Belgium	-0.0159	0.1117	-0.14
Constant	0.7799	0.1150	6.78

	Sigma		
	Coef.	*Std.err.*	*Z-value*
Constant	1.3455	0.0432	31.12

	Rho		
	Coef.	*Std.err.*	*Z-value*
Constant	0.0778	0.0224	3.48

Number of observations =	2,980
Log-likelihood =	-7,104.6934

Table A2 - 4: HOPIT estimation results for the breath domain.

	Latent own health		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.0583	0.1322	0.44
Age60to65	0.0622	0.1397	0.45
Age65to70	0.3726	0.1443	2.58
Age70to75	0.4835	0.1514	3.19
Age75to80	0.5936	0.1726	3.44
Age80to85	0.4990	0.2054	2.43
Female	0.2750	0.0836	3.29
Low_educ	0.1639	0.1220	1.34
Middle_educ	0.0417	0.1111	0.38
Sweden	0.7565	0.1968	3.84
Netherlands	-0.3513	0.2022	-1.74
Spain	-0.9238	0.2402	-3.85
Italy	-0.0433	0.2062	-0.21
France	1.5338	0.1750	8.76
Greece	-0.3696	0.1943	-1.90
Belgium	0.8888	0.1803	4.93
Constant	-2.8626	0.2054	-13.94

	Cut-point 1		
	Coef.	*Std.err.*	*Z-value*
Age55to60	-0.0345	0.0915	-0.38
Age60to65	-0.1380	0.1002	-1.38
Age65to70	0.0179	0.1009	0.18
Age70to75	-0.2194	0.1116	-1.97
Age75to80	-0.0211	0.1233	-0.17
Age80to85	-0.3119	0.1525	-2.05
Female	0.1982	0.0605	3.28
Low_educ	-0.0122	0.0864	-0.14
Middle_educ	-0.0208	0.0791	-0.26
Sweden	-0.3281	0.1660	-1.98
Netherlands	-0.2098	0.1615	-1.30
Spain	-0.5232	0.1974	-2.65
Italy	0.3147	0.1528	2.06
France	1.4364	0.1288	11.15
Greece	-0.3582	0.1577	-2.27
Belgium	0.8029	0.1334	6.02
Constant	-1.8075	0.1466	-12.33

	Cut-point 2		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.0335	0.0756	0.44
Age60to65	0.0811	0.0794	1.02
Age65to70	0.1503	0.0838	1.79
Age70to75	0.0409	0.0891	0.46
Age75to80	0.2000	0.1034	1.93
Age80to85	0.0873	0.1221	0.72
Female	0.1255	0.0487	2.58
Low_educ	-0.1085	0.0712	-1.52
Middle_educ	-0.0712	0.0642	-1.11
Sweden	-0.0961	0.1224	-0.79
Netherlands	0.3821	0.1116	3.42
Spain	-0.3768	0.1434	-2.63
Italy	0.4860	0.1169	4.16
France	1.5177	0.1045	14.53
Greece	0.3634	0.1082	3.36
Belgium	1.0373	0.1058	9.81
Constant	-1.0199	0.1125	-9.06

	Cut-point 3		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.0390	0.0780	0.50
Age60to65	0.0914	0.0819	1.12
Age65to70	0.1519	0.0875	1.74
Age70to75	0.1027	0.0931	1.10
Age75to80	0.2332	0.1112	2.10
Age80to85	0.0452	0.1331	0.34
Female	0.1020	0.0510	2.00
Low_educ	0.0320	0.0757	0.42
Middle_educ	0.1627	0.0674	2.41
Sweden	-0.0459	0.1084	-0.42
Netherlands	0.1919	0.1044	1.84
Spain	-0.6167	0.1184	-5.21
Italy	0.2127	0.1123	1.89
France	1.3291	0.1129	11.78
Greece	0.1507	0.1009	1.49
Belgium	0.8059	0.1076	7.49
Constant	0.1350	0.1072	1.26

	Cut-point 4		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.2177	0.1330	1.64
Age60to65	0.3109	0.1474	2.11
Age65to70	0.4510	0.1668	2.70
Age70to75	0.1116	0.1567	0.71
Age75to80	0.1908	0.1877	1.02
Age80to85	0.1344	0.2367	0.57
Female	0.0726	0.0913	0.79
Low_educ	0.0323	0.1371	0.24
Middle_educ	0.1451	0.1184	1.23
Sweden	-0.2405	0.2010	-1.20
Netherlands	-0.3855	0.1898	-2.03
Spain	-0.2214	0.2184	-1.01
Italy	0.0129	0.2198	0.06
France	0.9096	0.2457	3.70
Greece	-0.1065	0.1967	-0.54
Belgium	0.3116	0.2136	1.46
Constant	1.5641	0.1976	7.92

	Sigma		
	Coef.	*Std.err.*	*Z-value*
Constant	1.4717	0.0519	28.36

	Rho		
	Coef.	*Std.err.*	*Z-value*
Constant	0.0205	0.0244	0.84

Number of observations =	2,980
Log-likelihood =	-6,344.8095

Table A2 - 5: HOPIT estimation results for the work disability domain.

	Latent own health		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.0132	0.1445	0.09
Age60to65	0.2515	0.1468	1.71
Age65to70	0.2882	0.1551	1.86
Age70to75	0.5475	0.1605	3.41
Age75to80	0.6349	0.1808	3.51
Age80to85	1.1283	0.2080	5.43
Female	0.1121	0.0878	1.28
Low_educ	1.0157	0.1377	7.37
Middle_educ	0.3203	0.1270	2.52
Sweden	-0.5462	0.2016	-2.71
Netherlands	-0.8034	0.1964	-4.09
Spain	-0.5757	0.2133	-2.70
Italy	-0.3156	0.1958	-1.61
France	-0.5611	0.1755	-3.20
Greece	-1.3101	0.1931	-6.78
Belgium	-0.4666	0.1812	-2.57
Constant	-2.7240	0.2046	-13.31

	Cut-point 1		
	Coef.	*Std.err.*	*Z-value*
Age55to60	-0.2319	0.1298	-1.79
Age60to65	-0.0436	0.1294	-0.34
Age65to70	-0.1392	0.1382	-1.01
Age70to75	-0.3995	0.1505	-2.65
Age75to80	-0.2560	0.1685	-1.52
Age80to85	-0.0792	0.1876	-0.42
Female	-0.1071	0.0803	-1.33
Low_educ	0.4982	0.1248	3.99
Middle_educ	0.1385	0.1178	1.18
Sweden	0.0624	0.1830	0.34
Netherlands	-0.3289	0.1847	-1.78
Spain	-0.2935	0.2000	-1.47
Italy	0.0458	0.1792	0.26
France	-0.0718	0.1622	-0.44
Greece	0.0663	0.1701	0.39
Belgium	-0.5610	0.1769	-3.17
Constant	-2.2745	0.1824	-12.47

	Cut-point 2		
	Coef.	*Std.err.*	*Z-value*
Age55to60	-0.1898	0.0952	-1.99
Age60to65	-0.0149	0.0952	-0.16
Age65to70	-0.0129	0.1016	-0.13
Age70to75	-0.1094	0.1087	-1.01
Age75to80	-0.1019	0.1247	-0.82
Age80to85	-0.0698	0.1457	-0.48
Female	-0.0282	0.0597	-0.47
Low_educ	0.2463	0.0907	2.71
Middle_educ	0.0063	0.0839	0.08
Sweden	-0.3274	0.1428	-2.29
Netherlands	-0.1685	0.1345	-1.25
Spain	-0.4363	0.1547	-2.82
Italy	0.1675	0.1322	1.27
France	-0.0728	0.1196	-0.61
Greece	-0.0796	0.1245	-0.64
Belgium	-0.2433	0.1277	-1.91
Constant	-1.2201	0.1325	-9.21

	Cut-point 3		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.0688	0.0726	0.95
Age60to65	0.1488	0.0759	1.96
Age65to70	0.2426	0.0800	3.03
Age70to75	0.1177	0.0856	1.37
Age75to80	0.1521	0.0998	1.52
Age80to85	0.0927	0.1185	0.78
Female	0.0900	0.0468	1.92
Low_educ	0.1635	0.0699	2.34
Middle_educ	0.0009	0.0629	0.01
Sweden	-0.8473	0.1144	-7.40
Netherlands	-0.4324	0.1025	-4.22
Spain	-0.5780	0.1177	-4.91
Italy	-0.1207	0.1065	-1.13
France	-0.0367	0.0936	-0.39
Greece	-0.4260	0.0991	-4.30
Belgium	-0.3000	0.0977	-3.07
Constant	-0.2097	0.1036	-2.02

	Cut-point 4		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.2621	0.0828	3.17
Age60to65	0.3334	0.0889	3.75
Age65to70	0.3977	0.0966	4.12
Age70to75	0.2452	0.0988	2.48
Age75to80	0.5646	0.1290	4.38
Age80to85	0.2806	0.1404	2.00
Female	0.0416	0.0560	0.74
Low_educ	-0.0630	0.0843	-0.75
Middle_educ	-0.1022	0.0747	-1.37
Sweden	-0.9448	0.1467	-6.44
Netherlands	-1.0918	0.1396	-7.82
Spain	-0.3646	0.1650	-2.21
Italy	-0.5133	0.1514	-3.39
France	0.0230	0.1492	0.15
Greece	-0.8548	0.1390	-6.15
Belgium	-0.7424	0.1389	-5.34
Constant	1.3828	0.1438	9.62

	Sigma		
	Coef.	*Std.err.*	*Z-value*
Constant	1.5303	0.0518	29.56

	Rho		
	Coef.	*Std.err.*	*Z-value*
Constant	0.0778	0.0232	3.35

Number of observations =	2,980
Log-likelihood =	-6,834.7501

Table A2 - 6: HOPIT estimation results for the depress domain.

