Comparing health measures through a model for health care use

Master Thesis

MSc in Econometrics and Management Science, specialization: Econometrics
Erasmus School of Economics - Erasmus University

A. Exterkate
310472

August 11, 2011

Supervisors:

Dr. T.M. Marreiros Bago d'Uva
Dr. T.G.M. van Ourti

Co-reader:

Dr. C. Heij

Abstract

We consider two different methods of obtaining health measures. The first method consists of a health measurement model with a latent variable for general health. This model includes 23 self-assessed indicators of individual health, as well as the more objective health measure of grip strength to construct a health index that is corrected for differences in reporting styles across countries. The second method uses six subjective self-assessment questions on health problems in six different health domains, where responses are corrected for reporting style by making use of vignettes. To assess the usefulness of both methods, the alternative health measures are included in a model for health care use. The data that is used for implementation comes from the Survey of Health, Ageing and Retirement in Europe (SHARE). From this dataset, we use information from the Netherlands and Spain to perform all analyses. We find that both methods contain useful information on individuals' latent (unobserved) health, and that the health measures obtained from both methods are more useful in explaining health behavior than the uncorrected health indicators that are used to construct the health measures are. Once the six ill-health measures from the vignette method are included in the model for health care use, the latent health index does not improve explanatory power. Although the method using the health measurement model with a latent variable for health is found to be more helpful in finding relations between health indicators and "true" health, the method using vignettes seems to explain health-related behavior better.

Keywords:
health care use, health measurement, latent variables, reporting heterogeneity, self-assessed health, vignettes

Contents

1 Introduction

One of the major challenges in international research on health is the difficulty to measure an individual's health. This is mostly due to the fact that health is known to be a multidimensional concept, and thus health cannot be measured accurately by only considering an individual's physical symptoms and diagnoses; how an individual feels also seems to be of great importance. Therefore, it is challenging to obtain a general, accurate measure for individual health.

Another issue is that individual health across countries cannot be directly compared; if individuals' self-assessment of health is used, large social and cultural differences exist. In many surveys, due to the absence of medical reports, only self-assessments of health are available, which makes it even more difficult to compare individual health scores. The issue of social and cultural differences in self-assessed health across countries is supported by Zimmer, Natividad, Lin and Chayovan [2000]. They focused on surveys in the Philippines, Taiwan and Thailand, where a question on the self-assessment of health was asked, with responses on a categorical scale. They found that in some countries the responses were more positive than in other countries, even after controlling for more objective health measures. Jürges [2006] shows a similar result, using the Survey of Health, Ageing and Retirement in Europe (SHARE, the dataset that is also used in this thesis). He found that individuals in Scandinavian countries tend to overrate their health, while for example inhabitants of Germany and Southern Europeans tend to underrate their health. These results suggest that individuals in different countries have different perceptions when reporting self-assessed health, which makes it complicated to compare health across countries.

Surveys are very useful to measure an individual's health, particularly because they offer the possibility to obtain subjective information about the way an individual feels, next to objective information on individual health. According to the aforementioned results, however, there may exist large differences in reporting styles among individuals. So, on the one hand, self-assessment questions on health contain very useful information to measure individual health, but on the other hand, there is the issue of heterogeneity in reporting styles which makes it harder to use the information correctly.

So, we know that it is challenging to measure individual health, but the main question is why we care about this. The most important matter is that good measures of health are very useful in models for health-related behavior. Therefore, if we can construct health indices that are constructed from self-assessed health questions and that are corrected for heterogeneity in reporting styles, they will likely be good at explaining and predicting doctor visits, the use of medication, major health events, and many other health-related behaviors. Therefore, it is both interesting from a scientific point of view to tackle the problem of obtaining good health indices, as it is from a society's point of view if we can predict health-related behavior with these health indices.

The next step is how to overcome the above issues. First of all, we want to find out how to construct health indices that reflect only an individual's health and not also reporting style or country-specific differences. Secondly, if we find methods that are able to construct such health measures, we have to test which method should be used to obtain health measures that are actually usable, in the sense that the obtained measures explain and predict health-related behavior well. Otherwise, we might as well just use health information from the survey, instead of applying methods to correct this information. That is, we need to find out whether the applied methods lead to any improvement in explaining health behavior.

To analyze these research topics and find solutions for the above issues, in this thesis we use individuals' self-assessed information on health. This (subjective) information is known to be very useful, once it is corrected for differences in reporting styles across countries and individuals. To be more precise, we implement two different existing methods for obtaining health measures, to see how they perform and to find out how well they are at explaining and predicting health-related behavior.

One of these two methods is one by Meijer, Kapteyn and Andreyeva [2011] that uses a structural equation model with a latent variable for general health. The model constructs one index for general health that is cross-country comparable. For this purpose, 23 different self-assessed health indicators are included in the model, as well as a more objective indicator of individual health: grip strength. By doing this, we can get more insight in how the health indicators relate to actual (latent) health.

The other method is one that was implemented by Bago D'Uva, Lindeboom, O'Donnell and Van Doorslaer [2011a]. This method uses subjective self-assessment questions regarding health problems in six different health domains, where individuals were asked to indicate how many difficulties they experience in each of the six domains. We implement a hierarchical ordered probit model that uses vignettes to correct for heterogeneity in reporting styles. In this way, we can obtain indices for individual health problems that are corrected for reporting heterogeneity out of the self-assessment questions.

In the above two methods, we obtain individual (ill-)health indices that are corrected for differences in reporting styles across individuals and across countries. We make use of subjective information on health from surveys, that is known to be very useful in predicting behavior. To see how well the methods actually perform, we implement a model for health care use and include the obtained health measures. In this way we can find out whether the measures are actually reflecting an individual's health or health problems and thus capable of explaining and predicting whether an individual goes to see a medical doctor, and given that s/he does, how often. We do also include the obtained indices from both described methods simultaneously, to test whether the indices obtained from the two methods are complementary. That is, we test whether the health measures from one method still lead to an improvement in explanatory power, given that the measures from the other method are already included in the model. To compare the health indices from the two methods with the raw data from the survey, raw information on health will also be included, to consider whether these health variables contain useful information, and whether we might just use the raw information on health, instead of applying techniques to obtain health indices that might not even lead to an increase in explanatory power.

This thesis proceeds as follows. In Section 2, the two used methods for obtaining health measures, as well as the model for health care use, are described in more detail. Then, Section 3 gives an overview of the used data, as well as a description of the sample selection and it provides descriptive statistics. Section 4 shows and discusses the results of the two methods and of the model for health care use, and finally, Section 5 concludes.

2 Methods to measure health and to model health care use

This section describes two different methods that are used to obtain measures of health that are comparable across countries and across individuals. We present a model for health care use, which is used to assess the usefulness of the two different methods and to compare them.

2.1 The latent health method

The first method that will be investigated is a health measurement model, following Meijer, Kapteyn and Andreyeva [2011]. This model states that many health indicators, the dependent variables in the model, can be explained by an individual's (unobserved, latent) true health, which is represented as one factor that is the explanatory variable in this model. This latent health factor, called the "health index", in turn is influenced by multiple other explanatory variables, known as "predictors" in this model. The goal of the model is to obtain one health index, as a proxy for the latent health factor, that is comparable across individuals (where corrections for, for example, cultural differences and other differences in reporting styles are made) and across countries. To ensure this cross-country comparability, the parameters of the most objective health indicator are fixed and equal for each country. The most objective health indicator is chosen to be grip strength, as it is the health variable of the SHARE dataset that is assumed to be the most objective one; this will become clear later on. By fixing the parameters for grip strength, the parameters that indicate the relation of latent health with the other health indicators are relative to this one. Because the grip strength parameter is fixed to one in all countries and for all individuals, this does not only ensure comparability across countries, but also across individuals within a country; for each individual, the loadings of all other health indicators al relative to grip strength.

This health measurement model is a so-called structural equation model with a latent variable, where the latent variable is unobserved health in this case. This particular type of model is suitable, because we observe several health indicators, which are assumed to be affected by an individual's general health, which is unobserved (latent). At the same time, we assume that latent health is influenced by various factors, like age and socioeconomic status. So, we have one variable, namely latent health, that is the dependent variable in one equation, but an explanatory variable in another equation. These structural equations represent causal relationships among latent health, the health indicators and demographic and socioeconomic variables in the model. One of the strengths of structural equation modeling is the ability to construct latent variables, which is desirable in this case, because we want to know "true" individual health, but it is not directly measurable. Therefore, Meijer et al. [2011] chose to use this type of model. More detailed information on structural equation models with one or more latent variables is provided by Bollen [1989] and below, where we describe the model specification.

First of all, all used health indicators are captured in the vector y_{ci}, where c denotes the country and i the individual. The used health indicators say something about an individual's capability (or limitation) related to either mobility, an activity of daily living (ADL), or an instrumental activity of daily living (IADL). Most of the health indicators that are used here are binary; 0 if an individual does not have any difficulties with a particular indicator of mobility or (instrumental) activity of daily living, and 1 if s/he encounters some difficulty. The only exception here is grip strength, which is a continuous variable that measures the grip strength of an individual's hands and uses the maximum of all grip strength measures of that individual (it is possible that in one of the measurements, the interviewers did not obtain the full possible grip strength of an individual, and therefore the maximum is used to use the best possible grip strength of each individual).

Let n denote a particular indicator. We assume that the latent, continuous version of a health indicator, y^*_{cin}, depends structurally on an individual's unobserved (latent) true health, η_{ci}:

$$y^*_{cin} = \tau_{cn} + \lambda_{cn}\eta_{ci} + \varepsilon_{cin}, \tag{1}$$

where τ_{cn} is a country- and health indicator-specific intercept and λ_{cn} is a country- and health indicator-specific factor loading. These parameters are all country-specific, to allow for differences in reporting health difficulties across countries (for example, in some countries, individuals indicate mobility problems as soon as they encounter *some* problems, where in other countries, individuals face a much higher threshold for indicating difficulties). Further, ε_{cin} is an error term which is often referred to as the "measurement error", which is assumed to be independent and normally distributed, with variances captured in the diagonal covariance matrix Ω_c. However, the assumption of a diagonal covariance matrix (i.e., the assumption of uncorrelated measurement errors across the indicators y_{cin}) is questionable, because it is likely that the errors of measurement for the various health indicators are correlated (there might be systematic biases present in the responses of individuals) and the covariances of the measurement errors may not be equal to zero. Following Bollen [1989], however, structural equation models are more realistic in their allowance for measurement error than standard regression models, in that the error in measuring one variable is allowed to correlate with that of another. The assumption of normally distributed measurement errors is of less importance, as in light of previous research, results are often found to be insensitive to the distributional assumption and another distribution could also be assumed.

So, according to (1), latent health η_{ci} is different for all individuals in different countries, but it is the same index of general health that affects all health indicators y^*_{cin}. To allow for differences in response styles across countries, all parameters are country-specific.

As said before, the one health indicator for which the least country-specific deviations are expected, is the one that measures grip strength. Instead of asking individuals about certain difficulties with walking, kneeling, bathing and so on, grip strength is measured objectively with an electronic device and therefore cannot differ across countries due to heterogeneity in response styles. Of course, there are some country-specific deviations in grip strength possible, if for example in one country individuals do more manual work than in another country. However, considering the SHARE dataset, the grip strength variable is the variable that has the least expected country-specific deviations. Because it is expected that grip strength directly depends on an individual's height and weight, we add height and weight to equation (1) for the grip strength indicator:

$$y^*_{ci,GS} = \tau_{GS} + \lambda_{GS}\eta_{ci} + \beta' p_{ci} + \varepsilon_{ci,GS}, \tag{2}$$

where GS denotes grip strength and p_{ci} is a vector that includes height, weight, their squares and the product of height and weight. Here, the parameters τ_{GS}, λ_{GS} and β are restricted to be the same across countries. This restriction ensures cross-country comparability.

