WORKING
PAPER

Health Indexes and Retirement Modeling in International Comparisons

ERIK MEIJER
ARIE KAPTEYN
TATIANA ANDREYEVA

WR-614

August 2008

This paper series made possible by the NIA funded RAND Center for the Study of Aging (P30AG012815) and the NICHD funded RAND Population Research Center (R24HD050906).

This product is part of the RAND Labor and Population working paper series. RAND working papers are intended to share researchers' latest findings and to solicit informal peer review. They have been approved for circulation by RAND Labor and Population but have not been formally edited or peer reviewed. Unless otherwise indicated, working papers can be quoted and cited without permission of the author, provided the source is clearly referred to as a working paper. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.
RAND® is a registered trademark.

LABOR AND POPULATION

Health Indexes and Retirement Modeling in International Comparisons

Erik Meijer* and Arie Kapteyn
RAND Corporation

Tatiana Andreyeva
Yale University

July 28, 2008

Abstract

It is widely believed that health plays a major role in retirement decisions. The most important problem in including health in retirement models is the lack of availability of a good measure of health at the individual level in existing data sets. This problem is exacerbated when a model spanning multiple countries is desired, because self-reports on health may not be comparable across countries. Arguably, physical measures are less influenced by cultural and linguistic differences than self-reports on general health or even on health conditions. We develop a cross-country measurement model for health in which the relations between functional limitations, self-reports, and physical measures like grip strength are used to construct health indexes. Comparability across countries is achieved by using the physical measurements to define the measurement scales, and allowing other parameters to vary across countries to account for cultural and linguistic differences in reponse patterns. The usefulness of the health indexes is then investigated by including it in some simple retirement models.

1 Introduction

Many countries around the world face an aging population, with at the same time decreasing average retirement ages (see, e.g., Gruber & Wise, 1999, 2004, 2005, 2007). This pattern has substantial economic effects, the most obvious of which is that more pensions have to be paid by fewer workers, at least for pay-as-you-go pension systems, and the question arises whether such systems can be afforded in the future. A solution that is often mentioned in the public debate is to *increase* rather than decrease retirement ages. As a result, the study of the determinants of (early) retirement has become a key focus in economic research. A problem with studying retirement

*RAND Corporation, 1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138, meijer@rand.org

is that many determinants that are expected to play a major role in this decision vary little or sometimes not at all within a country. An obvious example is the age at which one becomes eligible for public pensions and other aspects of social security systems, but the same holds for tax laws and other economic institutions. Cross-country variation in institutions and retirement ages then forms a source of information about the role of these institutions. These mechanisms are under government control and are thus important public policy tools to influence retirement decisions. Hence, cross-country study of retirement has high public policy relevance.

In order to study the role of institutions in retirement decisions across countries, other determinants of retirement must be accounted for as well. If these determinants also vary across countries, excluding them from a model leads to incorrect attribution of cross-country differences in retirement to institutions. Of course, these other determinants may be of considerable interest in their own right as well.

One of the determinants of retirement most often mentioned is worsening of health with increasing age. Gordon and Blinder (1980) mention two main ways in which health may affect retirement decisions. The first is that deteriorating health leads to decreasing productivity and hence possibly to decreasing (real) wages and more generally less attractive employment opportunities. The second hypothesis is that decreasing health may shift the relative preferences for leisure versus work, e.g., because work becomes more burdensome. The latter is a fairly direct effect of health on labor supply, whereas the former leads to intricate market interactions between labor supply and labor demand. On the other hand, if affordable health insurance is linked to employment, the direction of the preference shift becomes ambiguous a priori. An incomplete list of papers discussing the relationships between health and retirement (usually in a broader context) is Anderson and Burkhauser (1985), Sammartino (1987), Bound, Schoenbaum, Stinebrickner, and Waidmann (1999), Coile (2004), French (2005), Bound, Stinebrickner, and Waidmann (2008), as well as the overview by Currie and Madrian (1999) and the more general overviews about retirement by Hurd (1990) and Lumsdaine and Mitchell (1999).

One of the most intractable problems in international research on health is the comparability (or incomparability) of health measures across countries or cultures. The conventional approach to evaluating health within and across nations relies heavily on using measures of subjective health assessment such as self-reports of health and health conditions. Arguably, these measures are conditioned by cultural or social norms, differences in thresholds for medical diagnosis and access to health care resources, so that comparisons of health across different populations may be difficult or impossible with such gauges. In response to this issue, research on modeling comparable health measures has focused on finding objective measurement tools that would provide consistent evaluations of health across and within nations.

The ability to compare health across countries is a prerequisite for understanding the role of national policies and institutions in influencing behavior. Health plays a substantial role in many economic models, including models of labor force participation, retirement, or savings decisions. Omitting health for a lack of comparable health measures may produce biased estimates of model parameters if health is correlated with the variables of interest. Although economic models differ

greatly in what categorization and specific pecuniary factors they use, the reality is that economic incentives (e.g., disability benefits) are often conditioned on health. In a cross-national study of economic behavior, the use of comparable health measures helps not only to provide unbiased assessment of behavior but also to predict the effects of changes in policies. Based on comparable measures of health, we can evaluate effectiveness of different policy initiatives, assess health interventions across countries, and set priorities for intervention.

The analysis of inequalities in health within and across nations points at another dimension of research that needs comparable health measures. Health inequalities that are generally traced to inequalities in income, education and other socioeconomic categories persist in all countries but there are cross-national differences in their level, rate of change and strength of association (Carlson, 1998; Kopp, Skrabski, & Szedmak, 2000; Kunst et al., 2005; Macinko, Shi, Starfield, & Wulu, 2003; van Doorslaer et al., 1997).