	Latent own health		
	Coef.	*Std.err.*	*Z-value*
Age55to60	-0.3463	0.1228	-2.82
Age60to65	-0.2045	0.1256	-1.63
Age65to70	-0.1352	0.1321	-1.02
Age70to75	-0.1527	0.1399	-1.09
Age75to80	-0.1520	0.1622	-0.94
Age80to85	0.4286	0.1763	2.43
Female	0.1037	0.0779	1.33
Low_educ	0.8827	0.1229	7.18
Middle_educ	0.2944	0.1155	2.55
Sweden	-0.3737	0.1831	-2.04
Netherlands	-0.9433	0.1884	-5.01
Spain	-0.8327	0.1999	-4.16
Italy	0.0750	0.1672	0.45
France	-0.5711	0.1582	-3.61
Greece	-0.3204	0.1591	-2.01
Belgium	-0.4054	0.1620	-2.50
Constant	-2.4444	0.1787	-13.68

	Cut-point 1		
	Coef.	*Std.err.*	*Z-value*
Age55to60	-0.3240	0.1221	-2.65
Age60to65	-0.1328	0.1241	-1.07
Age65to70	-0.0121	0.1291	-0.09
Age70to75	-0.1952	0.1379	-1.42
Age75to80	-0.3232	0.1622	-1.99
Age80to85	0.3380	0.1664	2.03
Female	-0.3018	0.0773	-3.91
Low_educ	0.6772	0.1204	5.62
Middle_educ	0.2155	0.1146	1.88
Sweden	-0.9441	0.1886	-5.01
Netherlands	-0.8160	0.1816	-4.49
Spain	-0.8164	0.1952	-4.18
Italy	-0.0885	0.1592	-0.56
France	-0.5040	0.1513	-3.33
Greece	-0.4804	0.1553	-3.09
Belgium	-0.4500	0.1557	-2.89
Constant	-2.0443	0.1689	-12.10

	Cut-point 2		
	Coef.	*Std.err.*	*Z-value*
Age55to60	-0.1630	0.1090	-1.50
Age60to65	-0.0091	0.1102	-0.08
Age65to70	0.1044	0.1145	0.91
Age70to75	-0.0236	0.1227	-0.19
Age75to80	-0.0794	0.1437	-0.55
Age80to85	0.4379	0.1504	2.91
Female	-0.2087	0.0681	-3.06
Low_educ	0.5582	0.1075	5.19
Middle_educ	0.1884	0.1029	1.83
Sweden	-0.5106	0.1632	-3.13
Netherlands	-0.4886	0.1646	-2.97
Spain	-0.6543	0.1757	-3.72
Italy	0.1031	0.1422	0.72
France	-0.3944	0.1367	-2.88
Greece	-0.3759	0.1382	-2.72
Belgium	-0.1947	0.1391	-1.40
Constant	-1.6040	0.1505	-10.66

	Cut-point 3		
	Coef.	*Std.err.*	*Z-value*
Age55to60	-0.0855	0.0860	-0.99
Age60to65	0.0169	0.0881	0.19
Age65to70	0.0903	0.0920	0.98
Age70to75	0.0968	0.0979	0.99
Age75to80	-0.0741	0.1171	-0.63
Age80to85	0.3747	0.1287	2.91
Female	-0.0780	0.0545	-1.43
Low_educ	0.4091	0.0837	4.89
Middle_educ	0.1600	0.0781	2.05
Sweden	-0.5096	0.1351	-3.77
Netherlands	-0.3720	0.1247	-2.98
Spain	-0.4837	0.1399	-3.46
Italy	-0.0435	0.1209	-0.36
France	-0.1513	0.1096	-1.38
Greece	-0.3370	0.1135	-2.97
Belgium	-0.1118	0.1127	-0.99
Constant	-0.9723	0.1212	-8.02

	Cut-point 4		
	Coef.	*Std.err.*	*Z-value*
Age55to60	0.1077	0.0714	1.51
Age60to65	0.2087	0.0749	2.79
Age65to70	0.3023	0.0799	3.79
Age70to75	0.3086	0.0859	3.59
Age75to80	0.4068	0.1003	4.06
Age80to85	0.4982	0.1218	4.09
Female	0.0928	0.0467	1.99
Low_educ	0.1092	0.0698	1.56
Middle_educ	0.0718	0.0622	1.15
Sweden	0.1096	0.1108	0.99
Netherlands	-0.6236	0.1038	-6.01
Spain	-0.2657	0.1175	-2.26
Italy	-0.2984	0.1084	-2.75
France	0.2720	0.0977	2.78
Greece	-0.4485	0.0992	-4.52
Belgium	-0.1461	0.0990	-1.48
Constant	0.0026	0.1036	0.02

	Sigma		
	Coef.	*Std.err.*	*Z-value*
Constant	0.9896	0.0414	23.88

	Rho		
	Coef.	*Std.err.*	*Z-value*
Constant	0.0116	0.0240	0.48

Number of observations = 2,980

Log-likelihood = -6,577.0742

Appendix 3 The simulation exercise

In the simulation exercise conducted in the empirical part of this thesis we draw random numbers for the two error terms of the bivariate probit model, i.e., u_i and v_i, 1,000 times for each individual. These random numbers are drawn from a bivariate normal distribution. When we have n random variables, we can describe their multivariate normal probability density function using matrix notation as in Equation A3-1 (Heij et al., 2004).

$$f(v) = \frac{1}{(2\pi)^{n/2}\left(det(\Sigma)\right)^{1/2}} e^{-\frac{1}{2}(v-\mu)'\Sigma^{-1}(v-\mu)} \quad \text{(A3-1)}$$

where v denotes the n variables, μ is a $n \times 1$ vector with the means of the random variables, Σ is the covariance matrix. We have two random variables and therefore consider the bivariate normal distribution. In that case v denotes the two variables, u and v, μ is a 2×1 vector with the means of these two error terms. The covariance matrix Σ now is a 2×2 matrix with the variances and covariances between the two error terms for which we draw random numbers. The covariance matrix is defined as in Equation A3-2.

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma \cdot \rho \\ \sigma \cdot \rho & 1 \end{bmatrix} \quad \text{(A3-2)}$$

The means of the two error terms were assumed to be zero. Besides, we have obtained estimates for the variance of u and the correlation between the two error terms. These numerical values are used in the specification of the bivariate normal distribution for the two error terms.

Then we draw random numbers for u and v from this bivariate normal distribution for each of the respondents in the first wave. We do this 1,000 times. For each respondent and for each round of the simulation we generate a variable h_i^s that is defined as in Equation A3-3.

$$h_i^s = x_i'\hat{\beta} + u_i^s \quad \text{(A3-3)}$$

where u_i^s is the drawn random number[54], $\hat{\beta}$ is the vector of estimated coefficients from the latent health index function for the respondent as described in Equation 19 in section 8.1.1 and x_i is a vector of individual characteristics. This variable h_i^s thus equals the sum of the predicted corrected latent health level resulting from the bivariate probit model and the drawn random number for the error term of the latent health index function for the respondent. In addition we generate a variable labelled c_i that equals 1 if the vignette person is evaluated correctly and 0 otherwise. This variable is described as in Equation A3-4.

$$c_i^s = \begin{cases} & vig = 1 \cap v_i^s < \tau_i^1 \\ & vig = 2 \cap \tau_i^1 \leq v_i^s < \tau_i^2 \\ 1 \ if & vig = 3 \cap \tau_i^2 \leq v_i^s < \tau_i^3 \\ & vig = 4 \cap \tau_i^3 \leq v_i^s < \tau_i^4 \\ & vig = 5 \cap v_i^s \geq \tau_i^4 \\ 0 \ if & otherwise \end{cases} \quad \text{(A3-4)}$$

where $vig = 1$ means that the observed vignette evaluation is 1, i.e., the respondent has evaluated the vignette person to have no limitations. We can then count the number of times that a respondent evaluates the vignette person correctly over all 1,000 draws for the error terms, indicated by N_i. This is described in Equation A3-5.

$$N_i = \sum_{s=1}^{1000} c_i^s \quad \text{(A3-5)}$$

[54] The superscript s indicates the round of the simulation. It reflects the fact that the variables differ across the various rounds of the simulation procedure.

In constructing our corrected latent health variable we only want to consider those draws of the error terms that are meaningful, i.e. for which the observed vignette evaluation is possible. To take this into account in the construction of the corrected latent health variable we use the variable N described in Equation A3-5. We will use the average predicted corrected latent health level for a respondent as our corrected health status variable. The predicted corrected latent health level of the respondent in each round of the simulation is already defined by the variable h_i^S. We generate a variable $hpred$ defined as in Equation A3-6.

$$hpred_i^S = c_i^S \cdot h_i^S \tag{A3-6}$$

We can then calculate the average predicted corrected latent health level as in Equation A3-7.

$$hpred_i = \frac{\sum_{s=1}^{1000} hpred_i^S}{N_i} \tag{A3-7}$$

This average predicted corrected latent health level is corrected for reporting heterogeneity. Therefore we can use this variable in order to test for the extent to which the vignettes method brings the self-reported health measures closer to the true health status.

Appendix 4 Estimation results for the test regressions

Table A4 - 1: Test regressions. Labour market states as outcome variables (pain domain).

	Working		Retired		Disabled	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	2.112 (0.101)***	2.118 (0.102)***	1.888 (0.083)***	1.892 (0.083)***	1.765 (0.185)***	1.777 (0.187)***
w1_self_report	0.014 (0.045)	0.038 (0.045)	-0.034 (0.035)	-0.049 (0.035)	0.291 (0.052)***	0.304 (0.053)***
w1_age55to60	-0.600 (0.096)***	-0.595 (0.098)***	0.867 (0.111)***	0.868 (0.116)***	-0.168 (0.155)	-0.150 (0.156)
w1_age60to65	-1.246 (0.124)***	-1.241 (0.126)***	1.540 (0.125)***	1.547 (0.130)***	-0.499 (0.194)***	-0.486 (0.196)**
w1_age65to70	-1.567 (0.174)***	-1.534 (0.177)***	1.655 (0.143)***	1.653 (0.154)***	-0.628 (0.212)***	-0.593 (0.211)***
w1_age70to75	-1.656 (0.251)***	-1.592 (0.254)***	1.551 (0.152)***	1.541 (0.171)***	-0.577 (0.204)***	-0.530 (0.203)***
w1_age75to80			1.676 (0.168)***	1.667 (0.197)***	-0.465 (0.231)**	-0.398 (0.233)*
w1_age80to85			1.329 (0.184)***	1.390 (0.212)***	-0.736 (0.327)**	-0.669 (0.339)**
w1_female			-0.226 (0.075)***	-0.239 (0.118)**		
w1_married			0.168 (0.079)**	0.175 (0.079)**	-0.373 (0.120)***	-0.400 (0.122)***
w1_lincome	0.036 (0.016)**	0.035 (0.016)**	0.059 (0.012)***	0.059 (0.012)***		
w1_low_educ	-0.472 (0.119)***	-0.450 (0.124)***	0.322 (0.099)***	0.325 (0.113)***	0.410 (0.186)**	0.454 (0.203)**
w1_middle_educ	-0.118 (0.100)	-0.091 (0.100)	0.185 (0.088)**	0.174 (0.092)*	0.317 (0.175)*	0.346 (0.179)*
w1_sweden	0.711 (0.198)***	0.346 (0.304)	-0.101 (0.154)	-0.032 (0.359)	0.032 (0.277)	-0.184 (0.461)
w1_netherlands	0.207 (0.188)	0.102 (0.203)	-0.435 (0.160)***	-0.421 (0.189)**	0.371 (0.242)	0.301 (0.266)
w1_spain	0.253 (0.209)	0.121 (0.231)	-0.617 (0.167)***	-0.605 (0.209)***	0.420 (0.257)	0.334 (0.303)
w1_italy	0.076 (0.201)	-0.009 (0.217)	-0.191 (0.159)	-0.186 (0.194)	0.026 (0.273)	-0.045 (0.312)
w1_france	0.410 (0.172)**	0.374 (0.180)**	-0.144 (0.137)	-0.153 (0.152)	-0.561 (0.251)**	-0.559 (0.264)**
w1_greece	0.752 (0.173)***	0.741 (0.176)***	-0.642 (0.139)***	-0.654 (0.143)***	-0.207 (0.253)	-0.225 (0.261)
w1_belgium	0.163 (0.184)	0.133 (0.186)	-0.332 (0.145)**	-0.349 (0.148)**	0.219 (0.235)	0.213 (0.240)
w1_corr_health		-0.180 (0.109)*		0.039 (0.150)		-0.108 (0.151)
constant	-1.382 (0.243)***	-1.448 (0.246)***	-2.133 (0.226)***	-2.098 (0.240)***	-2.582 (0.291)***	-2.619 (0.292)***
N	2,568	2,537	2,940	2,903	2,940	2,903

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., working, retired or disabled); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 2: Test regressions. Labour market states as outcome variables (mobility domain).