Because of the above mentioned assumption that grip strength does not have problems of cross-country differences, we use grip strength as the reference indicator in our model, as we need one indicator which is the one that all other health indicators are relative to. This means that for location and scaling purposes, for grip strength, the factor loading λ_{GS} will be normalized to 1 and the intercept τ_{GS} is set to 0, to identify the mean of the latent health index, η_{ci}. This means that the latent health index is directly linked to grip strength. By fixing the parameters for grip strength, the parameters of all other health indicators are relative to these $\lambda_{GS} = 1$ and $\tau_{GS} = 0$. This means that, for example, a 0.1 increase in latent health is linked to a 0.1 increase in grip strength in each country (except for the measurement error). This 0.1 increase in latent health can have different effects in different countries on the other health indicators, as λ_{cn} is different for each country and for each health indicator.

Again following Meijer et al. [2011], to model latent health, we assume the following linear relationship between an individual's unobserved health, η_{ci}, and demographic and socioeconomic characteristics, x_{ci}. This relationship is also referred to as the "predictive health equation" (and in this equation, latent health is the dependent variable, where it was the explanatory variable in equation (1)):

$$\eta_{ci} = \gamma_c' x_{ci} + \zeta_{ci}. \tag{3}$$

Again, all parameters are country specific. The main reasons for this are that some variables are difficult to compare across countries (for example, different countries have different education systems, and also because of institutional differences across countries). Secondly, health production functions may be different across countries, as mentioned before; the magnitude of the relation between changes in latent health and changes in indicating health difficulties may be different across countries.

The precise way in which the model is estimated is the following. We use a two-step method, where in the first step we correct grip strength for height and weight, for all countries jointly. In the second step, all other parameters are obtained for all countries of interest separately.

First, the predictive health equation (3) is inserted into the grip strength equation (2). This equation is then estimated for all countries of interest jointly, where again the intercept is set to 0 and the factor loading of latent health is set to 1. Then, the estimated height-weight polynomial is subtracted from grip strength and the residual is scaled by subtracting the mean and dividing by ten, to obtain better scaling. In this way, the grip strength residuals are small, so that they are more comparable (i.e., more in the same range) to the other health indicators, which are all binary. This makes the factor loadings for the other health indicators easier to interpret and we have to care less about approximation errors due to small parameter values.

In the second step, the obtained grip strength residual is used as reference health indicator (instead of using the raw grip strength variable that has not been corrected for height and weight). The parameters for all other health indicators are estimated by using maximum likelihood for each country separately, because of the fact that there are no joint parameters left. From these estimates, the health index can be computed by calculating the linear combination in (3) by multiplying the obtained estimates with the corresponding covariates.

In their paper, Meijer et al. [2011] not only estimate the parameters of the model country by country, but also separately for males and females. They argue that they are not able to make assumptions about equal response patterns for males and females and therefore all analyses have to be done separately. In this way, it is not possible to compare health of males with health of females.

Although indeed, further research is needed to be able to make assumptions about equal (or different) response patterns between males and females, in this thesis the analysis will be done jointly for males and females. This is simply done by adding a dummy for gender to the regression variables. Firstly, this eases optimization, as it was verified in preliminary analysis that there are many convergence problems due to relatively small sample sizes in a model of this complexity. By doing the analysis for males and females jointly, sample sizes are doubled. Secondly, for the purpose of this thesis - comparing different methods for health measures - the joint estimation makes results more comparable to results of the other method, the vignette method, where the estimation is also performed jointly for males and females. However, the other decision we could have made, is to perform the analysis of the vignette method also separately for males and females. For this method, however, there exists the same problem of relatively small sample sizes and therefore we decided to do all analyses jointly for males and females. Further research should be done to clarify whether there are substantial differences in response patterns between males and females regarding their own health. This can be done by comparing subjective responses to self-assessed health questions and controlling for more objective health measures. In this way, for example, we can see whether males are more optimistic or pessimistic about their own health, compared to females. For now, we make the assumption that there are no such significant differences.

For the same reasons of optimization problems and comparability with the results of the vignette method, in contrast to what Meijer et al. [2011] did, we also do not include respondents weights in our analysis. We do not expect the respondents weights to have a very large effect on the results. Furthermore, respondent-level weights are missing for some individuals in our particular dataset, so including them in our analysis would lead to the loss of an additional part of the sample.

2.2 The vignette method

The second method, the vignette method, is one that was originally developed by King, Murray, Salomon and Tandon [2004]. However, we build on a paper by Bago d'Uva, Lindeboom, O'Donnell and Van Doorslaer [2011a], who implemented the model. This method aims to construct six health measures, one in each of the following six health domains: *mobility problems*, *cognition problems*, *pain*, *sleeping problems*, *breathing problems* and *emotional health problems*. Respondents' self-reported scores on a five-point scale in each of the six domains are used to construct these indices, by controlling for individual characteristics and more objective health indicators. More important, the method also corrects for heterogeneous reporting behavior using vignettes. In these vignettes, respondents were asked to rate the degree of severity of health-related problems in the six different health domains that persons in the vignette stories encounter. By correcting individuals' assessments of their own health for their reporting style in the vignettes, true health effects are purged of reporting effects and more useful measures for health problems can be obtained.

The above can be done by using a hierarchical ordered probit (HOPIT) model (King, Murray, Salomon and Tandon [2004]) that makes use of these described vignette ratings. However, two assumptions have to hold in order for this HOPIT model to be usable: the assumption of *vignette equivalence* and the assumption of *response consistency*. The former assumes that "all respondents understand the vignette description as corresponding to the same level of functioning on a uni-dimensional scale" (Bago d'Uva et al. [2011a]). *Response consistency* is an assumption that says that "respondents rate the vignettes in the same way as they do their own health". The latter has to hold in order to be able to apply reporting styles derived from the vignettes to individuals' rating of their own health. Although some additional research by Bago d'Uva et al. [2011b] finds that these two vignette assumptions often do not hold in practice, we will still assume their correctness in this research.

In line with the assumption of *vignette equivalence*, we assume that the latent health level of the vignette in a certain domain does not depend on the characteristics of the respondent, but only on an indicator of the vignette and some random error. This means that when an individual rates a vignette, the latent health level in this vignette (regarding another person than the respondent) does not depend on the characteristics of the respondent. Following an ordered probit specification, the perceived latent health level of the person that is described in the vignette in domain d (where d is one of the six described health domains) for respondent i, V^*_{di}, relates to the observed ratings of the vignettes, V_{di}, in the following way[1]:

$$V_{di} = \begin{cases} 1 & \text{if} \quad V^*_{di} \leq \tau_{di,1} \\ 2 & \text{if} \quad \tau_{di,1} < V^*_{di} \leq \tau_{di,2} \\ 3 & \text{if} \quad \tau_{di,2} < V^*_{di} \leq \tau_{di,3} \\ 4 & \text{if} \quad \tau_{di,3} < V^*_{di} \leq \tau_{di,4} \\ 5 & \text{if} \quad \tau_{di,4} < V^*_{di}, \end{cases} \tag{4}$$

where the thresholds are allowed to differ across health domains and respondents; they are domain- and individual-specific. Note that there is no subscript for country, but the analyses for this method are done separately for all countries of interest, so that the effects are allowed to be different for each country.

More specifically on the thresholds, these thresholds are defined as functions of demographic and socioeconomic (W_i), as well as health (Z_{di}) covariates. The demographic and socioeconomic variables are the same for all health domains, while the health covariates are variables that differ depending on the health domain. These covariates relate to the thresholds in the following way:

$$\tau_{di,k} = W_i \phi_{d,k} + Z_{di} \theta_{d,k}, \tag{5}$$

with $k = 1, ..., 4$. In this way, we can correct for individual-specific differences in response styles, as these thresholds will differ across respondents due to different demographic and socioeconomic characteristics. So, the obtained thresholds will be such that they are, for example, lower for someone that only responds 1, 2 or 3 to the vignettes (the best, healthiest categories), compared to someone that will only respond 3, 4 or 5 (the worst categories). By applying these different thresholds, the thresholds will correct for the difference in reporting style so that ratings of respondents' own health problems will be corrected for this.

Recall that under the assumption of *response consistency*, we assume that individuals rate the health problems in the vignettes in the same way as they rate their own health problems. Therefore, the same thresholds can be used to relate individuals' responses to the categorical questions on own health problems in the six domains, H_{di}, to their unobserved latent individual ill-health, H^*_{di}, using an ordered probit specification as in (4). To be clearer, this means that individual health problems also depend on the same thresholds:

$$H_{di} = \begin{cases} 1 & \text{if} \quad H^*_{di} \leq \tau_{di,1} \\ 2 & \text{if} \quad \tau_{di,1} < H^*_{di} \leq \tau_{di,2} \\ 3 & \text{if} \quad \tau_{di,2} < H^*_{di} \leq \tau_{di,3} \\ 4 & \text{if} \quad \tau_{di,3} < H^*_{di} \leq \tau_{di,4} \\ 5 & \text{if} \quad \tau_{di,4} < H^*_{di}. \end{cases} \tag{6}$$

[1] In their paper, Bago d'Uva et al. [2011a] define the observed vignette ratings as V_{jdi}, where j denotes the jth vignette in a certain domain. They do this, because with their data, they have three different vignettes per health domain; in our data, as we will see, we have only one vignette in each domain and thus the subscript j can be left out here. See Section 3 for a further description of the data and the vignettes.

This unobserved latent individual "level of health problems", H^*_{di}, also depends on the same covariates W_i and Z_{di}:

$$H^*_{di} = W_i \gamma_d + Z_{di} \delta_d + \varepsilon_{di}. \tag{7}$$

So, the model consists of two parts: the vignette part, in which the thresholds are determined, and the part concerning the individuals' own health problems. These two parts are estimated simultaneously using maximum likelihood, where the log-likelihoods for the two parts are derived similar to a standard ordered probit model. For both parts, the same thresholds are used and the log-likelihoods are simply added up.

Under the two mentioned assumptions, latent individual indices for health problems per health domain (corrected for reporting style) can then be approximated by computing the linear combination of the covariates in (7). If we did not allow for reporting heterogeneity, that is, if we assumed that the thresholds were constant across individuals, then the coefficients in (7) were a mixture between health effects and reporting style effects. Because now we allowed for thresholds to differ across individuals, the coefficients in (7) only reflect health effects.

2.3 Health care use model

To find out how well the above two methods explain health behavior, we apply a model for health care use and include the obtained health measures from the two methods. The health care use model that is used in this paper is similar to the one that Bago d'Uva et al. [2011a] used. The aim of the model is to explain the number of doctor visits an individual made in the past year. As will be discussed in further detail in Section 3.1, the number of doctor visits is defined as the number of times an individual has seen or talked to a medical doctor in the last twelve months. This process contains two different parts, namely the decision of whether one goes to the doctor or not, and the decision of how often one visits the doctor once s/he decided to go to the doctor. Therefore, the model is called a two-part model or a hurdle model and consists of two different steps.

For the first decision, the decision whether or not to visit a doctor at all, we use a logit specification of the probability of going to the doctor. For the individuals who actually went to the doctor in the last twelve months, we use a truncated negative binomial specification to determine their number of doctor visits. As the process of how often an individual goes to a medical doctor is a count process, we might use a simple Poisson model for this. However, for the distribution of number of doctor visits we expect the sample variance to be larger than the sample mean (which is called overdispersion). In this case, the Poisson distribution is not the appropriate distribution, because the Poisson distribution has equal mean and variance. As the negative binomial distribution accounts for this larger variance by including a parameter for overdispersion, the negative binomial specification is better to use. We use the truncated version of this specification, because a negative number of doctor visits is not possible.

In this two-step model, we allow for the possibility that the zeros and the positive numbers of doctor visits come from different underlying distributions. The motivation for this particular hurdle model is that it handles excess zeros well (Cameron and Trivedi [2005]) and there is a substantial proportion of individuals in our data that did not visit a medical doctor at all.