Efforts to develop comparable, composite measures of population health have a long history. Yet, despite many efforts to develop a consistent instrument to measure health, there seems to be no standard that is universally accepted (Murray, Salomon, Mathers, & Lopez, 2002). The measures developed to date differ methodologically (on, for example, the use of weights for health problems or coverage of health domains) and conceptually (composite measures of individual health vis-à-vis population health). From a conceptual point of view, indices designed to capture the detailed components of individual health require a different set of considerations than more general, population-oriented health status measures. The latter include generalizable data on mortality, the prevalence, incidence and natural history of non-fatal conditions, prevalence-based valuations for the disability weights associated with these conditions (Murray et al., 2002).

One of the best known summary measures of population health is the disability-adjusted life year (DALY) that made its debut in the World Development Report 1993 (World Bank, 1993). The DALY measures the gap between a population's actual health and some explicit goal, and is calculated as the present value of the future years of disability-free life that is lost, to all causes, whether from premature mortality or from some degree of disability during a period of time (Erickson, Wilson, & Shannon, 1995; Murray et al., 2002). Another common summary measure of population health used by the World Health Organization (WHO) is the Disability-Adjusted Life Expectancy (DALE) that measures the expected number of years of life in full health, or healthy life expectancy. DALE estimates are based on the estimates of severity-weighted disability prevalence developed for the non-fatal component of disease and injury burden (Murray et al., 2002).

There is no single instrument to monitor population health in the US. Measures used by the US government agencies include the Centers for Disease Control and Prevention Health-Related Quality-of-Life 14-Item Measure (CDC HRQOL-14) "Healthy Days Measures"[1] and the Health

[1]The CDC HRQOL-14 combines three modules: "Healthy Days core module" that evaluates self-reported general health, number of days over the last 30 days in ill physical health, mental health or health-related limited functional ability; "Activity Limitations Module" with 5 questions about physical, mental, or emotional problems or limitations in daily life, and "Healthy Days Symptoms Module" with 5 questions about the number of days in the past 30 days

and Activity Limitation Index (HALex), also known as the Years of Healthy Life (YHL) that is based on age-specific mortality rates, activity limitation and self-rated overall health (Erickson, 1998; Sondik, 2002; Stewart, Woodward, & Cutler, 2005). YHL was used as a summary health measure in Healthy People 2000, the primary prevention initiative in the US. For the Healthy People 2010 program no single summary measure has been identified. About 20 leading health indicators serve as a summary set of nation's health measures (Sondik, 2002).

While summary measures like DALE, YHL, or HRQOL are useful for comparisons of overall health across countries and for the measurement of progress of one nation's health, they offer limited power in measuring the current health status of an individual that is essential in economic models. Self-rated health and reports of doctor-diagnosed chronic conditions have been common measures of the current individual health status in most types of modeling. On the one hand, health self-assessment is a good measure of health because the question has high response rates and predictive power for other health measures and mortality. On the other hand, self-reported health evaluations are subject to many biases related to differences in culture, language and institutional environment. In international comparisons of health, it is impossible to separate the observed variation in the subjective health responses into the variation in genuine health and the differences related to cultural or social norms. Furthermore, self-reported measures may be unstable over short periods of time.

Recent innovations in the design and data collection of some household surveys make it possible to construct internationally comparable health measures using a more objective and accurate evaluation of health than self-perceived health. These advances include collecting information on physical measures like grip strength and walking speed in cross-national multidisciplinary studies such as the U.S. Health and Retirement Study (HRS), the English Longitudinal Study of Ageing (ELSA), and the Survey of Health, Ageing and Retirement in Europe (SHARE). Interviewers take physical measures of health such as grip strength using the same protocol across all countries. Such assessments are therefore less likely to be subject to the biases affecting self-reports of health, and may overcome the measurement issues of cultural differences in how people evaluate their health. The importance of using objective measures of health was stressed by Bound (1991) and Burkhauser and Cawley (2006).

The primary objective of this paper is to construct internationally comparable measures of health. To address the scaling issues inherent in cross-national comparability of subjective health questions, we develop a model that relies on objective health indicators like grip strength to arrive at comparable scales. We next incorporate the health measures in a simple model of retirement. It is often stated that an objective measure of health is essential because of *justification bias*, which asserts that people who have retired early rate their health lower to justify their exit from the labor force. However, empirical evidence for the existence of justification bias seems very limited (Dwyer & Mitchell, 1999). Therefore, we include self-rated health measures in addition to an objective measure in our models. But, as shown by Lindeboom and van Doorslaer (2004),

when experiencing specific symptoms. More detail about the CDC HRQOL-14 is available at the CDC website `http://www.cdc.gov/hrqol/hrqol14_measure.htm`

self-reports may differ between populations because of differences in reporting patterns that are unrelated to health. They regress self-reports on another health variable that is considered more objective. Jürges (2007) takes the same approach, but includes a large number of health variables as explanatory variables. Between-population differences in the resulting coefficients are then interpreted as reflections of differing reporting patterns. Our approach is based on similar ideas as those of Lindeboom and van Doorslaer, but different in the operationalization: we treat our objective measures not as infallible measures of health, but as imperfect indicators that are still subject to measurement error, but not to differences in reporting patterns. We also use more indicators of health than Lindeboom and van Doorslaer do, and different ones than Jürges does.

Section 2 describes the health data we use, while section 3 presents our model for health. Once the model has been estimated, one can use this to construct health measures. This is discussed in section 4. Section 5 then presents the empirical results, in the form of the measurement model estimates, the distribution of the health index, and its relation to age and socio-economic status. After a prelude presenting data on retirement patterns across countries in section 6, section 7 presents a simple retirement model that includes our health measures next to demographics and economic incentive measures. At this point the retirement model is primarily illustrative. Section 8 concludes.