	Working		Retired		Disabled	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	2.097 (0.100)***	2.109 (0.101)***	1.889 (0.083)***	1.893 (0.084)***	1.685 (0.184)***	1.672 (0.186)***
w1_self_report	-0.058 (0.058)	-0.068 (0.060)	-0.069 (0.037)*	-0.068 (0.037)*	0.331 (0.054)***	0.343 (0.056)***
w1_age55to60	-0.594 (0.096)***	-0.590 (0.097)***	0.861 (0.110)***	0.854 (0.111)***	-0.174 (0.154)	-0.155 (0.156)
w1_age60to65	-1.250 (0.123)***	-1.270 (0.129)***	1.551 (0.124)***	1.615 (0.137)***	-0.553 (0.195)***	-0.452 (0.202)**
w1_age65to70	-1.595 (0.177)***	-1.658 (0.226)***	1.665 (0.142)***	1.940 (0.259)***	-0.698 (0.225)***	-0.243 (0.280)
w1_age70to75	-1.645 (0.251)***	-1.730 (0.315)***	1.572 (0.153)***	1.926 (0.340)***	-0.772 (0.216)***	-0.139 (0.321)
w1_age75to80			1.705 (0.168)***	2.259 (0.499)***	-0.661 (0.242)***	0.324 (0.474)
w1_age80to85			1.364 (0.186)***	2.084 (0.586)***	-0.883 (0.319)***	0.334 (0.579)
w1_female			-0.237 (0.075)***	-0.111 (0.117)		
w1_married			0.162 (0.079)**	0.169 (0.079)**	-0.371 (0.121)***	-0.417 (0.122)***
w1_lincome	0.037 (0.016)**	0.041 (0.016)**	0.057 (0.012)***	0.056 (0.012)***		
w1_low_educ	-0.450 (0.118)***	-0.598 (0.313)*	0.329 (0.099)***	0.811 (0.422)*	0.358 (0.187)*	1.227 (0.440)***
w1_middle_educ	-0.105 (0.099)	-0.177 (0.185)	0.183 (0.088)**	0.448 (0.248)*	0.275 (0.174)	0.760 (0.268)***
w1_sweden	0.696 (0.199)***	0.826 (0.323)**	-0.076 (0.154)	-0.504 (0.393)	-0.150 (0.282)	-0.886 (0.474)*
w1_netherlands	0.196 (0.188)	0.287 (0.258)	-0.427 (0.160)***	-0.726 (0.292)**	0.337 (0.248)	-0.186 (0.347)
w1_spain	0.233 (0.209)	0.297 (0.255)	-0.629 (0.167)***	-0.847 (0.261)***	0.433 (0.261)*	0.045 (0.340)
w1_italy	0.049 (0.203)	0.084 (0.207)	-0.210 (0.159)	-0.225 (0.161)	0.016 (0.279)	0.023 (0.282)
w1_france	0.386 (0.175)**	0.532 (0.283)*	-0.188 (0.138)	-0.570 (0.344)*	-0.407 (0.254)	-1.067 (0.418)**
w1_greece	0.713 (0.176)***	0.904 (0.427)**	-0.676 (0.140)***	-1.332 (0.560)**	-0.104 (0.256)	-1.254 (0.600)**
w1_belgium	0.148 (0.187)	0.171 (0.215)	-0.334 (0.145)**	-0.505 (0.201)**	0.283 (0.240)	0.005 (0.290)
w1_corr_health		0.131 (0.289)		-0.496 (0.406)		-0.866 (0.388)**
constant	-1.257 (0.244)***	-0.876 (0.927)	-2.054 (0.227)***	-3.718 (1.381)***	-2.464 (0.277)***	-5.264 (1.263)***
N	2,574	2,540	2,946	2,905	2,946	2,905

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., working, retired or disabled); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 3: Test regressions. Labour market states as outcome variables (sleep domain).

	Working		Retired		Disabled	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	2.107 (0.100)***	2.107 (0.101)***	1.886 (0.083)***	1.888 (0.083)***	1.800 (0.175)***	1.809 (0.177)***
w1_self_report	0.016 (0.043)	0.025 (0.044)	-0.011 (0.034)	-0.018 (0.034)	0.239 (0.050)***	0.263 (0.053)***
w1_age55to60	-0.600 (0.096)***	-0.625 (0.097)***	0.867 (0.111)***	0.881 (0.112)***	-0.165 (0.151)	-0.168 (0.152)
w1_age60to65	-1.254 (0.124)***	-1.261 (0.126)***	1.545 (0.125)***	1.540 (0.129)***	-0.518 (0.195)***	-0.475 (0.197)**
w1_age65to70	-1.570 (0.173)***	-1.573 (0.175)***	1.652 (0.143)***	1.658 (0.148)***	-0.559 (0.214)***	-0.507 (0.214)**
w1_age70to75	-1.657 (0.250)***	-1.652 (0.253)***	1.550 (0.152)***	1.541 (0.157)***	-0.549 (0.206)***	-0.484 (0.204)**
w1_age75to80			1.672 (0.167)***	1.640 (0.197)***	-0.420 (0.231)*	-0.234 (0.231)
w1_age80to85			1.325 (0.183)***	1.335 (0.212)***	-0.676 (0.298)**	-0.461 (0.292)
w1_female			-0.234 (0.076)***	-0.269 (0.124)**		
w1_married			0.169 (0.079)**	0.187 (0.079)**	-0.371 (0.118)***	-0.410 (0.121)***
w1_lincome	0.037 (0.016)**	0.040 (0.016)**	0.059 (0.012)***	0.059 (0.012)***		
w1_low_educ	-0.473 (0.119)***	-0.399 (0.182)**	0.316 (0.099)***	0.207 (0.214)	0.399 (0.183)**	0.774 (0.258)***
w1_middle_educ	-0.123 (0.099)	-0.075 (0.110)	0.186 (0.088)**	0.135 (0.116)	0.312 (0.171)*	0.466 (0.187)**
w1_sweden	0.700 (0.199)***	0.424 (0.510)	-0.078 (0.154)	0.184 (0.655)	0.044 (0.264)	-1.120 (0.634)*
w1_netherlands	0.202 (0.188)	0.162 (0.205)	-0.424 (0.161)***	-0.396 (0.191)**	0.312 (0.236)	0.120 (0.256)
w1_spain	0.252 (0.209)	0.193 (0.249)	-0.612 (0.167)***	-0.535 (0.239)**	0.480 (0.250)*	0.164 (0.307)
w1_italy	0.070 (0.202)	0.098 (0.205)	-0.192 (0.159)	-0.200 (0.162)	0.005 (0.270)	0.048 (0.271)
w1_france	0.405 (0.173)**	0.434 (0.175)**	-0.135 (0.138)	-0.152 (0.139)	-0.622 (0.252)**	-0.606 (0.249)**
w1_greece	0.754 (0.175)***	0.693 (0.220)***	-0.634 (0.140)***	-0.559 (0.224)**	-0.188 (0.248)	-0.526 (0.294)*
w1_belgium	0.158 (0.184)	0.154 (0.185)	-0.327 (0.146)**	-0.346 (0.147)**	0.166 (0.234)	0.153 (0.238)
w1_corr_health		-0.105 (0.166)		0.096 (0.224)		-0.417 (0.191)**
constant	-1.382 (0.242)***	-1.676 (0.435)***	-2.188 (0.225)***	-1.934 (0.622)***	-2.418 (0.263)***	-3.429 (0.522)***
N	2,572	2,538	2,944	2,905	2,944	2,905

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., working, retired or disabled); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 4: Test regressions. Labour market states as outcome variables (breath domain).