Using the two decision processes, and letting y_i denote the number of doctor visits, following Bago d'Uva et al. [2011a], the probability of observing a given number of doctor visits can be written down as

$$f(y_i) = (1 - f_{1i})^{1-d_i}[f_{1i} f_T(y_i|y_i > 0)]^{d_i}, \tag{8}$$

where d_i is an indicator that is 1 if an individual visited a medical doctor in the past twelve months (that is, $y_i > 0$), and zero otherwise. Further, f_{1i} is the probability of going to the doctor and $f_T(y_i|y_i > 0)$ is the distribution of how many times one visits the doctor, given that $y_i > 0$. So, the first part gives us the probability of not going to the doctor, where the second part gives the probability of going to a medical doctor, and conditional on that, how many times. To be more precise,

$$f_{1i} = \Pr[y_i > 0; \beta_1|X_i] = \tfrac{\exp(X_i\beta_1)}{1+\exp(X_i\beta_1)}, \tag{9}$$

where X_i is a vector that includes demographic and socioeconomic characteristics, as well as health variables, for example the ones obtained from the previously discussed two methods. In (8), $f_T(y_i|y_i > 0)$ is the part following a truncated negative binomial specification:

$$f_T(y_i|y_i > 0) = \frac{f_{2i}(y_i)}{1 - f_{2i}(0)}, \tag{10}$$

where $f_{2i}(y_i)$ is a negative binomial specification that is truncated at zero in (10). The negative binomial specification is defined as follows (see also Cameron and Trivedi [2005]):

$$f_{2i}(y_i) = \Pr[y_i; \alpha, \beta_2|X_i] = \tfrac{\Gamma(\alpha^{-1}+y_i)}{\Gamma(\alpha^{-1})\Gamma(y_i+1)} \left(\tfrac{\alpha^{-1}}{\alpha^{-1}+\exp(X_i\beta_2)}\right)^{\alpha^{-1}} \left(\tfrac{\exp(X_i\beta_2)}{\exp(X_i\beta_2)+\alpha^{-1}}\right)^{y_i}, \tag{11}$$

where $\Gamma(.)$ is the gamma function and $\alpha > 0$ denotes a parameter for flexibility, that is, a parameter that allows for greater or smaller variability in the model. In our case, we expect the variance to be larger than the mean, and thus we expect the presence of overdispersion.

To obtain the log-likelihood function of the hurdle model for doctor visits, we take the logarithm of the probability of observing a number of doctor visits, as in (8), and add over all individuals. The log-likelihood function then becomes:

$$\log(\ell) = \sum_{\{d_i=0\}} \log(1 - f_{1i}) + \sum_{\{d_i=1\}} [\log(f_{1i}) + \log(f_T(y_i|y_i > 0))], \tag{12}$$

where the first summation is over all individuals that did not visit a medical doctor (and hence we do not have to take into account the truncated negative binomial part for the number of visits) and the second summation is over all individuals that visited a medical doctor at least once. We assume independence of the two processes (the logit part and the truncated negative binomial part), and therefore the two parts can be estimated separately by means of maximum likelihood and the log-likelihoods of the two models can simply be added, as is shown in (12).

This hurdle model is widely used (Cameron and Trivedi [2005]), and, when used with the negative binomial specification, it is quite flexible, compared to when it is used with, for example, the Poisson specification. In particular, the negative binomial specification works well for doctor visits (Cameron and Trivedi [2005]). A large drawback of this model is, however, according to Cameron and Trivedi [2005], that it is not very parsimonious, as by implementing these two steps instead of just one, the number of parameters is doubled. In a simple one-step model, we would just have one vector β to estimate. Here, we would assume that the same factors influence both the decision of whether or not to visit a medical doctor, and the decision of how often to go to a medical doctor, given that an individual goes. However, in the two-step model we use, we allow for different factors to influence these two decisions. To give an intuitive example: the decision to visit a medical doctor can be based on health conditions, because bad health eventually leads to the decision to visit a doctor. The number of times that an individual visits a doctor, however, could be affected by factors of wealth and income, as an individual with a low income does not have the ability to visit a medical doctor as often as an individual with a higher income does. To allow for these different influences, in the two-step model we have to estimate both β_1 and β_2.

3 Data

The data are taken from the Survey of Health, Ageing and Retirement in Europe (SHARE)[2]. This is a cross-national biennial survey of individuals living in several countries in many parts of Europe who were aged 50 or over when first interviewed, and their spouses. Individuals were interviewed on several topics, including social and demographic background, physical and mental health, health care, employment and financial situation. The SHARE data is very useful for cross-country comparisons, both because so many domains are covered in the survey and because many different countries participate in SHARE. Next to comparing the latent health method with the vignette method, we also want to compare our results to earlier results by Meijer et al. [2011] and Bago d'Uva et al. [2011a], who all used the first wave of SHARE. Therefore, for this research, the second wave of SHARE is used, for which data was collected in 2005-2006 from over 34,000 Europeans living in thirteen different countries. The SHARE dataset is available on the SHARE website to all registered users.

For our particular research, the SHARE data is used because of several reasons. First of all, the SHARE is a survey that is exactly the same in many countries within Europe. Therefore, it is fairly easy to do cross-country comparisons, as exactly the same data is available for several countries, and in each country, between 2,000 and 3,000 individuals are interviewed. Secondly, the SHARE dataset consists of personal information on many different topics. Of course, in this research the field of health is the most important one, but we also want to control for many different demographic and socioeconomic characteristics, which is all available on respondent-level in SHARE. Furthermore, concerning the information on health, SHARE allows us to not only use objective health measures such as symptoms or diseases that are diagnosed by a medical doctor; in SHARE, there is also a lot of subjective information on individual health available, information that cannot easily be found in medical records (for example, the way an individual feels, cognitive status, etc.). This subjective information gives us more insight in individual "true" health, which is much more than just diagnoses.

However, using survey responses as data also has disadvantages. The SHARE is a survey that consists of hundreds of questions, and respondents have the choice of refusing to answer. This means that for some questions/variables, there are many missing data points. In analyses, we have to delete such respondents from our dataset. A second disadvantage is that the vignettes that we use in the vignette method are only asked to a third of the respondents. This reduces our sample size dramatically. Until now, however, the SHARE data is the largest European survey on health and therefore it is very suitable for this thesis.

3.1 Variables in our analysis

Table 8 in the Appendix gives an overview of all variables that are used in the latent health method. Most of these variables are similar to the ones that Meijer et al. [2011] used. The first part of this table gives an overview of the 24 used health indicators, the ones that are collected in the vector y_{ci}. All of these indicators, except for the grip strength residual,

[2] Condition of use of the SHARE dataset: "This thesis uses data from SHARE release 2.5.0, as of May 24th 2011. The SHARE data collection has been primarily funded by the European Commission through the 5th framework programme (project QLK6-CT-2001- 00360 in the thematic programme Quality of Life), through the 6th framework programme (projects SHARE-I3, RII-CT- 2006-062193, COMPARE, CIT5-CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th framework programme (SHARE-PREP, 211909 and SHARE-LEAP, 227822). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01 and OGHA 04-064, IAG BSR06-11, R21 AG025169) as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions)."

are binary; 1 if a respondent encounters any difficulty with the activity and 0 otherwise. The indicators are divided into difficulties with mobility, difficulties with activities of daily living (ADL) and instrumental activities of daily living (IADL). The grip strength residual is computed as explained in Section 2.1. There is one health indicator, however, that Meijer et al. [2011] used in their analysis and that we do not use in ours; self-assessed health. There were two main reasons for why we did not include self-assessed health as a health indicator in the health measurement model. The first reason is that in preliminary analysis, we found that due to relatively small sample sizes and many health indicators, inclusion of a categorical health indicator led to problems with optimization. Therefore, except for grip strength which needed to be included because it is the reference variable, we decided to only include binary health indicators. A second reason is that, as already said, we already have many health indicators, which means many parameters in the latent health method, as well as in the health care use model. Inclusion of the self-assessed health variable would lead to, next to an additional factor loading (λ_{cn}), four additional threshold parameters because of the five response categories of self-assessed health. Therefore, we decided to exclude self-assessed health from the health measurement model.

The second part of the table gives an overview of the variables that are included in x_{ci} in equation (3), besides a constant. These variables are basically the same as the ones used in the article by Meijer et al. [2011]. The only difference is that we added a dummy for gender, as they did their analyses separately for males and females whereas we do a joint analysis for both sexes. They included a third-degree age polynomial, educational achievement (secondary and tertiary education, with primary or no education as reference category),[3] household size and living with a spouse or partner.

To have some measure of a respondent's wealth, household net worth is included, which is adjusted for the difference in purchasing power of money across countries and over time. This information is also included, because earlier research found that socioeconomic status and individual wealth are positively related to individual health (see, for example, Meer, Miller and Rosen [2003]). Because some households do have a negative net worth, we cannot simply use the logarithm of household net worth, and therefore Meijer et al. [2011] chose to use the inverse hyperbolic sine (IHS) of net worth: $\text{IHS}(x) = \log(x + \sqrt{1 + x^2})$. This gives approximately the same result as the logarithmic transformation for those who have a positive household net worth not close to zero.

The final variables in the table correspond to an individual's body mass index (BMI), where BMI is calculated as the weight of a person in kilograms divided by the square of his/her height in meters. Some dummies on different BMI categories are included; these are relative to individuals with a normal, considered healthy weight ($18.5 \leq \text{BMI} < 25$). Because for a reasonable number of individuals BMI could not be computed due to missing height and/or weight, we also included a dummy for missing BMI. One could question whether we should include BMI as an explanatory variable, given that it is clearly health-related and thus could also be on the left-hand side of equation (1) as a health indicator. The reason why we included BMI as an explanatory variable, is because obesity is expected to have (negative) effects on physical health and thus on the probability of documenting difficulties with mobility or (I)ADLs.

[3] To make the levels of education cross-country comparable, the SHARE dataset makes use of the 1997 International Standard Classification of Education (ISCED-97) as designed by the United Nations Educational, Scientific and Cultural Organization (UNESCO; see United Nations Educational, Scientific and Cultural Organization [2006] for details on ISCED-97 coding) as coding for education; this coding ranges from 0 (pre-primary education) to 6 (upper tertiary education).

Table 9 in the Appendix lists all used variables for the vignette method. First of all, as already explained, respondents rate problems with their own health in six different domains, where they have to indicate how much difficulties they experience when doing activities related to the six health domains. Their responses are on a five-point scale, ranging from "no difficulties" to "extreme difficulties". During the interview, they also get six stories about non-existing persons (one story in each health domain) and they have to rate the difficulties that these non-existing persons encounter from their view, using the same five-point scale. For example, in the domain *mobility problems*, they get the following story and question:

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

which they then have to answer with either *"1. None"*, *"2. Mild"*, *"3. Moderate"*, *"4. Severe"* or *"5. Extreme"*.

Following Bago d'Uva et al. [2011a], we use the variables in the second part of the table as explanatory variables in both the thresholds and in the health equation. The first few variables, on age, gender and education, are the same variables for all six health domains and are also closely related to the first explanatory variables that are used in the method by Meijer et al. [2011]. The other variables are the ones that differ according to the health domain of the vignette and the self-assessed health question. These variables are similar to the variables Bago d'Uva et al. [2011a] used; they followed expert medical advice and so these variables are assumed to really affect the degree of difficulties in a particular health domain. We did not use all of their health indicators, as the main difference between this thesis and their article is, as already noted, that this thesis only has one vignette available in each health domain, instead of three, which makes the optimization process more difficult in our research. That is, Bago d'Uva et al. [2011a] have three vignettes in each health domain to correct for heterogeneity in response styles, where we only have one third of the information on response styles that they have. The consequence is that we use slightly less health indicators, because with many health indicators, it becomes more difficult to find relations between all the health indicators and reporting heterogeneity. In each domain, we chose some of the most important health indicators, which are mostly dummy variables.

For the model for health care use, as dependent variable we use the number of visits to a medical doctor (general practitioner and/or specialist) in the past twelve months. As a visit, we both count a physical visit to a medical doctor and a conversation with a medical doctor (can also be by telephone).

As explanatory variables, we use demographic and socioeconomic variables that are similar to the aforementioned variables. To be more specific, we use variables that do closely match the ones in a research by Majo [2010]. We use the variables from this specific research, because Majo [2010] argues that these covariates are in line with the factors usually considered to explain the demand for health care. This includes age, gender, living with spouse or partner, household net worth, employment status and education. The precise variables can be found in Table 10 in the Appendix. Additionally, we include several different health variables each time the model is estimated; this is explained in further detail in Section 4.3.