2 Data on Health

We use information collected in the first wave of the Survey of Health, Ageing and Retirement in Europe (SHARE), which is a multidisciplinary cross-national longitudinal survey of continental Europeans over the age of 50 and their spouses. The baseline SHARE study includes data on 12 countries providing a balanced representation of the different European regions from Scandinavia (Denmark and Sweden) through Central Europe (Austria, France, Germany, Switzerland, Belgium, The Netherlands) to the Mediterranean (Spain, Italy, Greece, and Israel). We use data from SHARE Wave 1, Release 2.0.1 (July 5, 2007).

Designed after the role models of HRS and ELSA, SHARE combines information on health (e.g., self-reported health, physical and cognitive functioning, health behaviors, health care utilization and expenditure), psychological conditions (e.g., mental health, well-being, life satisfaction), socio-economic status (e.g., work activity, job characteristics, income, wealth and consumption, housing, education), and social support (e.g., social networks, volunteer activities).

The SHARE Wave 1, Release 2.0.1 sample includes 31,115 respondents. The survey has been administered by means of computer assisted personal interviewing (CAPI), mostly in 2004, with additional data collected in 2005 and 2006, to probability samples of individuals of 50 and over in all participating countries. The sampling plan follows a complex probabilistic multistage design to produce estimates representative of the non-institutionalized population aged 50 and above in each country. The study also interviews spouses younger than 50. The response rate varies by country but on average is 62% for households and 85% for individuals within participating

Table 1: Sample composition by country and gender.

Country	Male	Female	Total
Austria	777	1,072	1,849
Belgium	1,696	1,921	3,617
Denmark	757	857	1,614
France	1,365	1,671	3,036
Germany	1,364	1,566	2,930
Greece	1,235	1,424	2,659
Israel	1,135	1,349	2,484
Italy	1,125	1,382	2,507
The Netherlands	1,344	1,515	2,859
Spain	984	1,357	2,341
Sweden	1,406	1,588	2,994
Switzerland	448	497	945
Total	13,636	16,199	29,835

Note. Unweighted number of respondents.

households. A detailed description of the SHARE data and methodology has been published elsewhere (Börsch-Supan et al., 2005; Börsch-Supan & Jürges, 2005). The data are available to registered users from the SHARE website (`http://www.share-project.org`).

Sample selection

Because we estimate the models using sampling weights, we removed all cases for which the individual sampling weight was missing. These are mostly persons younger than 50 years of age, and a few persons whose age and/or gender was missing. We decided not to remove cases with missing data on any of the other variables. How we deal with the missings is described below. The resulting analysis data set contains 29,835 cases. Table 1 gives a breakdown by country and gender.

Missing data

There are 129 cases with all dependent variables for the health model missing. These do not contribute anything to the loglikelihood and thus have no influence on the estimation of the parameters of the health model. Nevertheless, it may be useful to include them in the retirement model, because we will be able to compute health indexes for them, although these will evidently be less precise than for observations with fewer missing values.

There are 686 cases with missing height and/or weight, including a few cases with unlikely values (height < 110 cm or weight < 10 kg).

SHARE uses multiple imputations for various missing variables; most important for our purposes are household income, assets, and education. Imputations for Israel were not yet

available in our data set. We use the means of the five imputations as our income and asset variables. In the modeling, we have treated zero mean incomes as missing (negative incomes do not occur). For education we use the first imputation, but there are only 83 observations for which education had to be imputed. We have classified "other" education as missing.

Health

SHARE contains extensive modules on physical health, combining information on subjective health assessment (based on the US categorization on the five-point scale from "poor" to "excellent" and the European categorization on the five-point scale from "very bad" to "very good"), indicators of doctor-diagnosed chronic conditions (heart disease, high blood cholesterol, hypertension, stroke, diabetes, lung disease, asthma, arthritis/rheumatism, osteoporosis, cancer, ulcer, Parkinson's disease, cataracts, hip fracture), a battery of functional limitations from more severe limitations with activities of daily living (ADL) to less disabling problems with instrumental activities of daily living (IADL) and mobility limitations. In addition, SHARE contains a limited number of physical measures, including self-reported body weight and height, interviewer-measured walking speed (for respondents aged 76 and older and those who had indicated having difficulty with walking 100 m) and grip strength (for all respondents). The appendix gives a detailed list of the variables we use.

Grip strength is a core physical measure of health that potentially enables cross-national comparability of health estimates and avoids some of the endogeneity problems inherent in more subjective health measures like self-rated health. It also helps to overcome the measurement issues related to biases that arise from subjectivity of self-reported health and health conditions due to cultural differences across and within countries, differential physician contacts or cross-national differences in the criteria for thresholds of medical diagnosis. At the same time, predictive validity of grip strength for assessing health was established in studies that found grip strength to be a better predictor of future medical problems than self-reported health (Christensen et al., 2000; Rantanen et al., 1999, 2000; Al-Snih, Markides, Ray, Ostir, & Goodwin, 2002).

SHARE asked respondents to report whether they had any difficulties doing various activities because of a health or physical problem in the last month before the interview (difficulties expected to last less than three months were excluded). We have selected 25 indicators to measure health and functional ability in SHARE, including reports of limitations with 10 activities of mobility, arm function and fine motor function, 6 ADLs, 7 IADLs, self-reported health, and grip strength. Table 2 summarizes the distribution of our analysis sample across countries and gender for selected health indicators. We report four measures of subjective health assessment in SHARE: any limitation with ADL, IADL, mobility limitations, and self-reports of fair or poor general health. The distribution of the data on self-reported health is particularly illustrative of the large cross-country differences embedded in self-reports. For example, the percentage of men who rate their health as poor or fair is more than three times as large in Germany as in Sweden, whereas approximately the same proportion of men in both countries report having some chronic health

Table 2: Percentage reporting health-related limitations or fair/poor self-reported health.