	Working		Retired		Disabled	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	2.106 (0.101)***	2.105 (0.102)***	1.896 (0.083)***	1.908 (0.084)***	1.902 (0.173)***	1.912 (0.173)***
w1_self_report	0.026 (0.057)	0.013 (0.056)	-0.001 (0.042)	-0.007 (0.043)	0.088 (0.063)	0.094 (0.063)
w1_age55to60	-0.616 (0.096)***	-0.628 (0.098)***	0.881 (0.111)***	0.888 (0.114)***	-0.119 (0.147)	-0.137 (0.150)
w1_age60to65	-1.267 (0.123)***	-1.272 (0.125)***	1.556 (0.125)***	1.564 (0.128)***	-0.456 (0.189)**	-0.436 (0.189)**
w1_age65to70	-1.614 (0.177)***	-1.657 (0.189)***	1.666 (0.143)***	1.699 (0.187)***	-0.550 (0.208)***	-0.461 (0.226)**
w1_age70to75	-1.681 (0.243)***	-1.734 (0.279)***	1.559 (0.152)***	1.589 (0.217)***	-0.530 (0.202)***	-0.421 (0.222)*
w1_age75to80			1.660 (0.165)***	1.728 (0.265)***	-0.360 (0.228)	-0.218 (0.277)
w1_age80to85			1.339 (0.183)***	1.415 (0.253)***	-0.586 (0.286)**	-0.457 (0.307)
w1_female			-0.246 (0.075)***	-0.222 (0.112)**		
w1_married			0.168 (0.078)**	0.161 (0.079)**	-0.356 (0.114)***	-0.372 (0.117)***
w1_lincome	0.035 (0.016)**	0.040 (0.016)**	0.056 (0.012)***	0.054 (0.012)***		
w1_low_educ	-0.465 (0.118)***	-0.502 (0.128)***	0.308 (0.099)***	0.321 (0.114)***	0.447 (0.178)**	0.517 (0.193)***
w1_middle_educ	-0.109 (0.099)	-0.107 (0.100)	0.167 (0.088)*	0.180 (0.090)**	0.360 (0.171)**	0.364 (0.174)**
w1_sweden	0.685 (0.205)***	0.619 (0.294)**	-0.084 (0.156)	0.004 (0.292)	-0.211 (0.277)	-0.018 (0.353)
w1_netherlands	0.228 (0.188)	0.278 (0.214)	-0.422 (0.162)***	-0.461 (0.192)**	0.302 (0.239)	0.216 (0.267)
w1_spain	0.256 (0.209)	0.388 (0.340)	-0.628 (0.169)***	-0.744 (0.334)**	0.436 (0.252)*	0.195 (0.404)
w1_italy	0.084 (0.203)	0.104 (0.205)	-0.214 (0.161)	-0.187 (0.162)	0.061 (0.265)	-0.018 (0.270)
w1_france	0.423 (0.173)**	0.256 (0.465)	-0.160 (0.139)	0.021 (0.516)	-0.512 (0.243)**	-0.154 (0.530)
w1_greece	0.757 (0.174)***	0.819 (0.210)***	-0.643 (0.141)***	-0.686 (0.185)***	-0.244 (0.251)	-0.355 (0.286)
w1_belgium	0.189 (0.184)	0.052 (0.306)	-0.345 (0.146)**	-0.214 (0.330)	0.194 (0.236)	0.409 (0.370)
w1_corr_health		0.122 (0.281)		-0.124 (0.324)		-0.244 (0.328)
constant	-1.371 (0.238)***	-1.076 (0.749)	-2.163 (0.228)***	-2.492 (0.968)***	-2.069 (0.272)***	-2.717 (0.913)***
N	2,569	2,536	2,941	2,899	2,941	2,899

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., working, retired or disabled); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 5: Test regressions. Labour market states as outcome variables (depress domain).

	Working		Retired		Disabled	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	2.115 (0.100)***	2.123 (0.101)***	1.880 (0.083)***	1.905 (0.084)***	1.802 (0.173)***	1.806 (0.173)***
w1_self_report	0.017 (0.042)	0.020 (0.043)	-0.043 (0.034)	-0.047 (0.034)	0.211 (0.051)***	0.227 (0.051)***
w1_age55to60	-0.625 (0.096)***	-1.011 (0.199)***	0.878 (0.111)***	1.248 (0.217)***	-0.064 (0.145)	-0.312 (0.291)
w1_age60to65	-1.268 (0.124)***	-1.507 (0.161)***	1.555 (0.125)***	1.763 (0.165)***	-0.391 (0.186)**	-0.510 (0.225)**
w1_age65to70	-1.609 (0.178)***	-1.775 (0.204)***	1.669 (0.142)***	1.810 (0.165)***	-0.502 (0.208)**	-0.583 (0.244)**
w1_age70to75	-1.663 (0.250)***	-1.850 (0.266)***	1.567 (0.151)***	1.702 (0.176)***	-0.487 (0.202)**	-0.582 (0.246)**
w1_age75to80			1.684 (0.164)***	1.808 (0.184)***	-0.335 (0.229)	-0.425 (0.254)*
w1_age80to85			1.343 (0.181)***	0.894 (0.292)***	-0.575 (0.303)*	-0.316 (0.405)
w1_female			-0.226 (0.075)***	-0.342 (0.094)***		
w1_married			0.165 (0.078)**	0.147 (0.079)*	-0.313 (0.116)***	-0.322 (0.118)***
w1_lincome	0.034 (0.016)**	0.028 (0.016)*	0.057 (0.012)***	0.055 (0.012)***		
w1_low_educ	-0.474 (0.119)***	0.478 (0.453)	0.307 (0.099)***	-0.618 (0.487)	0.436 (0.184)**	1.018 (0.658)
w1_middle_educ	-0.115 (0.099)	0.201 (0.174)	0.170 (0.088)*	-0.131 (0.185)	0.360 (0.173)**	0.536 (0.275)*
w1_sweden	0.699 (0.200)***	0.272 (0.273)	-0.091 (0.155)	0.294 (0.255)	-0.187 (0.268)	-0.429 (0.373)
w1_netherlands	0.208 (0.189)	-0.843 (0.511)*	-0.417 (0.162)***	0.593 (0.528)	0.348 (0.235)	-0.258 (0.684)
w1_spain	0.261 (0.210)	-0.662 (0.474)	-0.620 (0.167)***	0.244 (0.479)	0.407 (0.253)	-0.141 (0.627)
w1_italy	0.068 (0.203)	0.177 (0.207)	-0.192 (0.159)	-0.262 (0.166)	-0.032 (0.270)	-0.077 (0.274)
w1_france	0.432 (0.175)**	-0.181 (0.333)	-0.158 (0.139)	0.424 (0.339)	-0.497 (0.244)**	-0.881 (0.477)*
w1_greece	0.758 (0.175)***	0.390 (0.243)	-0.629 (0.141)***	-0.298 (0.222)	-0.312 (0.254)	-0.547 (0.337)
w1_belgium	0.195 (0.186)	-0.258 (0.274)	-0.341 (0.146)**	0.072 (0.268)	0.206 (0.235)	-0.063 (0.363)
w1_corr_health		-1.101 (0.496)**		1.064 (0.537)**		-0.635 (0.686)
constant	-1.365 (0.239)***	-3.932 (1.168)***	-2.109 (0.228)***	0.525 (1.330)	-2.400 (0.276)***	-3.923 (1.662)**
N	2,572	2,538	2,946	2,904	2,946	2,904

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., working, retired or disabled); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 6: Test regressions. Labour market states as outcome variables (work disability domain).

	Working		Retired		Disabled	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	2.082 (0.102)***	2.087 (0.102)***	1.876 (0.083)***	1.885 (0.083)***	1.499 (0.180)***	1.494 (0.182)***
w1_self_report	-0.057 (0.046)	-0.057 (0.046)	0.011 (0.034)	0.001 (0.034)	0.320 (0.052)***	0.326 (0.053)***
w1_age55to60	-0.636 (0.097)***	-0.623 (0.097)***	0.889 (0.111)***	0.876 (0.112)***	-0.147 (0.153)	-0.143 (0.153)
w1_age60to65	-1.279 (0.125)***	-1.135 (0.136)***	1.559 (0.125)***	1.504 (0.141)***	-0.503 (0.188)***	-0.456 (0.211)**
w1_age65to70	-1.599 (0.174)***	-1.423 (0.181)***	1.672 (0.143)***	1.629 (0.163)***	-0.664 (0.226)***	-0.605 (0.248)**
w1_age70to75	-1.668 (0.252)***	-1.336 (0.271)***	1.575 (0.153)***	1.473 (0.200)***	-0.738 (0.209)***	-0.636 (0.274)**
w1_age75to80			1.687 (0.166)***	1.558 (0.229)***	-0.569 (0.235)**	-0.444 (0.293)
w1_age80to85			1.348 (0.184)***	1.232 (0.331)***	-0.809 (0.333)**	-0.611 (0.474)
w1_female			-0.228 (0.075)***	-0.244 (0.080)***		
w1_married			0.180 (0.078)**	0.184 (0.079)**	-0.362 (0.120)***	-0.367 (0.121)***
w1_lincome	0.040 (0.016)**	0.036 (0.016)**	0.059 (0.012)***	0.057 (0.012)***		
w1_low_educ	-0.471 (0.119)***	0.131 (0.263)	0.295 (0.100)***	0.147 (0.263)	0.366 (0.184)**	0.533 (0.406)
w1_middle_educ	-0.109 (0.100)	0.085 (0.124)	0.176 (0.088)**	0.124 (0.116)	0.297 (0.170)*	0.356 (0.207)*
w1_sweden	0.709 (0.201)***	0.376 (0.236)	-0.101 (0.155)	-0.041 (0.211)	-0.113 (0.281)	-0.176 (0.329)
w1_netherlands	0.205 (0.190)	-0.285 (0.265)	-0.424 (0.160)***	-0.336 (0.256)	0.383 (0.243)	0.275 (0.319)
w1_spain	0.266 (0.211)	-0.086 (0.245)	-0.611 (0.168)***	-0.537 (0.222)**	0.357 (0.257)	0.288 (0.315)
w1_italy	0.072 (0.205)	-0.102 (0.217)	-0.187 (0.160)	-0.145 (0.181)	-0.017 (0.272)	-0.034 (0.282)
w1_france	0.412 (0.174)**	0.084 (0.214)	-0.137 (0.139)	-0.091 (0.199)	-0.650 (0.257)**	-0.666 (0.303)**
w1_greece	0.732 (0.176)***	-0.060 (0.349)	-0.624 (0.142)***	-0.465 (0.351)	-0.142 (0.256)	-0.335 (0.497)
w1_belgium	0.158 (0.184)	-0.125 (0.210)	-0.346 (0.146)**	-0.306 (0.193)	0.199 (0.237)	0.154 (0.278)
w1_corr_health		-0.593 (0.226)***		0.138 (0.236)		-0.172 (0.330)
constant	-1.260 (0.239)***	-2.808 (0.644)***	-2.242 (0.226)***	-1.807 (0.684)***	-2.473 (0.276)***	-2.969 (0.952)***
N	2,562	2,524	2,936	2,885	2,936	2,885

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., working, retired or disabled); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 7: Test regressions. Various types of health care usage as outcome variables (pain domain).