3.2 Sample selection

Although the latent health method and the vignette method have never been compared before, the basic analysis of the two methods is already done for many different countries in previous research (although it is using wave 1 of the SHARE data). Therefore, for implementing the second wave of SHARE and for comparing the different methods of obtaining health measures, which is the primary focus of this thesis, we do not include all available countries in our analysis. Particularly, we only make use of the Netherlands and Spain. We chose these particular countries because of known differences in various socioeconomic characteristics, such as level of educational attainment and household income (both expected to be higher in the Netherlands). Also, health and the use of health care is expected to be different in these two countries, where we expect that individuals in the Netherlands are, on average, more healthy than Spanish individuals and thus are less in need of health care.

We deleted individuals that have missing information on almost all demographic, socioeconomic and health variables. For the Netherlands, the remaining sample consists of 2661 individuals, where the sample is a little bit smaller for Spain, with 2228 individuals.

The method by Bago d'Uva et al. [2011a] makes use of vignettes. Unfortunately, these vignettes are only offered to a random third of the total sample, so that the sample sizes for this analysis are smaller. In particular, after deleting respondents with a lot of missing information, the vignette sample sizes are equal to 523 for the Netherlands and 517 for Spain.

For the health care use model, we want to be able to compare different models (including different health measures), and therefore sample sizes become even smaller. In the sample sizes mentioned above, individuals that have missing information on, for example, only one health variable are still included, because they can be used in the latent health method or in the vignette method. They can, however, not be used in all of the implemented models for health care use, because in this model we use additional information on health that is not used in either the latent health method or the vignette method. Individuals with missing information on at least one used variable in the model for health care use, will therefore be deleted in Section 4.3. Sample sizes then reduce to 2390 for the Netherlands and 1757 for Spain in the "full" sample, and 471 for the Netherlands and 424 for Spain in the vignette sample.

3.3 Descriptive statistics

Table 1 shows descriptive statistics for the main variables. Average age is approximately 2.5 years higher in Spain than it is in the Netherlands, where it is close to 64 years on average. In both countries just over half of the sample is female and in the Netherlands, average household size is close to two, where in Spain this number lies a bit higher (2.6). In both countries, over three quarters of the used sample is living with his/her spouse or partner. There exists a huge difference in average level of educational attainment between the two countries. In the Netherlands, about 60 percent of the respondents finished secondary education (i.e., they finished high school) and more than 20 percent has followed even tertiary education, where in Spain not even 30 percent finished secondary education and less than ten percent followed tertiary education. This is in line with what we expected, as about 50 years ago, Spanish people did not have as much opportunities for higher education as Dutch people did. Assuming that more education leads to a higher chance of having a job, the difference in level of education is also reflected in employment percentages in the sample; in

Table 1: Descriptive statistics of main variables

	Netherlands			Spain		
Variable	N	Mean	St.D.	N	Mean	St.D.
Demographic/socioeconomic						
Age	2660	63.9	9.9	2223	66.6	11.0
Female (%)	2661	54.5		2228	55.0	
Household size	2661	2.1	0.8	2228	2.6	1.1
Living with spouse or partner (%)	2661	80.3		2227	77.6	
Secondary education (%)	2661	60.6		2228	28.5	
Tertiary education (%)	2661	23.4		2228	8.8	
Employed (%)	2565	32.3		2183	21.0	
Household net worth (real, in 10,000 €)	2661	38.3	93.0	2228	35.3	76.3
Health						
Number of doctor visits	2636	4.8	8.0	2181	8.2	11.3
At least one limitation: mobility (%)	2646	36.9		2225	49.5	
At least one limitation: ADL (%)	2646	6.7		2225	13.0	
At least one limitation: IADL (%)	2646	14.1		2225	20.4	
Grip strength (kg)	2515	36.0	11.5	1927	30.0	11.6
Height (cm)	2598	171.7	9.1	2112	163.0	8.5
Weight (kg)	2629	77.0	14.4	2151	73.4	13.3
Obese (%)	2586	14.0		2083	24.3	
Self-assessed health: fair/poor (%)	2646	29.2		2225	46.6	
Mobility problems: severe/extreme (%)	522	5.6		517	10.6	
Cognition problems: severe/extreme (%)	523	2.7		517	8.7	
Pain: severe/extreme (%)	522	6.7		517	15.9	
Sleeping problems: severe/extreme (%)	523	4.8		517	13.2	
Breathing problems: severe/extreme (%)	521	1.9		516	4.7	
Emotional health pr.: severe/extreme (%)	523	1.0		516	14.0	

Note: (I)ADL: (instrumental) activity of daily living.

Spain, the proportion of individuals that is still employed is over ten percent lower than in the Netherlands. Average household net worth is a bit higher in the Netherlands, but the amounts in euros vary a lot (from almost -10 million to almost 18 million).

The second part of the table contains variables on health. First of all, on average, Spanish people have 3.5 visits to a medical doctor more in a year than people in the Netherlands have. The fact that Spanish individuals go to a doctor more often seems logical, as a much larger proportion of the Spanish individuals reports difficulties on mobility, ADLs and IADLs. Also, their grip strength is lower and a larger proportion is obese. The same pattern is visible for the self-rated questions. For each of these questions we report the proportion of individuals who answered one of the two poorest categories. For the general self-assessed health question, this is everyone who reported to be in less than good health: *fair* or *poor*. For the self-rated questions on health problems in the six health domains, we report the proportion of individuals who answered *severe* or *extreme* to the question how many difficulties they encounter in each of the six domains. For these six questions, a smaller sample is used, as only the vignette sample answered these questions. For all self-rated questions, the proportion of individuals reporting to be in poor health is substantially larger in Spain than in the Netherlands. This may explain for a large part why Spanish people visit the doctor more often.

4 Results

We first focus on the results from our first method, the latent health method, followed by the results from the second method, the vignette method. After that, we compare the two methods by applying a model for health care use and including the different health measures.

4.1 The latent health method

One of the objectives of this thesis is to compare the results of the method by Meijer et al. [2011] to the results of the same method when using the second wave of SHARE, instead of the first wave. Are the results the same for the two waves, with two years of time in between? As many individuals in wave 2 were also interviewed in the first wave of SHARE, most of the data should be more or less the same for the two waves. However, within a time frame of two years (between the first and second wave of SHARE), a lot can happen with respect to individual health conditions. Are the results comparable for the two different waves, or do they show large differences? For this objective, we reproduce and interpret tables that are similar to the ones in Meijer et al. [2011] to be able to compare the results.

As said before, in this thesis the analysis of the latent health method is done for males and females jointly, and without using respondents weights. First of all, we estimated the grip strength equation (the first step of the estimation) to obtain the grip strength residual. We do not present all the results of this step here, as these results are not the ones that we want to focus on. However, we provide some intuition. In Table 1, we have seen that there is a difference in grip strength of about six kilograms on average between the Netherlands and Spain. At the same time, individuals in the Netherlands are on average almost nine centimeters taller and 3.5 kilograms heavier. In the grip strength equation (equation (2)), we corrected grip strength for height and weight. The resulting grip strength residual has an average of 0.131 for the Netherlands and -0.252 for Spain, on a scale ranging from approximately -4 to 4 for both countries. This means that the apparent difference in grip strength between the Netherlands and Spain is not only due to height and weight differences, which is corrected for by computing the grip strength residual. It seems that, after controlling for height and weight, Spanish individuals are still weaker than Dutch individuals. We will use this height/weight-corrected grip strength residual in the remainder of the analysis.

Table 2 gives the estimation results for the intercept, τ_{cn}, and the factor loading, λ_{cn}, for each health indicator for both the Netherlands and Spain. We left out the t- and p-values to save space, as all t-values were more negative than -4 (generally between -10 and -20), indicating that all coefficients have a p-value that is smaller than 0.01. So, all obtained coefficients are significant.

The first two columns give estimates of the intercepts. The first thing to be noticed from these intercepts is that they are all negative. Recall that all coefficients are relative to the coefficient for the residual grip strength, which was set to zero for the intercept. These negative intercepts reflect the fact that only a small number of individuals experiences difficulties with the various health indicators: the more negative an intercept is, the less likely individuals are to report mobility difficulties. When we compare the Netherlands and Spain, the intercepts for the ADLs and IADLs are much lower in Spain, indicating a higher threshold for reporting a difficulty in this country. This means that, compared to the Netherlands, Spanish individuals start off from a smaller probability of indicating a problem with an ADL or IADL, given that they have the same health level. However, the different

Table 2: Estimates of intercepts ($\hat{\tau}_{cn}$) and factor loadings ($\hat{\lambda}_{cn}$), latent health method

	Intercepts ($\hat{\tau}_{cn}$)		Factor loadings ($\hat{\lambda}_{cn}$)	
Indicator	Netherlands	Spain	Netherlands	Spain
Mobility				
1) Walking 100 meters	-2.330	-3.675	-4.363	-3.709
2) Sitting two hours	-1.677	-2.487	-2.152	-2.025
3) Getting up from chair	-1.343	-2.344	-3.300	-2.993
4) Climbing several flights of stairs	-1.416	-1.973	-4.224	-3.225
5) Climbing one flight of stairs	-2.077	-3.032	-4.109	-3.336
6) Stooping, kneeling, crouching	-1.077	-2.219	-3.644	-3.476
7) Reaching arms above shoulder	-1.694	-2.833	-2.006	-2.686
8) Pulling/pushing large objects	-2.587	-2.775	-4.983	-3.861
9) Lifting/carrying weights > 5 kg	-1.256	-2.381	-3.765	-3.504
10) Picking up small coin from table	-2.366	-3.297	-2.204	-2.452
ADL				
11) Dressing	-2.698	-4.497	-3.832	-4.324
12) Walking across a room	-4.816	-6.412	-6.228	-5.059
13) Bathing/showering	-3.787	-7.379	-5.920	-7.174
14) Eating, cutting up food	-3.267	-5.547	-3.313	-4.176
15) Getting in or out of bed	-3.347	-6.237	-4.280	-5.400
16) Using the toilet	-4.484	-6.935	-5.356	-5.647
IADL				
17) Using map in strange place	-1.851	-3.111	-2.133	-3.016
18) Preparing hot meal	-2.912	-7.735	-3.969	-7.115
19) Shopping for groceries	-3.642	-8.946	-5.839	-8.684
20) Telephone calls	-2.905	-5.982	-2.858	-5.054
21) Taking medications	-3.558	-7.216	-3.617	-6.159
22) Doing work around house/garden	-2.255	-4.530	-4.957	-4.972
23) Managing money	-3.301	-5.800	-4.263	-5.264
24) Residual grip strength	0	0	1	1
Measurement error s.d.				
Residual grip strength	0.785	0.720		
Number of observations	2660	2222		

Note: All coefficients are significant at the 1%-level (p<0.01).

intercepts within a country are more or less in the same order for both the Netherlands and Spain. That is, for instance, looking at the IADLs, the intercept for *using a map in a strange place* is the highest among the IADLs in both countries, where the intercept for *shopping for groceries* is the lowest in both countries.

These intercepts are similar to the results of Meijer et al. [2011]. Although they do not present the intercepts for each country but only show the minimum and maximum for each intercept across all countries, they also find strongly negative and significant intercepts. The ordering of the intercepts is also more or less the same. For example, in their analysis the intercept for *using a map in a strange place* is also the highest among all IADLs and *shopping for groceries* has, at least for males, the most negative intercept. Table 2 also gives the estimated standard deviations for the grip strength equation; these are also comparable to each other and to the results by Meijer et al. [2011]. Meijer et al. [2011] found standard deviations between a minimum of 0.687 and a maximum of 0.945 for males, and between 0.545 and 0.659 for females. This all shows that the intercepts for both waves of SHARE (in the year 2004 for Meijer et al. [2011], and 2006 for this thesis) are very similar.