Country	At least one limitation			Fair/poor self-reported health
	Mobility	ADL	IADL	
Male				
Austria	44.1	7.8	11.8	28.2
Belgium	39.6	9.4	13.4	25.0
Denmark	34.2	9.9	11.7	24.6
France	38.3	12.8	13.0	32.7
Germany	47.5	8.5	11.1	37.1
Greece	44.4	6.5	11.3	25.2
Israel	40.9	11.8	18.5	38.0
Italy	43.8	10.1	9.6	34.6
The Netherlands	31.7	6.4	10.5	25.6
Spain	43.1	10.2	16.9	34.9
Sweden	35.5	8.0	11.5	11.1
Switzerland	28.4	4.5	4.7	13.8
Total	42.2	9.6	12.0	32.4
Female				
Austria	58.3	10.6	22.0	31.5
Belgium	58.1	16.2	24.2	29.5
Denmark	50.3	11.0	21.9	26.3
France	59.0	12.5	21.5	35.8
Germany	61.7	12.0	18.7	42.6
Greece	64.6	11.4	25.9	37.3
Israel	56.7	13.0	29.7	38.7
Italy	60.2	13.9	20.6	47.8
The Netherlands	51.5	10.7	21.6	29.7
Spain	64.9	15.1	30.1	49.4
Sweden	55.5	12.7	22.5	15.6
Switzerland	46.4	8.7	11.8	18.5
Total	60.0	12.8	21.9	40.2

Note. Weighted results.

condition (about 70%, not reported in the table). Another example is the male population of Denmark, whose life expectancy is on average one year less than of French men, but who are 20% less likely than the French to rate their health as poor/fair.

Table 3 presents the mean value and standard deviation of grip strength measurements by country and gender. The cross-national variation in grip strength is much smaller than the observed differences in self-reports of health. The difference between the highest and the lowest average national measurements is about 25% for both men and women. In all countries, the average grip strength of women is around two-thirds of the average level for men.

Table 3: Mean and standard deviation of maximum grip strength (kg).

	Male		Female	
Country	Mean	s.d.	Mean	s.d.
Austria	46.1	9.8	28.9	7.8
Belgium	44.0	10.2	26.2	7.1
Denmark	46.7	10.5	26.9	7.3
France	42.4	10.7	25.5	7.0
Germany	46.0	10.9	28.3	7.8
Greece	41.2	11.1	24.9	6.9
Israel	39.4	11.7	23.4	7.5
Italy	39.7	11.1	23.3	7.2
The Netherlands	45.5	10.4	27.7	7.6
Spain	37.4	10.5	22.3	7.6
Sweden	44.9	10.0	26.4	7.3
Switzerland	44.3	9.5	27.2	7.2
Total	42.6	11.2	25.6	7.8

Note. Weighted results.

Covariates

We use a set of standard socio-demographic covariates in modeling physical health and functional ability. These include a third degree age polynomial, educational achievement (secondary and tertiary education, primary or no education is the reference category), household size, and whether or not the individual is living together with a spouse or partner.

Our basic model includes household net worth (PPP adjusted) as an explanatory variable to reflect the opportunities for more investment in health with higher amounts of economic resources. In terms of the functional form, we use the inverse hyperbolic sine of net worth rather than the log, because a nonnegligible fraction of households have negative net worth, which indicates less access to funds for investing in health, so it is meaningful to take this into account. The inverse hyperbolic sine function is defined as $\text{IHS}(x) \equiv \log(x + \sqrt{1 + x^2})$. For positive numbers not close to zero, it is virtually indistinguishable from a logarithmic function, $\text{IHS}(x) \approx \log(2x)$. For negative numbers, $\text{IHS}(x) = -\text{IHS}(-x)$, which is approximately $-\log(-2x)$ for x not close to zero. It is zero for $x = 0$ and strictly increasing and continuously differentiable for all x.

We also include a measure of relative body weight to account for the well-documented effects of excessive body weight or obesity on physical health and functioning. Individuals are classified by relative weight based on their body mass index (BMI), calculated from self-reported weight and height as weight in kilograms divided by the square of height in meters. We use the evidence-based clinical guidelines for the classification of overweight and obesity in adults, published by the National Heart, Lung and Blood Institute of the National Institutes of Health (NIH) to stratify the study respondents into five weight classes: underweight (BMI $<$ 18.5), normal weight (18.5 $\leq$ BMI $<$ 25), overweight (25 $\leq$ BMI $<$ 30), moderate obesity (30 $\leq$ BMI $<$ 35), and

Table 4: Socio-economic and demographic variables: mean and s.d. for continuous variables and percentage for categorical ones.

	Male		Female	
Variable	Mean	s.d.	Mean	s.d.
Age (years)	64.4	10.0	66.6	11.0
Household size	2.3	1.1	2.0	1.0
Household income (€)[a,b]	44,000	48,000	38,000	46,000
Household net worth (€)[b]	460,000	1.3 million	360,000	1.1 million
Living with spouse/partner (%)	78.8		55.8	
Education (%):				
Primary	45.2		58.8	
Secondary	34.0		28.0	
Tertiary	20.8		13.1	
BMI class (%):				
Underweight	0.5		2.0	
Normal	33.0		43.7	
Overweight	50.2		36.3	
Moderately obese	13.4		13.6	
Severely obese	3.0		4.5	

[a] Annual, before taxes.

[b] PPP adjusted.

Note. Weighted results.

severe obesity (BMI $\geq$ 35). The sample size for extreme obesity (BMI $\geq$ 40), another weight class in the NIH guidelines, is too small to enable meaningful analysis. We divide the obesity group into moderate and severe obesity because there are differential health effects by degree of obesity. Severe obesity is associated with more chronic health problems than moderate obesity, and the onset is at earlier ages (Field et al., 2001; Hillier & Pedula, 2001; Must et al., 1999).