	In hospital?		Had outpatient surgery?		Out-of-pocket drugs expenses	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.495 (0.087)***	0.493 (0.088)***	0.659 (0.121)***	0.683 (0.122)***	0.265 (0.085)***	0.266 (0.085)***
w1_self_report	0.198 (0.031)***	0.204 (0.031)***	0.081 (0.035)**	0.071 (0.036)**	23.096 (4.224)***	20.001 (5.260)***
w1_age55to60	0.036 (0.100)	0.005 (0.104)	0.133 (0.121)	0.117 (0.122)	-50.209 (68.696)	-57.729 (74.106)
w1_age60to65	0.182 (0.101)*	0.173 (0.106)	0.114 (0.126)	0.102 (0.127)	-49.778 (73.424)	-58.811 (80.001)
w1_age65to70	0.075 (0.108)	0.074 (0.118)	0.260 (0.127)**	0.240 (0.131)*	-28.557 (68.158)	-43.028 (78.658)
w1_age70to75	0.302 (0.107)***	0.268 (0.129)**	0.189 (0.137)	0.177 (0.143)	-28.665 (69.619)	-50.689 (85.903)
w1_age75to80	0.356 (0.123)***	0.303 (0.153)**	0.156 (0.158)	0.127 (0.172)	-41.977 (68.530)	-73.916 (90.330)
w1_age80to85	0.420 (0.141)***	0.357 (0.174)**	0.225 (0.183)	0.167 (0.201)	-22.609 (68.027)	-52.956 (90.695)
w1_female	-0.134 (0.061)**	-0.163 (0.104)				
w1_sweden	-0.272 (0.140)*	-0.175 (0.303)	0.165 (0.167)	0.255 (0.262)	-38.876 (10.255)***	52.635 (65.856)
w1_netherlands	-0.301 (0.134)**	-0.243 (0.156)	0.234 (0.155)	0.262 (0.167)	-47.335 (11.157)***	-18.834 (19.956)
w1_spain	-0.294 (0.143)**	-0.241 (0.166)	0.042 (0.176)	0.060 (0.191)	-28.742 (12.036)**	1.388 (24.439)
w1_italy	-0.071 (0.126)	-0.035 (0.150)	-0.322 (0.187)*	-0.291 (0.193)	171.018 (131.593)	201.672 (152.954)
w1_france	-0.180 (0.113)	-0.130 (0.121)	0.058 (0.147)	0.068 (0.150)	-38.723 (10.737)***	-24.453 (12.908)*
w1_greece	-0.420 (0.125)***	-0.379 (0.127)***	-0.392 (0.174)**	-0.395 (0.175)**	46.349 (12.844)***	50.342 (13.326)***
w1_belgium	-0.152 (0.119)	-0.133 (0.122)	0.076 (0.153)	0.047 (0.155)	89.519 (17.119)***	92.879 (18.404)***
w1_corr_health		0.020 (0.125)		0.043 (0.093)		43.195 (31.157)
constant	-1.524 (0.131)***	-1.542 (0.146)***	-1.917 (0.167)***	-1.882 (0.168)***	42.775 (61.045)	51.188 (67.620)
N	2,963	2,907	2,964	2,908	2,976	2,919

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., a dummy for having been in hospital yes or no; a dummy for having had outpatient surgery yes or no; the amount of out-of-pocket expenses on prescribed medicines); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 8: Test regressions. Various types of health care usage as outcome variables (mobility domain).

	In hospital?		Had outpatient surgery?		Out-of-pocket drugs expenses	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.477 (0.087)***	0.492 (0.088)***	0.673 (0.121)***	0.695 (0.122)***	0.270 (0.083)***	0.273 (0.085)***
w1_self_report	0.196 (0.031)***	0.188 (0.031)***	0.071 (0.037)*	0.080 (0.038)**	17.161 (9.504)*	14.919 (11.712)
w1_age55to60	0.036 (0.100)	0.014 (0.101)	0.137 (0.120)	0.134 (0.121)	-51.179 (68.588)	-53.673 (71.870)
w1_age60to65	0.156 (0.101)	0.152 (0.103)	0.120 (0.126)	0.121 (0.128)	-51.285 (72.631)	-60.216 (81.961)
w1_age65to70	0.051 (0.108)	0.030 (0.124)	0.249 (0.127)**	0.275 (0.142)*	-31.729 (67.172)	-56.389 (93.316)
w1_age70to75	0.237 (0.108)**	0.205 (0.140)	0.174 (0.139)	0.248 (0.160)	-32.119 (66.270)	-69.736 (104.999)
w1_age75to80	0.267 (0.125)**	0.212 (0.176)	0.136 (0.161)	0.232 (0.207)	-44.850 (63.996)	-98.492 (118.116)
w1_age80to85	0.346 (0.144)**	0.272 (0.209)	0.198 (0.184)	0.295 (0.249)	-26.807 (63.119)	-86.363 (127.853)
w1_female	-0.104 (0.060)**	-0.125 (0.066)*				
w1_sweden	-0.406 (0.139)***	-0.333 (0.160)**	0.102 (0.165)	0.044 (0.184)	-53.434 (10.741)***	-21.651 (35.376)
w1_netherlands	-0.321 (0.132)**	-0.267 (0.142)*	0.228 (0.154)	0.194 (0.164)	-51.849 (11.673)***	-31.296 (20.070)
w1_spain	-0.311 (0.142)**	-0.272 (0.143)*	0.068 (0.174)	0.018 (0.177)	-30.714 (11.944)***	-26.321 (12.771)**
w1_italy	-0.056 (0.126)	-0.031 (0.131)	-0.322 (0.187)*	-0.294 (0.191)	169.883 (130.051)	160.883 (120.134)
w1_france	-0.139 (0.113)	-0.088 (0.127)	0.079 (0.148)	0.014 (0.159)	-34.145 (11.632)***	-11.783 (20.366)
w1_greece	-0.381 (0.125)***	-0.318 (0.160)**	-0.376 (0.176)**	-0.457 (0.213)	49.421 (13.623)***	90.858 (41.325)**
w1_belgium	-0.149 (0.119)	-0.126 (0.123)	0.080 (0.153)	0.039 (0.157)	89.079 (17.157)***	97.102 (20.246)***
w1_corr_health		0.031 (0.089)		-0.073 (0.095)		39.081 (40.897)
constant	-1.415 (0.124)***	-1.326 (0.310)***	-1.856 (0.162)***	-2.069 (0.319)***	64.896 (72.790)	182.507 (194.315)
N	2,969	2,909	2,970	2,910	2,982	2,920

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.
w1_y is the outcome variable in wave 1 (i.e., a dummy for having been in hospital yes or no; a dummy for having had outpatient surgery yes or no; the amount of out-of-pocket expenses on prescribed medicines); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 9: Test regressions. Various types of health care usage as outcome variables (sleep domain).

	In hospital?		Had outpatient surgery?		Out-of-pocket drugs expenses	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.495 (0.087)***	0.493 (0.088)***	0.675 (0.122)***	0.687 (0.122)***	0.276 (0.083)***	0.278 (0.085)***
w1_self_report	0.198 (0.031)***	0.204 (0.031)***	0.030 (0.034)	0.036 (0.034)	6.244 (14.745)	1.722 (18.823)
w1_age55to60	0.036 (0.100)	0.005 (0.104)	0.126 (0.121)	0.115 (0.122)	-49.913 (68.419)	-52.106 (70.803)
w1_age60to65	0.182 (0.101)*	0.173 (0.106)	0.112 (0.126)	0.117 (0.126)	-48.602 (73.582)	-60.792 (84.260)
w1_age65to70	0.075 (0.108)	0.074 (0.118)	0.264 (0.126)**	0.260 (0.129)**	-27.221 (68.347)	-41.000 (80.705)
w1_age70to75	0.302 (0.107)***	0.268 (0.129)**	0.202 (0.136)	0.217 (0.137)	-24.403 (68.806)	-49.044 (88.893)
w1_age75to80	0.356 (0.123)***	0.303 (0.153)**	0.171 (0.157)	0.175 (0.169)	-34.069 (66.007)	-75.258 (100.988)
w1_age80to85	0.420 (0.141)***	0.357 (0.174)**	0.243 (0.181)	0.217 (0.196)	-15.248 (66.986)	-56.303 (103.754)
w1_female	-0.134 (0.061)**	-0.163 (0.104)				
w1_sweden	-0.272 (0.140)*	-0.175 (0.303)	0.128 (0.165)	0.121 (0.294)	-49.023 (11.868)***	116.401 (138.176)
w1_netherlands	-0.301 (0.134)**	-0.243 (0.156)	0.207 (0.153)	0.214 (0.159)	-54.916 (11.347)***	-30.774 (21.633)
w1_spain	-0.294 (0.143)**	-0.241 (0.166)	0.039 (0.175)	0.030 (0.181)	-31.068 (12.006)***	-7.142 (23.248)
w1_italy	-0.071 (0.126)	-0.035 (0.150)	-0.328 (0.187)*	-0.323 (0.191)*	166.194 (133.051)	143.050 (112.720)
w1_france	-0.180 (0.113)	-0.130 (0.121)	0.046 (0.146)	0.046 (0.148)	-41.686 (10.711)***	-52.712 (14.522)***
w1_greece	-0.420 (0.125)***	-0.379 (0.127)***	-0.400 (0.174)**	-0.404 (0.183)**	42.637 (13.731)***	74.374 (26.564)***
w1_belgium	-0.152 (0.119)	-0.133 (0.122)	0.068 (0.152)	0.039 (0.154)	85.406 (17.272)***	76.986 (17.186)***
w1_corr_health		0.020 (0.125)		-0.005 (0.087)		61.063 (53.173)
constant	-1.524 (0.131)***	-1.542 (0.146)***	-1.793 (0.157)***	-1.808 (0.247)***	80.125 (85.938)	221.339 (208.126)
N	2,963	2,907	2,968	2,910	2,980	2,921

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., a dummy for having been in hospital yes or no; a dummy for having had outpatient surgery yes or no; the amount of out-of-pocket expenses on prescribed medicines); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 10: Test regressions. Various types of health care usage as outcome variables (breath domain).