The last two columns of Table 2 give the factor loadings, λ_{cn}, for both countries and for all health indicators. The latent health variable is defined to be low when an individual is unhealthy, and higher when s/he is more healthy. Because the factor loading of the grip strength residual is restricted to be equal to one, this means that better health is associated with a higher grip strength. All other health indicators, however, have an opposite relation with health: for the difficulties with mobility and (I)ADLs, a higher number (a one instead of a zero) is associated with worse health. Therefore, we expect all other factor loadings to be negative. Indeed, all factor loadings are negative, which suggests that a better underlying health leads to less difficulties with mobility and (I)ADLs. Also, as mentioned before, all obtained factor loadings are significant with t-values that are generally more negative than -6. Comparing across countries, the factor loadings for mobility are a little bit more negative for the Netherlands, whereas the factor loadings for (I)ADLs are generally more negative for Spain. The latter means that a similar increase in health in both countries leads to a larger decrease in the probability to indicate difficulties with (I)ADLs in Spain, compared to the Netherlands.

Meijer et al. [2011] show similar results for the factor loadings; they also found statistically significant negative factor loadings for all countries and for both genders. Again, we cannot see whether we obtain the same patterns for the Netherlands versus Spain, as they only show minimum and maximum factor loading values across all countries in their paper.

Table 3 gives the estimates of the predictive health equation for the latent health variable η (equation (3)) for both countries (the parameters are constant across health indicators). According to the first (linear) age variable, higher age is associated with lower (and thus poorer) health. In both countries females are likely to be in worse health than their male counterparts, although this relation seems to be a bit stronger in Spain than in the Netherlands. Further, higher education and higher wealth (indicated by household net worth) both have a significant positive effect on health. Living with a spouse or partner does not seem to have a significant influence on health. Relative to people with a normal weight ($18.5 \leq \text{BMI} < 25$), being underweight or being obese in the two most extreme categories are, as expected, the strongest negatively related to health out of all weight categories. The effects are more or less the same for both the Netherlands and Spain. The results are also consistent with the results of Meijer et al. [2011]: they also found significant effects for age, education, household net worth and extreme under- or overweight, mostly in the same directions as we found (again, Meijer et al. [2011] only provide minimum and maximum coefficients across al countries).

After estimating the model for two countries separately, the obtained health indices (obtained by multiplying the estimates in Table 3 with the corresponding variables and adding them up) are assumed to be cross-country comparable. Therefore, the estimated means and standard deviations of latent health in Table 4 can be directly compared. The numbers are consistent with what is widely known: average health is better among elderly Dutch individuals, compared to elderly Spanish individuals. This is consistent with higher average income and wealth in the Netherlands, compared to Spain. Spain has a larger standard deviation and therefore more variation in health. The numbers in this table are similar to the numbers

Table 3: Estimation results for predictive health equation (γ_c), latent health method

Predictor	Netherlands	Spain
$(Age - 65)/10$	-0.086***	-0.209***
$[(Age - 65)/10]^2$	-0.027***	0.035***
$[(Age - 65)/10]^3$	-0.006*	-0.020***
Female	-0.070***	-0.160***
Secondary education	0.082***	0.058**
Tertiary education	0.095***	0.129***
Missing education	-0.072	0.016
Household size	0.021	-0.003
Living with spouse	0.031	0.009
Household net worth	0.014***	0.011***
Underweight	-0.229***	-0.369**
Overweight	-0.023***	-0.042
Obese class I	-0.081**	-0.134***
Obese class II and III	-0.231***	-0.241***
Missing BMI	0.030	-0.149***
Constant	-0.113*	-0.333***
Residual s.d.	0.352***	0.405***

Note: * p<0.10; ** p<0.05; *** p<0.01.

Table 4: Estimated distribution of latent (true) health η and reliability of the health index

Country	Mean	St.Dev.	Reliability
Netherlands	0.106	0.390	0.70
Spain	-0.336	0.515	0.83

that Meijer et al. [2011] found. For the Netherlands, they found a mean of 0.15 for males and 0.05 for females. For Spain, they found a mean of -0.41 for males and -0.32 for females. Averaging the numbers of Meijer et al. [2011] over both genders per country (because our analysis was done jointly for males and females), we find means that are very close to the means that we found. The same holds for the corresponding standard deviations.

To give some idea of how well the health index reflects true health, Meijer et al. [2011] computed the squared correlation between the health index ($\hat{\eta}_{ci}$) and true health (η_{ci}), or as they call it, the R^2 of the hypothetical regression of η_{ci} on $\hat{\eta}_{ci}$. In their paper, they derive how they compute this squared correlation, also called the *reliability* of the health index, even though true health is unknown. This reliability is expressed as 1 - $\text{Var}(\eta_{ci}|y_{ci}, x_{ci})/\sigma^2_{\eta,c}$, where $\text{Var}(\eta_{ci}|y_{ci}, x_{ci})$ is the conditional variance of η_{ci}, which is, according to Meijer et al. [2011] asymptotically equal to the mean-squared prediction error of $\hat{\eta}_{ci}$. Furthermore, $\sigma^2_{\eta,c}$ is the unconditional variance of η_{ci}: $\sigma^2_{\eta,c}$ $=\text{Var}(\zeta_{ci}) + \gamma_c'\text{Cov}(x_{ci})\gamma_c$, which can be obtained from x_{ci} and the parameter estimates. This reliability for our estimates is given in the last column of Table 4. With 0.70 and 0.83 for the Netherlands and Spain, respectively, the reliabilities are of the same order as the numbers found by Meijer et al. [2011] (the Netherlands: 0.73 for males and 0.83 for females, Spain: 0.81 for males and 0.88 for females). These numbers show us that the information that is in the health indicators (y_{ci}) has much explanatory power additional to the other covariates (x_{ci}) for latent health.

To summarize, the latent health method obtains coefficients that are consistent with expectations. The results we found are also similar to the results in the paper by Meijer et al. [2011]. In terms of reliability, this method constructs an index that is highly related to individual true (unobserved) health. Whether this health index performs well in explaining and forecasting, will become clear in Section 4.3.

4.2 The vignette method

In this section, we use the second wave of the SHARE data in the so-called vignette method. Bago d'Uva et al. [2011a] also used this method, but they used the first wave of the SHARE dataset to obtain their results. Although they do not explicitly interpret the results of the HOPIT model in their paper, we will do this here to make the results of the model better understandable.

First of all, as already mentioned before, there is a major difference between the analysis by Bago d'Uva et al. [2011a] and this analysis. That is, in their paper, using the first wave of SHARE, there were three available vignettes per health domain, where we only have one vignette per domain, with very few individuals responding one of the worst two categories (*"4. severe"* and *"5. extreme"*). Due to very few cases in these two categories, there is a large chance that, for example, all individuals that answered *"extreme"* to a vignette, all indicated 1 to a binary explanatory variable. These occurrences make it very difficult to obtain estimation results with only one vignette. Therefore in all health domains, for both countries, we combined these two categories. This means that for the HOPIT model, we have four categories left and thus three individual-specific thresholds ($\tau_{di,1}$, $\tau_{di,2}$, $\tau_{di,3}$).

As explained in Section 2.2, the model is estimated six times, for the six different health domains, and each of these six domains has to be estimated twice, namely for the Netherlands and for Spain. This gives a total of twelve models, with a large amount of resulting coefficients. As the aim of this paper is to compare the usefulness of the two methods and not to find determinants for response styles in the different health domains using vignettes, we will not discuss the results for all of the twelve models. Particularly, we will only present and discuss the resulting coefficients for one of the six domains, for one country, to make the results better understandable.

We will focus on the results for the health domain *mobility problems* for the Netherlands. In this analysis, all demographic and socioeconomic variables in Table 9 in the Appendix are included, as well as the health variables in the domain of *mobility problems* that are mentioned in this table. Table 5 shows the results.

The first three columns in this table give the coefficients for the three thresholds. The first thing to notice here is that most coefficients are not significant. That is, the coefficients for most of the thresholds do not significantly differ from zero. However, we do not only want to know how far the coefficients lie from zero, but more importantly, how far apart the thresholds are located relative to each other. Therefore, the fourth column of the table gives χ^2-statistics for the test whether $\hat{\tau}_1 - \hat{\tau}_3$ differs significantly from zero.

For gender, there is a monotonically decreasing effect on the thresholds, going from τ_1 to τ_3, with τ_1 being positive and τ_3 being negative. This means that females, compared to males, on average have a narrower view when it comes to assessing another person's mobility problems (the person in the vignette). Females have a larger chance to rate the vignette with *"1. no difficulties"*, as for them the threshold for answering *"2. mild"* is higher. They will also sooner rate an individual to be in the worst category (*"4. severe"* and *"5. extreme"*

Table 5: Estimates HOPIT model, domain *mobility problems*, the Netherlands

Covariate	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\chi^2[\hat{\tau}_1 - \hat{\tau}_3]$	$\hat{H}_i^*$
W_i	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$		$\hat{\gamma}$
Age	-0.001	0.010	0.001	0.02	0.010
Female	0.422**	0.226	-0.029	3.39*	0.087
Lower sec. educ.	-0.011	-0.011	0.556**	3.73*	0.184
Upper sec. educ.	0.036	-0.096	0.360	1.02	0.022
Tertiary educ.	-0.267	-0.375*	-0.045	0.51	-0.160
Constant	-1.878**	-0.595	1.828**	11.21***	-2.131*
Z_{di}	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$		$\hat{\delta}$
Obese	-0.440	0.013	-0.107	1.27	0.581**
Grip strength	0.066	-0.022	-0.258***	7.70***	-0.175
Limitations	-0.047	0.043	0.059	2.41	0.512***
Diagnoses	0.143	0.254	0.340	0.37	0.744***
Symptoms	0.000	-0.288	-0.178	0.37	0.122

Note: The first three columns give estimates for the thresholds $\hat{\tau}_{mob,k}$. The fourth column gives test results for whether there is a difference between $\hat{\tau}_1$ and $\hat{\tau}_3$. The last column gives estimates for the own-health equation. * p<0.10; ** p<0.05; *** p<0.01.

combined) compared to males, as τ_3 will on average be a lower number for females. So, on average, for females the three thresholds will be closer to each other, implying a larger probability to answer one of the outer categories (1 or 4/5). To test whether these closer thresholds for females are significantly different from males, we use the χ^2-statistic of the test whether τ_1 and τ_3 differ significantly from each other. It indicates significance at the ten percent level. This means that, using a ten percent level of significance, there is a statistically significant difference in thresholds for females versus males.

There is an opposite relation visible for individuals with lower secondary education (relative to individuals in the lowest category of education, no/primary education). They have, on average, thresholds that are located further away from each other, giving them a larger probability to answer one of the middle categories (2 or 3). This difference in thresholds compared to individuals with a lower level of education is significant at the ten percent level. Furthermore, we see an overall shift to the left for all three thresholds for individuals in the highest education categories, tertiary education. This shift is, however, not statistically significant. The coefficients for grip strength show a pattern of thresholds that are moving closer to each other, as grip strength becomes a higher number. So, a stronger grip strength leads on average to narrower thresholds, such that stronger individuals are more likely to rate a vignette with one of the two extreme options (1 or 4/5). This pattern is significant at the one percent level. All other health variables in this model do not lead to a significant difference in the location of thresholds τ_1 and τ_3, relative to each other.

Now the question is, using these thresholds, which factors do influence an individual's own mobility problems. From the last column of Table 5, it is clear that the demographic and socioeconomic variables do not significantly influence this assessment. More important are the health variables; a positive and significant coefficient of 0.581 for obesity means that, if an individual is obese, s/he is likely to have worse mobility. The same holds for having more limitations with activities of daily living and for having diagnoses related to mobility problems. So, these numbers are (if significant) in the expected direction.

Although we do not present the results for the other five domains and also not for Spain, we provide some intuition on the results for all six domains in both countries.