Table 4 presents sociodemographic characteristics and the BMI distribution of the sample.

3 A measurement model for health

The data for which we develop our model consist of explanatory variables collected in a vector x_n for the n-th observation and corresponding dependent variables collected in a vector y_n. The explanatory variables include a constant, continuous variables, and dummy variables. Nonlinear relationships between health and age, and between grip strength and height and weight, are accommodated by including powers and products of the original variables in the vector x_n. Categorical explanatory variables like education are transformed into a set of dummy variables.

The dependent variables consist of a combination of continuous (grip strength), binary (most mobility limitations, ADLs, and IADLs), and ordinal (climbing stairs, self-reported health)

variables. For the dependent variables, limited dependent variables must be treated quite differently from continuous variables. As with standard limited dependent variables regression models (Maddala, 1983), we assume that the binary and ordinal variables in y_n are reflections of underlying *latent response variables* y_n^*. For grip strength, $y_n = y_n^*$. For the binary and ordinal dependent variables, the relationships between y_n and y_n^* are step functions, with steps at given or estimated thresholds, as in binary and ordinal probit models.

We assume that y_n^* follows a linear structural equation model, conditional on x_n. The resulting complete model is a special case of a LISCOMP model, although it is slightly differently parameterized for computational convenience and easier interpretation. See Muthén and Satorra (1995) or Wansbeek and Meijer (2000, section 11.4) for an extensive discussion of the LISCOMP model. Our model closely resembles the health measurement submodel of Börsch-Supan, McFadden, and Reinhold (1996), although we add a continuous physical measure (grip strength), which allows us to make cross-country comparisons with relatively weak assumptions. Our approach also resembles the approach of Soldo, Mitchell, Tfaily, and McCabe (2006), who estimate an item response theory (IRT) model to compare health across cohorts in the U.S. They use similar data as we do, but select a somewhat different set of health indicators to base their analysis on.

The work of Bound et al. (1999) for U.S. data is also related, but they use the limitations as explanatory variables in the health equation, rather than as dependent variables, they do not have grip strength, and their only dependent variable is self-reported health, whereas we have 24 dependent variables. The work of Jürges (2007) is similar to this in its approach, except that he mainly uses doctor-diagnosed chronic conditions as explanatory variables, but also adds treatment for depression, grip strength, and BMI. Below we describe our specific model structure in more detail.

The linear structural equation model

The central variable in our measurement model is health, denoted by the symbol η_n. This is a possibly multidimensional *latent* or unobserved variable. The latent response variables y_n^* are assumed to depend on η_n, and η_n in turn depends on the explanatory variables x_n. With the exception of grip strength (and the intercepts), the dependence of y_n^* on x_n is assumed to be channeled through η_n.

Specifically, we assume that the relation between the latent response variables and the latent health variables and observed explanatory variables is of the following form:

$$y_n^* = T x_n + \Lambda \eta_n + \varepsilon_n, \tag{1}$$

where ε_n is a vector of residuals, which we will usually call *measurement errors*. If y_n^* is continous and observable, and if $T = 0$ and $\Lambda = I$, then (1) is a traditional multivariate measurement error model, which explains our usage of this term.

The matrix T is a matrix of regression coefficients and Λ is a matrix of factor loadings. Compared to the most common model specifications for structural equation models, the term Tx_n is added, although it is also used in the model underlying the M*plus* program (Muthén, 1998–2004, Appendix 2). This term contains the intercepts. In addition, it contains a direct effect of height and weight on grip strength. This allows for the fact that grip strength will be affected by an individual's height and weight, irrespective of one's health. Preliminary exploration suggested that a second-degree polynomial captures this relation well.

We interpret (1) as a causal structural relation, with the possible exception of the polynomial in height and weight for grip strength, which has a more reduced form interpretation. But it is important in our model development that we assume that the latent response variables structurally depend on health.

The dependence of η_n on x_n is modeled through the (multivariate) regression equation

$$\eta_n = \Gamma x_n + \zeta_n, \tag{2}$$

where ζ_n is a vector of random errors (disturbances), and Γ is a matrix of regression coefficients. In contrast with (1), our assumptions in using (2) are quite modest. In particular, it makes little sense to view this as a structural health production function, because such a function should have a strong dynamic component, with current health depending on past investments in health over a longer period of time and not just a few contemporaneous covariates. Instead, (2) has the flavor of a reduced form model, although net worth cannot be assumed to be exogenous. Therefore, our term for this equation would be a "predictive equation". We will discuss the formal status and assumption in more detail below.

The observation function

The variables y_n^* are not necessarily observable. We will denote the relationship between the latent response variables y_n^* and the observed dependent variables y_n by the generic expression $y_n = H(y_n^*)$, and we call $H(\cdot)$ the *observation function*. The observation function is such that each y_{ni} depends only on its latent response counterpart y_{ni}^*, i.e., $y_{ni} = H_i(y_{ni}^*)$ for the i-th dependent variable. For grip strength, $H_i(y_{ni}^*) = y_{ni}^*$, the identity function, whereas for the binary dependent variables, $H_i(y_{ni}^*) = I(y_{ni}^* > 0)$, where $I(\cdot)$ is the indicator function, i.e., its value is 1 if its argument is true and 0 otherwise. Thus, $H_i(y_{ni}^*)$ is a step function with a single step from 0 to 1 at the threshold 0. There are two ordinal dependent variables: limitations with climbing stairs and self-reported health. For these, the observation function is a step function with multiple thresholds:

$$H_i(y_{ni}^*) = c_{ij} \Leftrightarrow \alpha_{i,j-1} < y_{ni}^* \leq \alpha_{ij},$$

where c_{ij} is the j-th category of the i-th variable ($\{0, 1, 2\}$ for climbing stairs, $\{1, 2, 3, 4, 5\}$ for self-reported health), and the α's are thresholds, with $\alpha_{i,0} \equiv -\infty$ and $\alpha_{i,J_i} \equiv +\infty$, where J_i is the number of categories of y_{ni} ($J_i = 3$ for climbing stairs and $J_i = 5$ for self-reported health). It is now convenient to treat the remaining thresholds as free parameters and normalize the intercepts for these ordinal variables to zero, as is common in ordinal probit models.