	In hospital?		Had outpatient surgery?		Out-of-pocket drugs expenses	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.558 (0.085)***	0.560 (0.086)***	0.680 (0.121)***	0.706 (0.123)***	0.272 (0.085)***	0.271 (0.086)***
w1_self_report	0.110 (0.036)***	0.116 (0.037)***	-0.018 (0.047)	-0.009 (0.047)**	17.340 (6.732)***	16.290 (7.305)**
w1_age55to60	0.042 (0.100)	0.024 (0.102)	0.139 (0.120)	0.127 (0.122)	-50.178 (68.993)	-54.339 (72.041)
w1_age60to65	0.174 (0.100)*	0.168 (0.104)	0.113 (0.126)	0.102 (0.127)	-49.933 (73.235)	-55.945 (77.311)
w1_age65to70	0.081 (0.107)	0.023 (0.153)	0.266 (0.127)**	0.290 (0.148)*	-29.814 (68.366)	-59.024 (86.091)
w1_age70to75	0.295 (0.107)***	0.212 (0.176)	0.213 (0.138)	0.231 (0.167)	-28.602 (68.805)	-67.489 (92.265)
w1_age75to80	0.355 (0.123)***	0.236 (0.215)	0.214 (0.157)	0.230 (0.202)	-37.184 (67.281)	-88.639 (96.736)
w1_age80to85	0.440 (0.140)***	0.326 (0.205)	0.251 (0.183)	0.254 (0.214)	-19.649 (67.291)	-57.477 (91.122)
w1_female	-0.057 (0.060)	-0.123 (0.094)				
w1_sweden	-0.467 (0.140)***	-0.584 (0.248)**	0.115 (0.165)	0.137 (0.224)	-62.465 (11.565)***	-122.851 (37.217)***
w1_netherlands	-0.364 (0.131)***	-0.276 (0.157)*	0.209 (0.153)	0.209 (0.166)	-54.226 (11.492)***	-28.487 (15.811)*
w1_spain	-0.307 (0.141)**	-0.134 (0.254)	0.060 (0.174)	0.006 (0.243)	-29.801 (11.960)**	35.194 (40.586)
w1_italy	-0.071 (0.124)	-0.053 (0.127)	-0.378 (0.190)**	-0.360 (0.192)*	169.997 (131.294)	174.486 (134.216)
w1_france	-0.244 (0.112)**	-0.471 (0.429)	0.051 (0.146)	0.101 (0.339)	-42.262 (10.787)***	-165.361 (79.707)**
w1_greece	-0.478 (0.123)***	-0.397 (0.146)***	-0.414 (0.173)**	-0.418 (0.187)**	42.123 (12.789)***	67.271 (20.115)***
w1_belgium	-0.190 (0.118)	-0.340 (0.270)	0.081 (0.151)	0.092 (0.227)	85.621 (17.251)***	14.345 (43.777)**
w1_corr_health		0.168 (0.261)		-0.026 (0.191)		77.909 (48.156)
constant	-1.249 (0.125)***	-0.795 (0.762)	-1.708 (0.158)***	-1.788 (0.536)***	69.076 (66.106)	280.725 (193.966)
N	2,964	2,903	2,965	2,904	2,977	2,914

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., a dummy for having been in hospital yes or no; a dummy for having had outpatient surgery yes or no; the amount of out-of-pocket expenses on prescribed medicines); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 11: Test regressions. Various types of health care usage as outcome variables (depress domain).

	In hospital?		Had outpatient surgery?		Out-of-pocket drugs expenses	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.560 (0.085)***	0.561 (0.086)***	0.662 (0.121)***	0.695 (0.122)***	0.279 (0.084)***	0.282 (0.086)***
w1_self_report	0.117 (0.030)***	0.120 (0.030)***	0.048 (0.038)	0.055 (0.038)	3.245 (14.595)	0.509 (17.757)
w1_age55to60	0.065 (0.099)	0.106 (0.104)	0.149 (0.120)	0.098 (0.127)	-49.580 (70.487)	-31.247 (52.004)
w1_age60to65	0.198 (0.100)**	0.212 (0.101)**	0.137 (0.125)	0.116 (0.127)	-48.805 (74.827)	-42.416 (68.814)
w1_age65to70	0.121 (0.107)	0.127 (0.107)	0.274 (0.126)**	0.292 (0.126)**	-27.139 (70.657)	-25.878 (70.793)
w1_age70to75	0.340 (0.107)***	0.316 (0.108)***	0.214 (0.136)	0.240 (0.136)*	-23.689 (70.322)	-28.382 (76.238)
w1_age75to80	0.380 (0.123)***	0.331 (0.125)***	0.208 (0.155)	0.230 (0.156)	-34.099 (67.776)	-41.204 (75.910)
w1_age80to85	0.478 (0.140)***	0.321 (0.160)**	0.246 (0.182)	0.326 (0.206)	-15.018 (67.419)	-56.083 (112.908)
w1_female	-0.091 (0.060)	-0.138 (0.063)**				
w1_sweden	-0.450 (0.137)***	-0.402 (0.141)***	0.087 (0.165)	0.058 (0.169)	-53.240 (12.144)***	-37.167 (24.433)
w1_netherlands	-0.349 (0.130)***	-0.177 (0.157)	0.223 (0.153)	0.119 (0.186)	-54.980 (12.381)***	0.006 (52.349)
w1_spain	-0.325 (0.141)**	-0.224 (0.147)	0.060 (0.174)	-0.032 (0.184)	-32.047 (11.980)***	-4.692 (30.600)
w1_italy	-0.115 (0.124)	-0.194 (0.134)	-0.342 (0.187)*	-0.266 (0.194)	165.944 (133.313)	145.303 (109.902)
w1_france	-0.227 (0.111)**	-0.138 (0.118)	0.053 (0.146)	0.015 (0.153)	-40.865 (10.966)***	-17.899 (22.672)
w1_greece	-0.507 (0.124)***	-0.484 (0.125)***	-0.425 (0.173)**	-0.428 (0.174)**	39.691 (12.980)***	42.555 (13.760)***
w1_belgium	-0.184 (0.118)	-0.125 (0.121)	0.083 (0.152)	0.030 (0.158)	85.078 (17.265)***	102.329 (25.873)***
w1_corr_health		0.177 (0.098)*		-0.137 (0.115)		62.997 (66.274)
constant	-1.300 (0.121)***	-0.894 (0.268)***	-1.827 (0.159)***	-2.152 (0.304)***	87.024 (84.061)	233.956 (238.873)
N	2,969	2,908	2,970	2,909	2,982	2,919

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., a dummy for having been in hospital yes or no; a dummy for having had outpatient surgery yes or no; the amount of out-of-pocket expenses on prescribed medicines); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 12: Test regressions. Various types of health care usage as outcome variables (work disability domain).

	In hospital?		Had outpatient surgery?		Out-of-pocket drugs expenses	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.470 (0.087)***	0.476 (0.088)***	0.672 (0.122)***	0.666 (0.124)***	0.273 (0.085)***	0.269 (0.085)***
w1_self_report	0.183 (0.028)***	0.174 (0.029)***	0.053 (0.034)	0.062 (0.035)*	13.578 (8.693)	9.815 (11.614)
w1_age55to60	0.009 (0.102)	-0.012 (0.103)	0.113 (0.121)	0.089 (0.123)	-50.594 (67.849)	-54.491 (72.195)
w1_age60to65	0.158 (0.101)	0.126 (0.106)	0.106 (0.127)	0.127 (0.129)	-50.878 (72.360)	-70.379 (89.311)
w1_age65to70	0.050 (0.109)	0.028 (0.116)	0.258 (0.127)**	0.296 (0.133)**	-30.211 (67.006)	-47.837 (86.956)
w1_age70to75	0.232 (0.108)**	0.174 (0.128)	0.189 (0.138)	0.246 (0.149)*	-30.462 (66.360)	-62.761 (102.896)
w1_age75to80	0.312 (0.123)**	0.237 (0.147)	0.180 (0.157)	0.242 (0.177)	-39.694 (64.946)	-76.649 (106.017)
w1_age80to85	0.333 (0.143)**	0.204 (0.188)	0.206 (0.185)	0.276 (0.234)	-23.252 (62.692)	-77.325 (122.729)
w1_female	-0.093 (0.060)	-0.105 (0.063)*				
w1_sweden	-0.418 (0.140)***	-0.351 (0.145)**	0.119 (0.165)	0.100 (0.169)	-51.318 (10.792)***	-37.372 (19.425)*
w1_netherlands	-0.348 (0.133)***	-0.271 (0.145)*	0.215 (0.154)	0.193 (0.169)	-52.487 (12.007)***	-26.928 (28.164)
w1_spain	-0.350 (0.142)**	-0.307 (0.143)**	0.030 (0.175)	-0.008 (0.179)	-32.835 (12.079)***	-29.779 (12.936)**
w1_italy	-0.110 (0.125)	-0.079 (0.128)	-0.319 (0.186)*	-0.297 (0.187)	169.510 (130.904)	169.329 (128.905)
w1_france	-0.205 (0.112)*	-0.161 (0.116)	0.044 (0.146)	0.031 (0.150)	-38.611 (11.185)***	-29.373 (13.696)**
w1_greece	-0.416 (0.125)***	-0.332 (0.147)**	-0.383 (0.174)**	-0.435 (0.203)**	47.446 (14.004)***	82.411 (38.661)**
w1_belgium	-0.216 (0.119)*	-0.157 (0.122)	0.090 (0.152)	0.059 (0.157)	86.188 (17.228)***	92.880 (22.294)***
w1_corr_health		0.061 (0.081)		-0.062 (0.093)		39.010 (44.325)
constant	-1.388 (0.124)***	-1.233 (0.260)***	-1.827 (0.153)***	-1.993 (0.282)***	69.260 (72.512)	178.064 (191.033)
N	2,959	2,889	2,960	2,890	2,971	2,901

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., a dummy for having been in hospital yes or no; a dummy for having had outpatient surgery yes or no; the amount of out-of-pocket expenses on prescribed medicines); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 13: Test regressions. Results of objective measured tests as outcome variables (pain domain).

	Walking speed		Grip strength	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.370 (0.124)***	0.350 (0.123)***	0.565 (0.023)***	0.560 (0.023)***
w1_self_report	-0.027 (0.026)	-0.031 (0.031)	-0.396 (0.137)***	-0.417 (0.139)***
w1_age55to60			-0.828 (0.383)**	-0.948 (0.400)**
w1_age60to65			-2.043 (0.404)***	-2.173 (0.428)***
w1_age65to70			-2.940 (0.422)***	-3.189 (0.466)***
w1_age70to75			-4.464 (0.456)***	-4.877 (0.541)***
w1_age75to80	0.058 (0.067)	0.076 (0.071)	-5.389 (0.562)***	-5.899 (0.694)***
w1_age80to85	0.108 (0.089)	0.141 (0.098)	-6.364 (0.666)***	-6.836 (0.792)***
w1_sweden	-0.081 (0.089)	-0.182 (0.154)	0.801 (0.611)	1.967 (1.225)
w1_netherlands	-0.106 (0.119)	-0.146 (0.134)	-1.009 (0.556)*	-0.602 (0.661)
w1_spain	-0.049 (0.094)	-0.081 (0.103)	-2.059 (0.583)***	-1.727 (0.682)**
w1_italy	-0.215 (0.081)***	-0.231 (0.090)**	-0.250 (0.522)	0.073 (0.617)
w1_france	0.031 (0.078)	0.023 (0.086)	-0.347 (0.436)	-0.299 (0.475)
w1_greece	-0.128 (0.105)	-0.132 (0.109)	0.516 (0.469)	0.471 (0.474)
w1_belgium	0.002 (0.060)	0.000 (0.066)	0.667 (0.506)	0.696 (0.512)
w1_corr_health		-0.050 (0.066)		0.620 (0.493)
constant	0.457 (0.155)***	0.491 (0.161)***	21.229 (1.313)***	21.792 (1.335)***
N	156	149	2,688	2,640