For the thresholds of the vignettes, in general age is not very important, although thresholds seem to be moving towards each other once an individual gets older. This means that an older individual has a larger probability of answering one of the two extreme categories for the vignettes (either 1, or 4/5). However, in most cases, this relation is not significant. Overall, females in the Netherlands are also likely to show this tendency of having a larger chance of answering either 1 or 4/5 to the vignettes. In many of the domains in the Netherlands this difference between τ_1 and τ_3 is significant, but it is generally not in Spain. Level of education does not play a significant role. For the diagnoses corresponding to a particular health domain it holds that, if one is diagnosed with related symptoms or diseases, the thresholds are likely to move away from each other, with τ_1 shifting to the more negative side and τ_3 shifting towards the positive side. This implies that, once one is bothered by the symptoms him/herself and thus knows how it feels, s/he is less likely to rate the vignette in its extremes and more likely to respond 2 or 3. Although the coefficients for the thresholds are mostly not significant themselves, the difference between the lowest and the highest threshold is significant in many cases. For the assessment of own-health problems (the last column of Table 5), in general the health variables associated with each domain, together with age, are the most important indicators. In some domains, gender is also a significant determinant. Surprisingly, in Spain, education is more important in the self-assessment of health problems than it is in the Netherlands. In most domains in Spain, education plays a significant role, where higher education leads to a healthier self-assessment of a particular domain. Apparently, in Spain education is stronger associated with an individual's rating of his/her own health problems than in the Netherlands. This is surprising, as we did not find such education-related differences between the Netherlands and Spain in the latent health method. Not surprisingly, the most significant determinants of own health are, like already shown in Table 5, the health variables for the corresponding health domain.

Concluding, many results are not significant, although testing for the difference between the lowest and the highest threshold reveals some significant effects. Results are more different between the Netherlands and Spain than they were in the latent health method (Section 4.1), but this may also be due to the smaller number of observations we have available for the vignette method. Therefore, it is difficult to find determinants, although there are some patterns related to age, education (for Spain) and domain-related health diagnoses.

4.3 Health care use model

In this section, the model for health care use is applied several times with different variables. As mentioned already in Section 3.1, in all models we include controls for demographic and socioeconomic characteristics of individuals. These are the variables as listed in Table 10. Additionally, because we want to explain the number of doctor visits, we include health variables. Each time we estimate the parameters of the model, we add different variables on health. These health variables can be the health indices that result from the latent health method and the vignette method, or other general health variables. In this way, we can compare the explanatory power of the health indices from the two different methods. We can both compare the explanatory power of these indices to each other, as well as to other health variables.

Table 6: Estimates β_1 (logit) and β_2 (truncated negative binomial) for nine models on doctor visits

	$\hat{\beta}_1$		$\hat{\beta}_2$	
Model	Netherlands	Spain	Netherlands	Spain
Self-assessed health				
Poor	2.700***	2.960***	1.673***	1.145***
Fair	1.700***	1.828***	1.015***	0.724***
Good	0.561***	1.266***	0.375***	0.315*
Very good	0.084	0.666**	0.171	-0.230
LR-test of joint significance (χ^2)	*122.621****	*55.796****	*262.602****	*179.176****
23 indicators and grip strength residual				
LR-test of joint significance (χ^2)	*107.946****	*40.750****	*158.491****	*178.003****
Mean of 23 indicators				
Mean of health indicators	7.119***	2.968***	2.827***	2.134***
Latent health method				
Health index	-2.178***	-1.284***	-1.216***	-0.958***
Latent health method: 5 quintiles				
0%-20%: most unhealthy	1.588***	1.410***	1.072***	1.175***
20%-40%	1.201***	1.181***	0.571***	0.888***
40%-60%	0.409**	1.034***	0.504***	0.503***
60%-80%	0.313**	0.148	0.141	0.342***
LR-test of joint significance (χ^2)	*85.115****	*30.478****	*134.723****	*141.622****
Self-rating in 6 domains (categorical)				
Pain	0.533**	0.596**	0.317***	0.333***
Sleeping problems	0.082	0.083	0.124	0.018
Mobility problems	0.519**	-0.456	0.143	-0.127
Cognition problems	0.034	0.087	-0.108	0.091
Breathing problems	-0.331	0.806**	0.101	-0.008
Emotional health problems	0.448*	-0.239	0.066	0.141**
LR-test of joint significance (χ^2)	*40.103****	*17.755****	*41.747****	*63.278****
Vignette method				
Pain	0.495***	0.465**	0.096	0.193***
Sleeping problems	-0.174	0.103	0.107	0.133***
Mobility problems	0.408*	-0.018	0.148*	0.113*
Cognition problems	0.502	0.203	0.154	0.302
Breathing problems	0.411	0.700**	0.133	0.176**
Emotional health problems	0.639**	0.208	0.187	0.113**
LR-test of joint significance (χ^2)	*42.058****	*17.668****	*28.719****	*92.069****
Uncorrected domain-indices				
Pain	0.689***	0.648*	0.155	0.260**
Sleeping problems	-0.154	0.173	0.094	0.244***
Mobility problems	0.512*	-0.079	0.211**	0.154*
Cognition problems	0.240	0.230	0.295	0.093
Breathing problems	0.606	0.775	0.085	0.261**
Emotional health problems	0.735***	0.268	0.221*	0.185**
LR-test of joint significance (χ^2)	*43.421****	*15.666****	*28.832****	*91.552****
Latent health method and vignette method				
Health index	-2.098*	-0.679	-0.431	-0.646**
Pain	0.474***	0.468**	0.074	0.209***
Sleeping problems	-0.175	0.118	0.122	0.146***
Mobility problems	-0.140	-0.225	0.063	-0.075
Cognition problems	0.558	0.256	0.158	0.315
Breathing problems	0.372	0.709**	0.117	0.161**
Emotional health problems	0.613**	0.196	0.171	0.101**
LR-test of joint significance (χ^2)	*45.618****	*18.239****	*29.436****	*98.196****
LR-test: latent health method (χ^2)	*3.560*	*0.571*	*0.717*	*6.127***
LR-test: vignette method (χ^2)	*20.038****	*14.220****	*11.887****	*62.111****

Note: * p<0.10; ** p<0.05; *** p<0.01.

The results for nine different sets of variables of the logit model and the truncated negative binomial model are presented in Table 6. Although in all of the nine sets of variables we include demographic and socioeconomic variables as controls, we only show the coefficients for the health variables here, because these have our primary interest. Next to this, we also provide likelihood-ratio tests for the joint significance of the health variables, in the case of more than one health variable. The models using variables from the vignettes (the last four models in the table) are estimated using the smaller vignette sample; the other models are estimated using the full available sample. However, in each model we use different explanatory health variables, and for each of these variables there is some missing information on a few individuals. Therefore, we use the sample that excludes all individuals with missing information on at least one of the health variables. This leaves us with sample sizes of 2390 for the Netherlands and 1757 for Spain for the full sample, and 471 for the Netherlands and 424 for Spain for the vignette sample. By using the smaller samples that have valid values for all variables that are used in all models, we obtain directly comparable results (as the estimations are done on exactly the same sample sizes).

To provide some general intuition on the control variables: the influences of the demographic and socioeconomic control variables generally do not differ, depending on which health variable is included in the model. In general, for the logit part of the model it holds that older individuals, females, highly educated and retired individuals have a larger probability of visiting a medical doctor. However, the age effect becomes insignificant when the obtained health measures from the latent health method and/or the vignette method are included. For the truncated negative binomial part of the model it holds that, given that an individual visits a medical doctor at least once, older individuals, employed individuals and highly educated individuals have generally less visits than younger, unemployed and low-educated individuals. These effects are similar for all different models, so the effects do not depend on which health measure is included in the model.

The first two columns of Table 6 show the results of the logit part of the model, to explain whether or not an individual visits a medical doctor. The first model includes (besides demographic and socioeconomic controls) self-assessed health and is used to get a general idea of how self-assessment of health affects the decision on whether or not to go to a medical doctor. This model does not involve variables obtained from either the latent health method or the vignette method discussed in this thesis. Individuals were asked to rate their own health on a scale ranging from *"1. Excellent"* to *"5. Poor"*. We include four of the five response options here as dummies, to make the results comparable to results in a later model that includes quintiles of the latent health method. With respect to individuals who answered *"Excellent"* to the self-assessed health question, individuals indicating poorer health have a larger probability to visit a medical doctor. The coefficients are very significant ($p<0.001$), except for the individuals who answered *"Very good"*, because this response option is very close to *"Excellent"*, the omitted category. The likelihood-ratio test for joint significance of the four dummies indicates that they are jointly significant (i.e. this model is preferred over the model with only demographic and socioeconomic controls, and no health variables). The results are similar for both countries.

The second model does not include the question on self-assessed health, but instead it includes the 23 health indicators on mobility and (I)ADLs that were used in the latent health method, plus the grip strength residual (so, all 24 dependent variables in the latent health method). We do not show the resulting coefficients for this method here in the interest of space. However, the likelihood-ratio tests for both countries show strong joint significance for the 24 variables in the logit part of the model.

The third model in the table includes the mean of the 23 binary health indicators. This model is estimated to be able to compare it to the health index that is obtained from the latent health method, which is also one continuous variable (instead of 23). In this way, we can compare the model with the health index from the latent health method to a model with the same number of variables. If computing a simple average over 23 indicators turns out to be a better predictor than the health index from the latent health method is, then the method by Meijer et al. [2011] is not useful for prediction purposes. We did not include the grip strength residual in the mean over the health indicators, because the grip strength residual is not binary. Including the mean over all 23 indicators is as expected significant, and the positive coefficients indicate that the more difficulties an individual has, the more likely s/he is to visit a medical doctor.

The coefficients of the latent health method for the logit part are also significant, but have the opposite sign, as expected. This is due to the fact that for the health index, a higher number indicates better health and thus a smaller probability of visiting a medical doctor. In the next model, the health index from the latent health method is divided into five quintiles ranging from poorest health to best health, and we include four of the five resulting dummies, to make results comparable to the first model, where self-assessed health was included. The coefficients and significances are similar to the ones in the model with self-assessed health, although the magnitudes of the coefficients in the model with the quintiles of the latent health index are smaller. According to the likelihood-ratio test, the coefficients of the quintiles are jointly also less significant than the coefficients of self-assessed health jointly are.

The next three models make use of the vignettes. As only a small part of the respondents filled out the vignettes, this model only uses the smaller vignette sample, whereas the above five models are estimated using the full available sample for both countries. First, we include the categorical responses on own-health problems for the six domains that are not corrected for reporting heterogeneity. Here, we could have included the responses in the six domains as dummies (like we did in the model with self-assessed health), but we want to compare it to a model where we include the health measures in the six domains that are corrected for reporting heterogeneity. In the way we did it here, both models have the same number of explanatory variables, which makes the results better comparable. In the second of these vignette models, called the “vignette method”, we included the measures on health problems in the six domains that are corrected for reporting heterogeneity using the vignette method, as obtained in Section 4.2. Overall, there seems to be a bit more significance for the Netherlands than for Spain. Out of the six domains, the domains *pain*, *mobility problems* (in the Netherlands), *breathing problems* (in Spain) and *emotional health problems* (in the Netherlands) seem to have explanatory power for the decision whether or not an individual goes to see a medical doctor.

The third model using vignettes, called “uncorrected domain-indices”, is a model that is used to see whether the correction for reporting heterogeneity is necessary and useful. What we did in this model, is basically applying the same method as the vignette method, but we did not allow for individual-specific thresholds in the model. So, instead of the HOPIT model that is used in the vignette method, we now used a simple ordered probit model, where we used the same explanatory variables as in the HOPIT model that we used for the vignette method. So, here we also obtain six measures of health problems, in the six health domains, but the difference is that these indices are not corrected for reporting heterogeneity. Overall, the same coefficients are significant as in the vignette method.