Covariance parameters, distributional assumptions, and the predictive equation

To complete the model, we define $\Psi \equiv \text{Cov}(\zeta_n)$ and $\Omega \equiv \text{Cov}(\varepsilon_n)$. The matrix Ψ is usually unrestricted, although for some models, it may be block diagonal. As usual (in factor analysis and its generalizations), we assume that Ω is a diagonal matrix, so that the dependent variables are conditionally independent given the latent variables. Furthermore, in line with the standard LISCOMP model and probit models, we assume that the residuals and measurement errors are normally distributed. However, it will turn out that with our estimation method, it is fairly straightforward to use other distributional assumptions for the residuals, which can be used for sensitivity analyses in assessing to what extent the results are driven by the normality assumption. We leave this for future research.

We can now be more precise about the formal assumptions concerning the explanatory variables x_n and the predictive health equation. The latter, i.e., (2) is now formally interpreted as meaning

$$(\eta_n \mid x_n) \sim \mathcal{N}(\Gamma x_n, \Psi).$$

Thus, it is an assumption about a conditional distribution, without being causal or structural. In addition, it now follows that we assume that the dependent variables are conditionally independent of the explanatory variables given the latent variables, with the exception of grip strength, which is allowed to depend on height and weight directly.

Identification, normalizations, and cross-country comparability

In models with latent variables, many restrictions on the parameters are typically needed to obtain an identified model. Our model is no exception. Each latent variable must be assigned a location and a scale. This holds for the latent response variables y^* as well. For these, we use the probit convention of fixing the scales by normalizing the variances of the residuals in the matrix Ω to 1 and, as already mentioned above, fixing the locations by normalizing the thresholds to 0 for the binary dependent variables, and normalizing the intercept to 0 for the ordinal ones. For grip strength, no such normalizations are necessary, because the location and scale of y^*_{ni} are determined by the identity $y_{ni} = y^*_{ni}$.

The location and scale of η can be assigned in different ways. For our purposes, the most convenient and useful normalization is to assign a reference variable from the list of indicators for each element of η. The factor loading relating the reference indicator to this element of η is normalized to 1, and the intercept (or one of the thresholds in case of an ordinal variable) of this reference indicator is normalized to 0.

The arbitrariness of the locations and scales of the latent variables affects the extent to which parameters, and derived statistics such as the marginal distributions of the latent variables or the constructed health indexes, can be compared across countries. For example, if the threshold for reporting a certain type of difficulty is higher in one country than in another country for the same true health, but these thresholds are normalized to the same value for this variable, then

the difference shows up as an apparent difference in genuine health. A similar story can be told for different factor loadings: if a certain activity is more sensitive to health in one country than in another, the factor loadings are different. But if they are normalized to be the same, this shows up as a difference in health distributions. We assume that grip strength does not suffer from such problems of cross-country differences, and therefore use grip strength as our first reference indicator. This should make the location and scale of the first health dimension comparable across countries. Furthermore, we assume that the relationship between grip strength and height and weight is the same for each country, i.e., that the coefficients of the height-weight polynomial are the same.

With more than one health dimension, selection of the other reference indicators is more problematic, because all of the remaining indicators may suffer from country-specific response patterns. Hence, cross-country comparability of the health indexes beyond the first dimension cannot be assumed. The current paper is limited to one health dimension and thus does not suffer from this problem, but in future extensions, this will become important.

All other parameters are allowed to be different across countries, to account for cultural variation in response patterns, different educational and health systems, and other cross-country attributes that may give rise to country-specific parameters.

3.1 Estimation

It is customary to estimate LISCOMP models in several steps. In the first step, the reduced form of the linear structural equation model is obtained: $y_n^* = \Pi x_n + u_n$, with $\Pi = T + \Lambda\Gamma$ and $u_n \sim \mathcal{N}(0, \Sigma)$, with $\Sigma = \Lambda\Psi\Lambda' + \Omega$. The elements of Π and the threshold parameters are estimated from univariate linear regressions and probits. Here, Π is unrestricted. In the second step, the estimates from the first step are fixed and the elements of Σ are estimated from bivariate likelihoods. An estimate $\hat{W}$ of the joint asymptotic covariance matrix of the estimators in the first two steps is obtained by writing the estimators as the joint solution of a system of generalized estimating equations (GEE) and applying the standard GEE theory. In the third step, the estimates of the elements of Π and Σ thus obtained are gathered in the vector $\hat{s}$. Furthermore, its population value σ is written as a function of the free parameters θ: $\sigma = \sigma(\theta)$. Then the parameter vector θ is estimated by minimizing the distance function $F = (\hat{s} - \sigma(\theta))' \hat{W}^{-1} (\hat{s} - \sigma(\theta))$. The asymptotic covariance matrix of the resulting estimator $\hat{\theta}$ can then be obtained from standard minimum distance theory. See Muthén and Satorra (1995) or Wansbeek and Meijer (2000, pp. 332–338) for more details.