	Chair stand		Peak flow	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y				
w1_self_report	-0.088 (0.186)	-0.067 (0.184)	0.80 (3.11)	-0.62 (3.21)
w1_age55to60	0.178 (0.610)	0.316 (0.590)	-0.63 (9.59)	-0.87 (9.78)
w1_age60to65	-0.536 (0.635)	-0.452 (0.607)	-2.65 (10.18)	-3.28 (10.42)
w1_age65to70	0.181 (0.707)	0.442 (0.676)	-4.31 (10.32)	-4.33 (10.70)
w1_age70to75	0.302 (0.773)	0.448 (0.774)	-6.48 (10.96)	-8.46 (11.85)
w1_age75to80	-0.494 (0.666)	-0.143 (0.646)	-9.47 (12.52)	-14.90 (13.69)
w1_age80to85	-1.045 (0.679)	-0.607 (0.682)	0.23 (14.37)	-3.27 (16.31)
w1_spain	1.673 (0.968)*	1.358 (0.929)	-23.81 (12.60)*	-17.63 (13.53)
w1_italy	-0.257 (0.595)	-0.497 (0.659)	16.48 (12.81)	21.28 (13.72)
w1_france	1.015 (0.597)*	0.852 (0.613)	-41.93 (11.37)***	-38.74 (11.72)***
w1_greece	2.329 (0.725)***	2.332 (0.721)***	-18.04 (11.77)	-17.71 (11.87)
w1_belgium	1.027 (0.638)	1.036 (0.650)	-35.74 (11.74)***	-33.78 (11.89)***
w1_corr_health		-0.448 (0.544)		7.67 (7.60)
constant	10.869 (0.871)***	10.777 (0.862)***	348.98 (13.34)***	350.51 (13.59)***
N	1,502	1,469	2,026	1,985

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., walking speed test, grip strength test, chair stand test or peak flow test); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 14: Test regressions. Results of objective measured tests as outcome variables (mobility domain).

	Walking speed		Grip strength	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.266 (0.115)**	0.246 (0.116)**	0.563 (0.023)***	0.559 (0.024)***
w1_self_report	-0.087 (0.022)***	-0.088 (0.022)***	-0.512 (0.159)***	-0.463 (0.161)***
w1_age55to60			-0.790 (0.383)**	-0.762 (0.387)**
w1_age60to65			-1.995 (0.402)***	-1.821 (0.421)***
w1_age65to70			-2.912 (0.420)***	-2.442 (0.485)***
w1_age70to75			-4.338 (0.458)***	-3.737 (0.569)***
w1_age75to80	-0.007 (0.092)	0.019 (0.099)	-5.241 (0.566)***	-4.365 (0.756)***
w1_age80to85	0.039 (0.108)	0.080 (0.117)	-6.207 (0.675)***	-5.096 (0.916)***
w1_sweden	0.015 (0.101)	-0.015 (0.117)	1.063 (0.606)*	0.439 (0.695)
w1_netherlands	-0.115 (0.121)	-0.120 (0.126)	-0.943 (0.551)*	-1.296 (0.594)**
w1_spain	-0.045 (0.109)	-0.040 (0.112)	-2.028 (0.580)***	-2.221 (0.588)***
w1_italy	-0.212 (0.077)***	-0.179 (0.090)**	-0.325 (0.520)	-0.164 (0.535)
w1_france	0.037 (0.089)	0.022 (0.102)	-0.490 (0.439)	-1.055 (0.495)**
w1_greece	-0.172 (0.108)	-0.224 (0.136)	0.344 (0.477)	-0.491 (0.616)
w1_belgium	0.046 (0.072)	0.044 (0.077)	0.618 (0.504)	0.382 (0.509)
w1_corr_health		-0.059 (0.081)		-0.692 (0.365)*
constant	0.730 (0.136)***	0.581 (0.229)**	21.299 (1.296)***	19.378 (1.667)***
N	157	153	2,695	2,645

	Chair stand		Peak flow	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y				
w1_self_report	-0.415 (0.177)**	-0.360 (0.178)**	0.48 (3.39)	-0.05 (3.49)
w1_age55to60	0.117 (0.608)	0.286 (0.596)	0.23 (9.59)	-0.92 (9.76)
w1_age60to65	-0.542 (0.638)	-0.257 (0.604)	-2.04 (10.15)	-5.58 (10.60)
w1_age65to70	0.205 (0.709)	1.033 (0.674)	-2.80 (10.29)	-7.56 (12.02)
w1_age70to75	0.520 (0.777)	1.520 (0.867)*	-6.71 (11.12)	-12.76 (14.31)
w1_age75to80	-0.162 (0.670)	1.347 (0.949)	-9.63 (12.71)	-22.18 (17.68)
w1_age80to85	-0.762 (0.701)	0.976 (1.074)	0.92 (14.52)	-9.27 (20.82)
w1_spain	1.609 (0.961)*	1.580 (0.946)*	-23.24 (12.56)*	-22.08 (12.71)*
w1_italy	-0.310 (0.595)	0.085 (0.670)	17.15 (12.79)	13.98 (13.31)
w1_france	0.761 (0.594)	0.167 (0.680)	-42.47 (11.46)***	-35.97 (12.41)***
w1_greece	2.204 (0.723)***	1.106 (0.878)	-18.63 (11.85)	-10.46 (14.63)
w1_belgium	0.963 (0.631)	0.759 (0.641)	-35.47 (11.73)***	-33.10 (11.94)***
w1_corr_health		-1.102 (0.667)*		7.91 (8.58)
constant	11.377 (0.711)***	7.995 (1.926)***	349.34 (12.72)***	373.06 (29.70)***
N	1,504	1,471	2,030	1,985

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., walking speed test, grip strength test, chair stand test or peak flow test); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 15: Test regressions. Results of objective measured tests as outcome variables (sleep domain).

	Walking speed		Grip strength	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.352 (0.126)***	0.326 (0.132)**	0.566 (0.023)***	0.558 (0.024)***
w1_self_report	-0.044 (0.029)	-0.043 (0.030)	-0.330 (0.114)***	-0.310 (0.116)***
w1_age55to60			-0.830 (0.383)**	-0.770 (0.385)**
w1_age60to65			-2.062 (0.403)***	-1.924 (0.419)***
w1_age65to70			-2.982 (0.422)***	-2.821 (0.438)***
w1_age70to75			-4.491 (0.458)***	-4.319 (0.484)***
w1_age75to80	0.022 (0.069)	0.044 (0.072)	-5.429 (0.563)***	-5.039 (0.633)***
w1_age80to85	0.068 (0.084)	0.104 (0.091)	-6.441 (0.668)***	-6.025 (0.726)***
w1_sweden	-0.088 (0.087)	-0.264 (0.177)	0.881 (0.607)	-1.062 (1.207)
w1_netherlands	-0.100 (0.111)	-0.175 (0.108)	-0.864 (0.552)	-1.144 (0.575)**
w1_spain	-0.039 (0.101)	-0.049 (0.106)	-2.005 (0.586)***	-2.399 (0.607)***
w1_italy	-0.225 (0.079)***	-0.166 (0.096)*	-0.168 (0.520)	0.055 (0.542)
w1_france	0.028 (0.078)	0.066 (0.092)	-0.253 (0.433)	-0.299 (0.439)
w1_greece	-0.165 (0.104)	-0.185 (0.110)	0.534 (0.476)	0.060 (0.517)
w1_belgium	0.001 (0.061)	0.022 (0.071)	0.762 (0.501)	0.770 (0.507)
w1_corr_health		-0.072 (0.063)		-0.687 (0.375)*
constant	0.607 (0.168)***	0.408 (0.218)*	20.915 (1.287)***	19.660 (1.554)***
N	157	152	2,691	2,640

	Chair stand		Peak flow	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y				
w1_self_report	-0.147 (0.175)	-0.083 (0.178)	-2.35 (2.86)	-3.49 (2.96)
w1_age55to60	0.177 (0.611)	0.278 (0.611)	-0.33 (9.58)	-1.09 (9.71)
w1_age60to65	-0.548 (0.643)	-0.343 (0.623)	-2.91 (10.17)	-4.66 (10.51)
w1_age65to70	0.138 (0.709)	0.456 (0.685)	-4.18 (10.26)	-4.89 (10.64)
w1_age70to75	0.278 (0.784)	0.518 (0.768)*	-6.24 (10.92)	-7.12 (11.69)
w1_age75to80	-0.383 (0.673)	0.150 (0.661)	-10.69 (12.45)	-17.77 (13.72)
w1_age80to85	-1.051 (0.690)	-0.586 (0.678)	0.69 (14.35)	-2.69 (16.02)
w1_spain	1.679 (0.967)*	1.341 (0.892)*	-24.03 (12.57)*	-20.10 (12.89)
w1_italy	-0.255 (0.596)	0.065 (0.676)	17.67 (12.77)	15.26 (13.30)
w1_france	1.028 (0.594)*	1.134 (0.625)*	-41.69 (11.32)***	-42.17 (11.63)***
w1_greece	2.349 (0.719)***	1.972 (0.707)***	-19.50 (11.78)*	-15.55 (12.32)
w1_belgium	1.060 (0.639)*	1.158 (0.656)*	-35.32 (11.71)***	-33.90 (11.86)***
w1_corr_health		-0.783 (0.556)		8.26 (7.54)
constant	10.969 (0.744)***	9.127 (1.336)**	355.53 (12.93)***	374.72 (22.36)***
N	1,504	1,474	2,029	1,988

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., walking speed test, grip strength test, chair stand test or peak flow test); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 16: Test regressions. Results of objective measured tests as outcome variables (breath domain).