One of the interesting questions of this research is whether the two methods, the latent health method and the vignette method, contain basically the same information, so that the

Figure 1: Scatter for latent health index against measure *mobility problems* vignette method

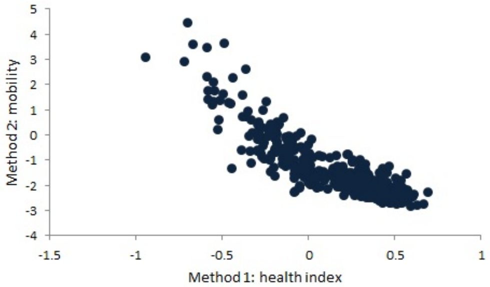
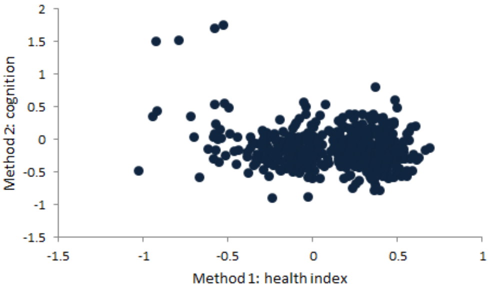


Figure 2: Scatter for latent health index against measure *cognition problems* vignette method



use of only one of the two methods in a model for health care use is sufficient. Or maybe the health measures from the two methods contain completely different information on individual health, so that they can be used both simultaneously. To research this, we first look at the pairwise correlations between the health index from the latent health method on one side, and the six health problems measures from the vignette method on the other side. These correlations are all significant, but differ in magnitude. The strongest correlation is visible between the latent health index and the index in the domain *mobility problems* from the vignette method: the correlation is equal to -0.866 ($p<0.001$). The smallest correlation is equal to -0.203 ($p<0.001$), which is the correlation between the latent health index and the index in the domain *cognition problems* from the vignette method. Scatterplots show the stronger correlation between the latent health index and the *mobility problems* measure, and the smaller correlation between the latent health index and the measure of *cognition problems*; see Figures 1 and 2) for scatterplots for the Netherlands.

The last model, at the bottom of Table 6, includes the health measures from both methods. Again, the smaller vignette sample is used for this analysis. To be more precise, X_i in equation (9) is now a vector consisting of the demographic and socioeconomic control variables, and next to these variables, the vector also contains the health index from the latent health method and the six reporting style-corrected ill-health measures from the vignette method. So, the measures from the two methods are incorporated in the model as additives. For the logit part, the health index from the latent health method by Meijer et al. [2011] is not very significant (and even insignificant for Spain), where it was significant if the six measures from the vignette method were not included. Generally the same vignette measures are significant as in the model without the latent health index, but concerning significances there is one difference: the coefficients for *mobility problems* are now insignificant for both countries (for the Netherlands, the p-value was equal to 0.061 in the model without the latent health index, which can be seen as somewhat significant, and it is equal to 0.685 in the model with the health measures from both methods, which is totally insignificant). This reflects the correlations that were discussed before; as the correlation between the latent health index and the measure of *mobility problems* is very high (-0.866), by including them both in a model, one of them becomes insignificant.

The likelihood-ratio test for joint significance of all seven (ill-)health measures shows that jointly they are significant, in both countries. However, if we perform a likelihood-ratio test for only adding the health index from the latent health method, when the six health measures from the vignette method are already included in the model, we see that the health index from the latent health method does not have significant additional explanatory power for the probability of visiting a medical doctor, as the likelihood-ratio tests for both countries are insignificant. Given that the health index from the latent health method is already incorporated in the model, adding the six measures from the vignette method still improves explanatory power. From this, we can conclude that for predicting whether an individual goes to see a medical doctor or not, including the six measures for health problems from the vignette method is sufficient, in that the inclusion of the health index from the latent health method does not improve prediction significantly.

The last two columns of Table 6 give the coefficients for the second step of the health care use model, the truncated negative binomial model. These coefficients try to explain how often an individual visits a medical doctor, given that s/he has at least one visit. Generally, the coefficients are significant and in the expected "direction", where a poorer health leads to a higher number of times an individual went to a medical doctor in the past twelve months. The only coefficients that are not all significant are, just as in the logit part of the health care use model, the coefficients of self-rated health problems in the six domains, in particular for the Netherlands. If both the health index from the latent health method and the six health measures from the vignette method are included in the model, at the bottom of Table 6, the same happens as in the logit part of the health care use model: the one domain-specific index that was significant before for the Netherlands, the *mobility problems* measure, has become insignificant. The likelihood-ratio tests indicate that for both countries, all seven health measures together have additional explanatory power, but the health index from the latent health method is not significant for the Netherlands once the six measures from the vignette method are already included in the model. For Spain, the likelihood-ratio test for additionally including the latent health index is significant at the five percent level. Similar to the results for the logit part of the health care use model, once the latent health index is already included in the model, including the six vignette measures does still improve explanatory power significantly for both countries. In general, it seems that for predicting the number of times an individual visits a medical doctor, the six measures of health problems that were obtained from the vignette method are less useful for the Netherlands, but they are useful for Spain. Apparently, how often one goes to the doctor is less related to how an individual feels (in the six domains) in the Netherlands, than it is in Spain.

How well do the health measures from the latent health method and the vignette method perform? Are they useful for prediction, or could we also just use the raw variables instead of applying the two relatively difficult methods as we did in this thesis? Table 7 gives an overview of all obtained log-likelihoods for the two steps of the models in total (by adding up the two log-likelihoods of the two parts, as explained in Section 2.3; see equation (12)), as well as two information criteria: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The log-likelihoods, as well as the AIC and BIC, are divided by the sample size of the estimation. This means that the numbers in the columns $\frac{\log(\ell)}{N}$ represent the average contributions to the log-likelihood per individual in the sample; for the information criteria we similarly show the average contribution per individual. As explained before, we use the largest possible sample size for which the information on all health variables is present, so that all models do have the same sample size, except for the models using the vignettes. To make the first five models comparable to the other four models (so that the latent health method (which uses the full sample) and the vignette method (which uses the vignette sample) can be compared), we also re-estimated the first five models with only the vignette subsample. These results are also presented in Table 7.

Table 7: Log-likelihood, AIC and BIC for all nine two-step models, full and vignette sample

	Netherlands					Spain				
Model	N	k	$\frac{\log(\ell)}{N}$	$\frac{AIC}{N}$	$\frac{BIC}{N}$	N	k	$\frac{\log(\ell)}{N}$	$\frac{AIC}{N}$	$\frac{BIC}{N}$
Full sample										
Self-assessed health	2390	27	-2.50	5.03	5.10	1757	27	-2.93	5.89	5.98
23 indicators & grip strength	2390	66	-2.53	5.11	5.27	1757	64	-2.93	5.94	6.14
Mean of 23 indicators	2390	21	-2.55	5.11	5.16	1757	21	-2.97	5.95	6.02
Latent health method	2390	21	-2.54	5.09	5.14	1757	21	-2.95	5.93	5.99
Latent health method: quint.	2390	27	-2.54	5.10	5.17	1757	27	-2.95	5.93	6.01
Vignette sample										
Self-assessed health	471	27	-2.38	4.88	5.12	424	27	-2.84	5.81	6.07
23 indicators & grip strength	471	56	-2.41	5.05	5.55	424	55	-2.89	6.03	6.56
Mean of 23 indicators	471	21	-2.44	4.97	5.16	424	21	-2.93	5.96	6.16
Latent health method	471	21	-2.43	4.94	5.12	424	21	-2.91	5.93	6.13
Latent health method: quint.	471	27	-2.43	4.97	5.21	424	27	-2.91	5.96	6.22
Self-rating in 6 domains (cat.)	471	31	-2.38	4.90	5.17	424	31	-2.87	5.88	6.17
Vignette method	471	31	-2.40	4.92	5.20	424	31	-2.83	5.81	6.11
Uncorrected domain-indices	471	31	-2.39	4.92	5.19	424	31	-2.83	5.82	6.11
Latent health & vign. method	471	33	-2.39	4.92	5.21	424	33	-2.82	5.80	6.12

Note: k = number of parameters.

Out of the five models in the upper part of the table, which shows the results for the five models that use the full sample, the model with self-assessed health as explanatory variable has the highest log-likelihood and the lowest information criteria. It is not very surprising that self-assessed health performs well, as it is widely known that an individual's self-assessment of health is a very useful variable to explain health related behavior.

The health index resulting from the latent health method has a higher log-likelihood and lower AIC and BIC than the average of the 23 health indicators. This is an argument for the use of the latent health method by Meijer et al. [2011], instead of just taking the average over the 23 indicators. The latent health method also performs better than the 23 raw health indicators and grip strength together; although the average log-likelihood contribution per individual is higher for the latter model, the AIC and BIC indicate lower values for the latent health method. This is due to the fact that in the latent health method, we only use one index for health (that appears twice in the model), whereas the other model includes 24 variables on health (that also appear two times in the model). Both information criteria penalize for this higher number of parameters. Including the quintiles obtained from the latent health index leads to approximately the same log-likelihood contribution per individual as just including the latent health index itself, but because of the higher number of parameters when including the quintiles, the AIC and BIC assess the model with just the latent health index to be better.

The lower part of Table 7 shows the log-likelihoods and AIC/BIC for all nine models using the smaller vignette sample. First of all, for the five models that were also estimated for the full sample, the coefficients and their significances are similar for both samples. Therefore, it is permissible to use the vignette sample to compare all nine models. The log-likelihoods are substantially less negative for the models using the vignette sample as compared to the models using the full sample, which is due to the smaller number of observations. The log-likelihoods per individual are also a bit less negative, reflecting that optimization is easier when including fewer individuals. Again, out of the first five models, apart from the model with self-assessed health, the information criteria indicate that the model with the latent health index by Meijer et al. [2011] has the best goodness of fit.

The model with the categorical self-assessed health problems in six domains performs almost as good as the model with self-assessed health in it; AIC and BIC are however a bit higher, due to the larger number of parameters. So, even without applying the vignette method that corrects for health indicators and reporting styles, these six categorical variables seem to have a lot of information on health behavior in them. This makes sense, as these indicators in the six health domains are also self-assessment questions on an individual's health and we know that self-assessment of health is valuable information in explaining health behavior.

The log-likelihoods for the models with these six categorical self-assessment variables included are higher than the log-likelihoods of the models where we included the 23 health variables plus grip strength. This means that the ill-health variables that are used in the vignette method by Bago d'Uva et al. [2011a] are more useful for predicting doctor visits, than the health indicators that are used in the latent health method by Meijer et al. [2011] are. So, without applying these methods, we can already see that the variables that are used for both methods are not equally useful for explaining health behavior.

Compared to the model with the six uncorrected health measures, the model that includes the ill-health measures from the vignette method, performs better in Spain, but worse in the Netherlands. Using the vignette method, a correction for reporting styles is applied, as well as a correction for demographic, socioeconomic and health variables. In Spain the log-likelihood is higher, and the AIC and the BIC are 0.07 and 0.06 points lower per individual, respectively. In the Netherlands, the log-likelihood of the uncorrected model was better, however. So, applying the correction for response style and health indicators, following Bago d'Uva et al. [2011a], seems to improve explanatory power a lot, but only in Spain.

The model with the uncorrected domain-indices (obtained from the ordered probit model, where we did not correct for reporting style) is also helpful to find this difference between the Netherlands and Spain. It turns out that, in terms of information criteria, the uncorrected domain-specific ill-health indices are better to use in the case of the Netherlands, compared to the corrected measures from the vignette method. However, for Spain, the correction for reporting heterogeneity leads to more useful health measures, as the AIC indicates that the uncorrected health indices perform worse.

In the last model, where we include the health index from the latent health method, as well as the six measures of health problems from the vignette method, the log-likelihood for the Netherlands becomes approximately equal to the log-likelihoods of the models where either self-assessed health or the self-rating in six domains is included. For Spain, the log-likelihood for this model including the health measures from both methods is even better than all other used models. The AIC and BIC are very close to the ones for the model that only includes the health measures from the vignette method. This means that, once we have controlled for the six corrected domain-specific variables, the inclusion of the latent health index does not lead to an improvement. This confirms what the likelihood-ratio tests in Table 6 already found: when the six ill-health measures from the vignette method are already included in the model, the inclusion of the latent health index does not lead to a significant improvement of the model. However, if only the health index from the latent health method is included in the model for health care use, inclusion of the six measures from the vignette method does lead to an improvement of the fit of the model, according to the higher log-likelihoods and lower AIC and BIC for the model with both methods included, compared to the model with only the latent health method included. This is also in accordance with the likelihood-ratio tests in Table 6. In general, the six corrected variables from the vignette method contain a lot of information and they are useful in predictions of health behavior, in this case visiting a doctor. According to the obtained log-likelihoods and information criteria, for prediction purposes, the latent health method is less useful than the vignette method. However, instead of including the 23 health indicators and grip strength that are used to obtain the latent health index, using the latent health index improves the fit of the model, according to the information criteria.