This procedure is very fast and tends to work very well in most applications. Therefore, this is the method implemented in structural equation modeling programs like LISREL (Jöreskog & Sörbom, 1993) and M*plus* (Muthén, 1998–2004). However, this procedure breaks down for our data. The problem is that many variables have very little variation. For several binary dependent variables, only a small percentage of respondents report limitations, and similarly small percentages are observed for some of the explanatory variables, most notably being underweight or severely obese. Therefore, it often happens that a certain cross-table of a binary dependent

variable and a dummy explanatory variable has an empty cell: only three out of the four possible combinations occur in the data. The result is that the corresponding coefficient in the reduced-form probit model is (plus or minus) infinity.

We have solved this problem by writing our own estimation program in `Stata`. It estimates the parameters directly by maximum simulated likelihood. The restrictions that the model imposes on Π and Σ are generally sufficient to estimate the model. However, this procedure is slow and reasonable starting values must be supplied in order to start the estimation process. Without the latter, the program is often not able to find feasible parameter values. We will now describe the method in more detail.

Loglikelihood

It is convenient to reparameterize the covariance matrices in the model. Let C be the Cholesky root of Ψ, so that $\Psi = CC'$ and C is a lower triangular matrix. Define $\xi_n \equiv C^{-1}\zeta_n$, so that $\xi_n \sim \mathcal{N}(0, I)$ and $\zeta_n = C\xi_n$. Analogously, let the diagonal element of Ω corresponding to grip strength be ω^2. The other diagonal elements of Ω are all equal to 1.

With this parameterization, the conditional likelihood of the n-th observation, given both x_n and ξ_n, is

$$\mathcal{L}_{n,\text{cond}} = \prod_{i=1}^{M} f_i(y_{ni} \mid x_n, \xi_n),$$

where $M = 24$ is the number of dependent variables in the model, and $f_i(y_{ni} \mid x_n, \xi_n)$ is the conditional probability mass function of the i-th variable, or the conditional density for grip strength. Specifically, let $\delta_n = (T + \Lambda\Gamma)x_n + \Lambda C\xi_n$, and let δ_{ni} be its i-th element. Then for the binary indicators,

$$f_i(y_{ni} \mid x_n, \xi_n) = [\Phi(\delta_{ni})]^{y_{ni}}[1 - \Phi(\delta_{ni})]^{1-y_{ni}},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. For the ordinal indicators, if $y_{ni} = c_{ij}$, the j-th category of this variable, then

$$\begin{aligned} f_i(y_{ni} \mid x_n, \xi_n) &= \Pr(\alpha_{i,j-1} < y^*_{ni} \leq \alpha_{ij}) \\ &= \Pr(\alpha_{i,j-1} - \delta_{ni} < \varepsilon_{ni} \leq \alpha_{ij} - \delta_{ni}) \\ &= \Phi(\alpha_{ij} - \delta_{ni}) - \Phi(\alpha_{i,j-1} - \delta_{ni}). \end{aligned}$$

For grip strength,

$$f_i(y_{ni} \mid x_n, \xi_n) = \frac{1}{\omega}\phi\left(\frac{y_{ni} - \delta_{ni}}{\omega}\right),$$

where $\phi(\cdot)$ is the standard normal density function.

The unconditional likelihood then simply becomes the product of the conditional likelihood and the density of ξ_n:

$$\mathcal{L}_{n,\text{uncond}} = \mathcal{L}_{n,\text{cond}} f_\xi(\xi_n) = \left[\prod_{i=1}^{M} f_i(y_{ni} \mid x_n, \xi_n)\right] f_\xi(\xi_n).$$

Let $K \geq 1$ be the number of health dimensions in the model. Then the density $f_\xi(\xi_n)$ is the product of K standard normal density functions, one for each element of ξ_n.

However, because ξ_n is not observed, $\mathcal{L}_{n,\text{uncond}}$ cannot be used directly. The likelihood of the observed variables is obtained by integrating ξ_n out. Thus we obtain

$$\mathcal{L}_n = \int \mathcal{L}_{n,\text{uncond}} \, \mathrm{d}\xi_n = \int \left[\prod_{i=1}^{M} f_i(y_{ni} \mid x_n, \xi_n)\right] f_\xi(\xi_n) \, \mathrm{d}\xi_n, \tag{3}$$

and the loglikelihood of the n-th observation is the logarithm of this: $L_n = \log \mathcal{L}_n$. The loglikelihood of the sample then becomes

$$L = \sum_{n=1}^{N} L_n,$$

where N is the sample size. However, SHARE is (largely) a probability sample, with sampling weights provided. Therefore, with w_n being the weight for the n-th individual, we instead use the log-pseudolikelihood function

$$L_w = \sum_{n=1}^{N} w_n L_n. \tag{4}$$

Missing data handling

As noted above, our data set contains observations with partly missing data. From preliminary analyses, we concluded that missingness is systematic. In particular, there is a fairly large group of respondents with missing grip strength, and this group reports substantially more limitations. Apparently, grip strength is often missing because of bad health. Consequently, removing all observations with missing data biases the sample and most likely will distort the parameter estimates as well.

For missing dependent variables (like grip strength), it is common in this type of model to assume "missing at random" (MAR), which is a weaker form of randomness than "missing completely at random" (MCAR). MCAR assumes that the distribution of all random variables is the same for observations with missings and observations without missings, whereas MAR allows systematic differences that depend on non-missing variables. See Rubin (1976) and Little and Schenker (1995) for a formal definition and analysis of MAR. The MAR assumption in our case means that the missingness of a variable may depend on the values of the observed variables

(dependent and explanatory), but not on the values of unobserved variables, in particular the value of the missing variable itself, and the latent variables (health). It is likely that this assumption is violated to some extent, but because the probability of missingness is allowed to depend on the values of the observed (non-missing) variables, the distortion is probably not large, because the observed variables will be able to account for most of the dependence of missingness on the unobserved variables. If this is not the case, selection models in the spirit of Heckman (1979) could be estimated, but Little and Schenker (1995) and Jamshidian and Bentler (1999) argued that these will be very sensitive to minor misspecifications and thus may make things worse rather than better. Simulation studies have found excellent properties of the MAR-based method if the MAR assumption is met, and reasonably good properties when it is not (e.g., Muthén, Kaplan, & Hollis, 1987; Jamshidian & Bentler, 1999). Therefore, we use this method. The practical implementation is that in the loglikelihood, the factor $f_i(y_{ni} \mid x_n, \xi_n)$ is replaced by 1 if the i-th variable is missing for the n-th respondent.