	Walking speed		Grip strength	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.372 (0.113)***	0.348 (0.113)***	0.566 (0.023)***	0.561 (0.023)***
w1_self_report	-0.053 (0.027)*	-0.054 (0.029)*	-0.445 (0.176)**	-0.406 (0.179)**
w1_age55to60			-0.804 (0.383)**	-0.653 (0.390)*
w1_age60to65			-2.083 (0.401)***	-1.918 (0.410)***
w1_age65to70			-2.948 (0.422)***	-2.125 (0.560)***
w1_age70to75			-4.422 (0.459)***	-3.384 (0.660)***
w1_age75to80	0.017 (0.079)	0.033 (0.086)	-5.401 (0.563)***	-4.061 (0.842)***
w1_age80to85	0.063 (0.092)	0.081 (0.101)	-6.343 (0.670)***	-5.135 (0.844)***
w1_sweden	-0.017 (0.093)	0.061 (0.152)	1.278 (0.613)**	2.849 (0.970)***
w1_netherlands	-0.060 (0.110)	-0.075 (0.119)	-0.897 (0.548)	-1.596 (0.641)**
w1_spain	-0.012 (0.098)	-0.072 (0.130)	-2.027 (0.585)***	-3.892 (1.034)***
w1_italy	-0.204 (0.084)**	-0.184 (0.099)*	-0.332 (0.522)	-0.288 (0.528)
w1_france	0.079 (0.082)	0.217 (0.231)	-0.227 (0.436)	2.975 (1.594)*
w1_greece	-0.123 (0.102)	-0.145 (0.107)	0.589 (0.471)	-0.146 (0.565)
w1_belgium	0.042 (0.055)	0.126 (0.135)	0.705 (0.503)	2.616 (1.023)**
w1_corr_health		-0.082 (0.120)		-2.124 (0.984)**
constant	0.486 (0.137)***	0.312 (0.321)	20.976 (1.298)***	15.147 (2.905)***
N	156	150	2,689	2,634

	Chair stand		Peak flow	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y				
w1_self_report	0.118 (0.220)	0.147 (0.225)	-8.23 (4.00)**	-8.53 (4.06)**
w1_age55to60	0.080 (0.614)	0.215 (0.601)	-0.02 (9.60)	-3.63 (9.72)
w1_age60to65	-0.604 (0.639)	-0.493 (0.622)	-1.99 (10.14)	-5.07 (10.35)
w1_age65to70	0.094 (0.706)	0.401 (0.696)	-2.53 (10.22)	-16.09 (12.23)
w1_age70to75	0.233 (0.773)	0.413 (0.849)	-3.34 (10.99)	-17.55 (14.10)
w1_age75to80	-0.571 (0.669)	-0.272 (0.777)	-6.46 (12.46)	-30.03 (16.60)*
w1_age80to85	-1.096 (0.686)	-1.099 (0.741)	3.55 (14.30)	-11.25 (17.93)
w1_spain	1.667 (0.962)*	1.338 (1.129)	-22.34 (12.54)*	6.07 (18.55)
w1_italy	-0.258 (0.595)	-0.272 (0.607)	19.13 (12.75)	16.94 (12.84)
w1_france	0.863 (0.590)	1.527 (1.766)	-39.15 (11.28)***	-93.34 (28.77)***
w1_greece	2.392 (0.720)***	2.226 (0.714)***	-16.82 (11.69)	-5.86 (12.81)
w1_belgium	1.032 (0.632)	1.425 (1.161)	-33.36 (11.66)***	-65.81 (19.12)***
w1_corr_health		-0.434 (1.030)		34.63 (16.59)**
constant	10.563 (0.725)***	9.306 (2.750)***	359.26 (12.64)***	452.75 (46.74)***
N	1,502	1,468	2,026	1,980

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., walking speed test, grip strength test, chair stand test or peak flow test); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 17: Test regressions. Results of objective measured tests as outcome variables (depress domain).

	Walking speed		Grip strength	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.367 (0.122)***	0.348 (0.121)***	0.565 (0.023)***	0.560 (0.024)***
w1_self_report	-0.026 (0.028)	-0.025 (0.030)	-0.518 (0.126)***	-0.517 (0.127)***
w1_age55to60			-0.864 (0.383)**	-1.007 (0.400)**
w1_age60to65			-2.122 (0.401)***	-2.168 (0.401)***
w1_age65to70			-3.080 (0.421)***	-3.080 (0.427)***
w1_age70to75			-4.573 (0.458)***	-4.592 (0.466)***
w1_age75to80	0.055 (0.070)	0.050 (0.072)	-5.473 (0.561)***	-5.447 (0.572)***
w1_age80to85	0.103 (0.089)	0.137 (0.095)	-6.482 (0.668)***	-6.093 (0.736)***
w1_sweden	-0.043 (0.093)	-0.041 (0.099)	1.126 (0.609)*	0.915 (0.624)
w1_netherlands	-0.088 (0.113)	-0.121 (0.125)	-0.914 (0.552)*	-1.320 (0.665)**
w1_spain	-0.017 (0.101)	-0.029 (0.110)	-2.008 (0.585)***	-2.335 (0.619)***
w1_italy	-0.201 (0.076)***	-0.156 (0.104)	-0.179 (0.523)	0.018 (0.554)
w1_france	0.051 (0.077)	0.044 (0.085)	-0.312 (0.438)	-0.633 (0.466)
w1_greece	-0.142 (0.102)	-0.132 (0.103)	0.720 (0.474)	0.625 (0.475)
w1_belgium	0.028 (0.057)	0.021 (0.062)	0.673 (0.505)	0.506 (0.514)
w1_corr_health		-0.061 (0.098)		-0.519 (0.405)
constant	0.537 (0.150)***	0.311 (0.265)	21.288 (1.291)***	20.339 (1.527)***
N	157	151	2,694	2,637

	Chair stand		Peak flow	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y				
w1_self_report	-0.134 (0.188)	-0.058 (0.189)	0.80 (3.22)	0.52 (3.27)
w1_age55to60	0.066 (0.619)	-0.339 (0.712)	0.09 (9.60)	0.41 (10.01)
w1_age60to65	-0.609 (0.646)	-0.695 (0.672)	-2.08 (10.13)	-2.41 (10.12)
w1_age65to70	0.087 (0.717)	0.249 (0.726)	-3.54 (10.22)	-3.83 (10.35)
w1_age70to75	0.268 (0.787)	0.398 (0.786)	-6.76 (10.93)	-4.42 (11.17)
w1_age75to80	-0.467 (0.672)	-0.262 (0.658)	-9.67 (12.45)	-13.57 (12.60)
w1_age80to85	-1.059 (0.691)	-0.028 (0.707)	1.18 (14.35)	-1.15 (16.78)
w1_spain	1.659 (0.963)*	1.135 (0.910)	-20.86 (12.55)*	-18.72 (13.25)
w1_italy	-0.229 (0.598)	0.539 (0.669)	19.28 (12.73)	15.61 (13.51)
w1_france	0.901 (0.588)	0.471 (0.623)	-39.87 (11.33)***	-37.07 (11.94)***
w1_greece	2.426 (0.723)***	2.523 (0.722)***	-16.60 (11.72)	-17.12 (11.80)
w1_belgium	1.037 (0.633)	0.740 (0.639)	-33.52 (11.72)***	-33.29 (12.04)***
w1_corr_health		-1.567 (0.631)**		7.32 (9.69)
constant	10.948 (0.767)***	7.094 (1.447)***	346.58 (12.72)***	364.14 (26.42)***
N	1,504	1,469	2,030	1,985

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., walking speed test, grip strength test, chair stand test or peak flow test); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.

Table A4 - 18: Test regressions. Results of objective measured tests as outcome variables (work disability domain).

	Walking speed		Grip strength	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y	0.329 (0.130)**	0.331 (0.135)**	0.567 (0.023)***	0.564 (0.024)***
w1_self_report	-0.048 (0.026)*	-0.047 (0.028)	-0.292 (0.134)**	-0.240 (0.138)*
w1_age55to60			-0.779 (0.384)**	-0.756 (0.390)*
w1_age60to65			-2.030 (0.403)***	-1.901 (0.431)***
w1_age65to70			-2.927 (0.423)***	-2.611 (0.450)***
w1_age70to75			-4.442 (0.460)***	-4.105 (0.530)***
w1_age75to80	0.024 (0.064)	0.006 (0.068)	-5.320 (0.562)***	-5.052 (0.637)***
w1_age80to85	0.087 (0.085)	0.130 (0.094)	-6.247 (0.672)***	-5.482 (0.826)***
w1_sweden	-0.043 (0.094)	-0.081 (0.096)	0.906 (0.605)	0.634 (0.625)
w1_netherlands	-0.101 (0.113)	-0.178 (0.113)	-0.920 (0.554)*	-1.252 (0.611)**
w1_spain	-0.054 (0.094)	-0.029 (0.099)	-1.976 (0.586)***	-2.076 (0.591)***
w1_italy	-0.232 (0.083)***	-0.190 (0.093)**	-0.264 (0.524)	-0.218 (0.531)
w1_france	0.021 (0.084)	0.039 (0.094)	-0.337 (0.438)	-0.603 (0.451)
w1_greece	-0.179 (0.111)	-0.215 (0.132)	0.439 (0.476)	-0.090 (0.563)
w1_belgium	0.008 (0.064)	0.006 (0.072)	0.730 (0.506)	0.490 (0.513)
w1_corr_health		-0.066 (0.069)		-0.488 (0.326)
constant	0.643 (0.169)***	0.562 (0.188)***	20.821 (1.303)***	19.625 (1.513)***
N	157	147	2,685	2,622

	Chair stand		Peak flow	
	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)	coef. (st.dev)
w1_y				
w1_self_report	-0.068 (0.177)	0.008 (0.185)	-2.74 (3.11)	-3.23 (3.21)
w1_age55to60	0.173 (0.618)	0.303 (0.612)	-0.25 (9.66)	-0.66 (9.84)
w1_age60to65	-0.563 (0.635)	-0.015 (0.617)	-2.58 (10.20)	-6.52 (10.90)
w1_age65to70	0.145 (0.713)	0.892 (0.673)	-3.60 (10.38)	-6.99 (11.20)
w1_age70to75	0.135 (0.768)	1.250 (0.798)	-4.96 (11.25)	-11.15 (13.39)
w1_age75to80	-0.521 (0.691)	0.740 (0.709)	-9.42 (12.57)	-19.56 (14.93)
w1_age80to85	-1.039 (0.706)	0.745 (0.798)	3.57 (14.55)	-10.17 (18.99)
w1_spain	1.715 (0.965)*	1.663 (0.951)*	-22.51 (12.58)*	-21.92 (12.72)*
w1_italy	-0.258 (0.594)	-0.007 (0.603)	17.59 (12.83)	15.50 (13.06)
w1_france	0.999 (0.593)**	0.715 (0.612)	-40.80 (11.38)***	-39.08 (11.74)***
w1_greece	2.342 (0.744)***	1.179 (0.805)	-18.89 (11.87)	-11.08 (13.85)
w1_belgium	0.945 (0.626)	0.661 (0.639)	-34.61 (11.75)***	-31.64 (12.08)***
w1_corr_health		-1.342 (0.513)***		9.57 (7.70)
constant	10.824 (0.719)**	7.120 (1.413)***	354.46 (12.57)***	379.99 (24.58)***
N	1,501	1,467	2,023	1,976

* Indicates significance at the 10% significance level; ** indicates significance at the 5% level; *** indicates significance at the 1% level.

w1_y is the outcome variable in wave 1 (i.e., walking speed test, grip strength test, chair stand test or peak flow test); *w1_self_report* is the observed self-reported health measures in wave 1 for the specific health domain under consideration; *w1_corr_health* is the generated corrected latent health variable in wave 1.