5 Conclusions

The main goals of this thesis are to construct useful health measures that are corrected for differences in reporting styles across individuals and across countries in surveys. Further, if we can construct such indices, we want to know whether they explain health-related behavior well. Otherwise, we might as well just use raw information from the survey as explanatory variables, instead of applying difficult methods to obtain health measures. With all of the results in this thesis, we have to keep in mind that the results may depend on the health indicators that are used to construct the (ill-)health measures in the two methods.

We implemented the latent health method by Meijer et al. [2011] to construct one index for general health. We found that the results in this thesis, where we used the second wave of the SHARE data, are similar to the results they found when using wave one. This method is useful to implement, as it gives us a good insight in the structure of individual health; the relations between 23 health indicators and latent health became visible through the use of this method. We found significant negative relations between all health indicators and latent health, meaning that reporting difficulties with each one of the health indicators leads to a significant decrease in general health. Especially in Spain, (I)ADLs have a stronger influence on health, as indicating difficulties with one of them leads to a relatively large decrease of health. Furthermore, older individuals, females, less-wealthy individuals and individuals with underweight or obesity are likely to be in worse health.

For the vignette method, we followed Bago d'Uva et al. [2011a], where we made use of six different domains of health problems. The results of the HOPIT model per health domain showed us that individuals' response styles are generally influenced by symptoms and diagnoses in the corresponding domain that individuals themselves encounter, when they rate vignettes. This means that, once an individual suffers from health domain-related symptoms him/herself, s/he rates the persons in the vignette stories differently; response style does not seem to depend much on demographic and socioeconomic characteristics. After correcting for reporting heterogeneity, the indicators that significantly influence own-health problems in this method were found to be age, together with health-domain related diagnoses that an individual has.

Using a model for health care use, we have several findings. First of all, when it comes to explaining health-related behavior (visiting a medical doctor in this case), the latent health index can better be used than the 23 health indicators and grip strength, which are used to obtain the latent health index. That is, the model fit is better when the one latent health index is used as explanatory variable, compared to a model that includes 23 health indicators and grip strength. The latent health index also outperforms the simple average of the 23 health indicators. So, for explaining health behavior, the latent health method constructs a health index that explains health behavior better than the raw data does.

Secondly, we compare the categorical responses to self-assessed ill-health questions in the six health domains to health measures from the vignette method, which are corrected for demographic, socioeconomic and health variables, as well as for reporting heterogeneity. We find that for the Netherlands, it is better to use the categorical (uncorrected) responses; including the six measures of health problems from the vignette method makes the model worse. Even if we do not correct for reporting style differences, but we only control for demographic, socioeconomic and health variables, the model fit is worse. In Spain, on the other hand, correcting the six ill-health measures for reporting heterogeneity makes the variables more useful in explaining doctor visits. So, we find that whether or not we should correct for reporting style differences, depends on the country that we are analyzing. We have to keep in mind, however, that the used vignette sample sizes here are relatively small, and therefore might not be representative.

If we compare the health indicators that are used for the latent health method on one side, and the health indicators that are used for the vignette method on the other side, we find that the six (uncorrected) ill-health measures from the vignette method are more useful in the model for health care use, than the 23 health indicators in the latent health method are. This implies that, without applying the two methods, the variables that are used for the vignette method in themselves contain more or better information on health behavior. Including the six corrected measures from the vignette method and the one health index from the latent health method all at once in the model for health care use, tests show that the seven variables are jointly significant. However, once the six variables from the vignette method are included in the health care use model, adding the latent health index does not significantly improve the model. The other way around, when the latent health index is already included as explanatory variable, we find that including the six measures from the vignette method still improves explanatory power. We can say that for prediction purposes, the ill-health measures from the vignette method are more useful than the latent health index is.

In the end, we cannot say that one of the two methods is systematically better than the other one, as we noted before that both methods use different information on health. Therefore, results may depend on the health indicators that were used. Both methods seem to perform better in explaining doctor visits than just using the raw information, although the vignette method can better be used for this purpose.

As said before, further research is needed to detect possible differences in response patterns between males and females. For now, for the latent health method we assumed that there are no such significant differences and this allowed us to do the analyses for males and females jointly, where Meijer et al. [2011] mentioned that, without further research, we may not make this assumption of equal response patterns. In the vignette method, however, we found that when it comes to vignettes, females respond slightly different than males (for example in the domain of mobility problems.

Meijer et al. [2011] also mentioned that there may be a need for more than one latent health variable; in the current analysis, we only included one latent variable in the latent health method to obtain one general health index. However, as they suggested, health is a multidimensional concept and therefore we may need to include more than one health index. There may be a structure of different health domains; we assumed this in the vignette method by Bago d'Uva et al. [2011a], where six different health domains are used. So, in future research, we suggest to consider the possibility of including more latent variables in the health measurement model and find out whether this is an improvement.

One more thing that we have to keep in mind regarding the results in this paper, is that the two assumptions of the HOPIT model may or may not hold: the assumptions of vignette equivalence and response consistency. Previous research by Van Soest et al. [2007] found that anchoring vignettes do a very good job in correcting for differences in reporting styles and in their research the assumption of response consistency is found to hold. However, as already noted, Bago d'Uva et al. [2011b] found that the two assumptions of the HOPIT model may not always hold. Further research is desirable on this topic by further exploring the vignettes.

6 References

Bago d'Uva, T., Lindeboom, M., O'Donnell, O. and Van Doorslaer, E. [2011a], "Education-Related Inequity in Health Care with Heterogeneous Reporting of Health," *Journal of the Royal Statistical Society Series A* 174(3), 639-664.

Bago d'Uva, T., Lindeboom, M., O'Donnell, O. and Van Doorslaer, E. [2011b], "Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity," *Journal of Human Resources* 46(4), 872-903.

Bollen, K.A. [1989], *Structural Equations with Latent Variables*, John Wiley & Sons, Inc., New York.

Cameron, A.C. and Trivedi, P.K. [2005], *Microeconometrics: Methods and Applications*, Cambridge University Press, New York.

Jürges, H. [2006], "True Health vs. Response Styles: Exploring Cross-country Differences in Self-Assessed Health," German Institute for Economic Research Discussion Paper 588.

King, G., Murray, C.J.L, Salomon, J. and Tandon, A. [2004], "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research," *American Political Science Review* 98(1), 184-191.

Majo, M.C. [2010], "A Microeconometric Analysis of Health Care Utilization in Europe," *Dissertation, Tilburg University.*

Meer, J., Miller, D.L. and Rosen, H.S. [2003], "Exploring the health-wealth nexus," *Journal of Health Economics* 22(5), 713-730.

Meijer, E., Kapteyn, A. and Andreyeva, T. [2011], "Internationally Comparable Health Indices," *Health Economics* 20(5), 600-619.

United Nations Educational, Scientific and Cultural Organization [2006], ISCED 1997. Downloaded from http://www.uis.unesco.org/Library/Documents/isced97-en.pdf, accessed July 29, 2011.

Van Soest, A., Delaney, L., Harmon, C.P., Kapteyn, A. and Smith, J. [2007], "Validating the Use of Vignettes for Subjective Threshold Scales," IZA Discussion Paper 2860.

Zimmer, Z., Natividad, J., Lin, H. and Chayovan, N. [2000], "A Cross-National Examination of the Determinants of Self-Assessed Health," *Journal of Health and Social Behavior* 41(4), 465-481.

A Appendix

Table 8: Used variables in the latent health method

Variable	Description
Health indicators (y_{cin})	
Mobility:	
1) Walking 100 meters	Dummy: 0 = no difficulty, 1 = any difficulty
2) Sitting two hours	
3) Getting up from chair	
4) Climbing several flights of stairs	
5) Climbing one flight of stairs	
6) Stooping, kneeling, crouching	
7) Reaching arms above shoulder	
8) Pulling/pushing large objects	
9) Lifting/carrying weights > 5 kg	
10) Picking up small coin from table	
Activities of daily living:	
11) Dressing	Dummy: 0 = no difficulty, 1 = any difficulty
12) Walking across a room	
13) Bathing/showering	
14) Eating, cutting up food	
15) Getting in or out of bed	
16) Using the toilet	
Instrumental activities of daily living:	
17) Using map in strange place	Dummy: 0 = no difficulty, 1 = any difficulty
18) Preparing hot meal	
19) Shopping for groceries	
20) Telephone calls	
21) Taking medications	
22) Doing work around house/garden	
23) Managing money	
Reference variable:	
24) Residual grip strength	Residual from grip strength equation
Demographic and socioeconomic variables (x_{ci})	
Age	(Age−65)/10 $[(\text{Age}-65)/10]^2$ $[(\text{Age}-65)/10]^3$
Gender	Dummy: 0 = male, 1 = female
Education	Dummies for secondary education, tertiary education and missing education; reference category is no/primary education
Household size	Number of individuals in household
Living with spouse or partner	Dummy: 1 = living with spouse or partner
Household net worth	Inverse hyperbolic sine function of net worth
BMI	Dummies for underweight (BMI<18.5), overweight (25≤BMI<30), obesity class I (30≤BMI<35), obesity class II/III (BMI≥35) and missing BMI; reference category is normal weight (18.5≤ BMI < 25)

Table 9: Used variables in the vignette method

Variable	Description
Self-rated own health & Vignette ratings (6 domains)	
Mobility problems	Difficulties: 1 = none, 2 = mild, 3 = moderate, 4 = severe, 5 = extreme
Cognition problems	
Pain	
Sleeping problems	
Breathing problems	
Emotional health problems	
Demographic and socioeconomic variables (W_i)	
Age	Age
Gender	Dummy: 0 = male, 1 = female
Education	Dummies for lower/upper secondary and tertiary education; reference category is no/primary education
Health variables, domain specific (Z_{di})	
Mobility problems	
Obese	Dummy: 1 = obese (BMI$\geq$30)
Grip strength	Maximum of grip strength measures
Limitations	Number of limitations with mobility
Diagnoses	Dummy: 1 = any of the following diagnoses: stroke, arthritis/rheumatism, hip/femoral fracture, Parkinson's disease
Symptoms	Dummy: 1 = any of the following symptoms: swollen legs, falling down, fear of falling down, dizziness, faints, blackouts
Cognition problems	
Date recall	Score on recalling day of week, day, month, year (0 = bad, 4 = good)
Word recall	How many words recalled out of a list of 10 words
Word recall, delayed	How many words recalled out of a list of 10 words, delayed
Pain	
Symptoms	Dummy: 1 = bothered by pain in back, knees, hips or other joints
Sleeping problems	
Euro-D score	Dummy: 1 = Euro-D score indicates trouble with sleeping
Symptoms	Dummy: 1 = bothered by sleeping problems
Medication	Dummy: 1 = takes medication for sleeping problems
Obese	Dummy: 1 = obese (BMI$\geq$30)
Diagnoses	Dummy: 1 = any of the following diagnoses: asthma, bronchitis
Breathing problems	
Symptoms	Dummy: 1 = bothered by breathlessness
	Dummy: 1 = bothered by persistent cough
Diagnoses	Dummy: 1 = diagnosed with chronic lung disease or asthma
Emotional health problems	
Euro-D score	Score on Euro-D depression scale
Medication	Dummy: 1 = takes medication for depression

Table 10: Used variables in the model for health care use

Variable	Description
Dependent variable (y_i)	
Doctor visits	Number of times seen or talked to a medical doctor in the last twelve months
Explanatory variables (X_i)	
Age	Age
Gender	Dummy: 0 = male, 1 = female
Living with spouse or partner	Dummy: 0 = living single, 1 = with spouse/partner
Household net worth	Inverse hyperbolic sine function of net worth
Employment	Dummy: 0 = unemployed, 1 = employed
Retirement	Dummy: 0 = not retired, 1 = retired
Education	Dummies for secondary and tertiary education; reference category is no/primary education