Missing covariates are more problematic in principle, because we would prefer to make no assumptions about their (conditional) distributions, and all dependent variables depend on them. A common practical solution, which we adopt here, is to set the value of a missing covariate to an arbitrary fixed value (zero) and add a dummy variable for "missingness".

Thus, if education is missing for the n-th respondent, we set secondary and tertiary education to zero for this respondent, and set the dummy `missedu` to 1, whereas this dummy is zero if education is observed. So we now have three education dummies: "secondary", "tertiary", and "missing", with primary or no education as the reference group. In this way, the `missedu` dummy picks up the (average) main effect of the missing variable. In addition to education, this method has also been applied for height, weight, and the BMI dummies. The grip strength equation contains a second degree polynomial in height and weight. In most situations where height was missing, body weight was missing as well and vice versa, so that all five terms (height, weight, the squares of both, and their product) are missing. In only a few cases, one of them was missing but not the other, mainly because we replaced an implausible value by missing. To avoid having to introduce additional dummies for only a few observations, we have taken a somewhat unorthodox approach of replacing the observed height or weight by missing as well if the other was missing. In this way, we can use a single dummy "missing height and weight" to capture the average main effect. Because the BMI dummies need both height and weight, they are missing as well if height and/or weight is missing. Therefore, the "missing height and weight" dummy is also used in the health equation as an additional BMI category.

Maximum simulated likelihood

In principle, obtaining maximum (pseudo-)likelihood estimators is now straightforward. We only have to maximize L or L_w as defined above. However, maximizing the loglikelihood is often a formidable task, because the integrals do not have a closed form solution. We use *maximum*

simulated likelihood (MSL) to solve this problem. Observe that (3) can be written as

$$\mathcal{L}_n = \mathrm{E}_{\xi}\left[\prod_{i=1}^{M} f_i(y_{ni} \mid x_n, \xi_n)\right], \tag{5}$$

where the expectation is over the distribution of ξ_n, i.e., the normal distribution with zero mean and identity covariance matrix. It now follows immediately that this can be approximated to any desired degree of accuracy by drawing a (pseudo-)random sample from the distribution of ξ_n and computing the resulting sample average of the term in brackets in (5):

$$\check{\mathcal{L}}_n = \frac{1}{R}\sum_{r=1}^{R}\prod_{i=1}^{M} f_i(y_{ni} \mid x_n, \check{\xi}_{nr}),$$

where $\check{\xi}_{nr}$ is the r-th drawing from the distribution of ξ_n for the n-th observation. The approximation can be made arbitrarily accurate by generating a larger sample (increasing R).

The simulated log-(pseudo)likelihood is now obtained by replacing $\mathcal{L}_n$ by $\check{\mathcal{L}}_n$ in the formula for L or L_w, and the MSL estimators are obtained by maximizing the resulting function. Properties of MSL estimators are derived in Gouriéroux and Monfort (1991), Lee (1995), and Train (2003, chap. 10). For our purposes, the theory implies that MSL estimators have the same properties as the estimators that use the exact integrals, as long as the approximation is close enough. To increase accuracy without unduly increasing the computation time, we use *Halton sequences*, which is a more systematic (nonrandom) method to generate drawings in a way that reduces variance and thus increases precision. Train (2003, pp. 224–238) gives a detailed description. A function generating Halton sequences is supplied with `Stata` (Drukker & Gates, 2006). Based on earlier experience, some experimentation, and the remarks in Train (2003, p. 231), we assumed that $R = 100$ Halton draws should be sufficient. However, we experimented a little (for Germany) with $R = 1000$ and $R = 5000$. The differences between 100 and 1000 draws are noticeable but relatively small. The differences between 1000 and 5000 draws are negligible. More importantly, none of these differences leads to substantively different conclusions, and the resulting health indexes are very highly correlated (0.999). Therefore, the results here have all been obtained using 100 draws, with the exception for Germany, where we use the results with 1000 draws.

Two-step approach to circumvent computational problems

Almost all parameters in the model are either fixed to 0 or 1, or are country-specific free parameters. If this would hold for all parameters, the model could be estimated for each country separately, which would be computationally preferable. However, the coefficients of the height-weight polynomial in the grip strength equation are assumed to be equal across countries. Because of these cross-country restrictions on the parameters, estimation should ideally be done jointly for all countries. Unfortunately, this is computationally prohibitive. Therefore, we take a two-step approach. In the first step, we insert the predictive health equation (2) into the

grip strength equation of (1) and estimate the resulting reduced-form parameters jointly for all countries. Then we subtract the estimated height-weight polynomial from grip strength. In the second step, we use the residual grip strength as an ordinary indicator. Because there are no joint parameters left, the remaining model is estimated separately for each country.

There are a few technical complications, however, including the treatment of missing grip strength. We solve it by estimating a Heckman-type sample selection model, with all other health indicators included in the selection equation. Details are given in Appendix B. This appendix shows that, apart from innocuous location and scale differences, the residual grip strength is extremely insensitive to variations in the model specification. Consequently, we can be confident that the residual grip strength obtained in the first step is a useful indicator in the second step.

4 Estimating the latent variables

In many situations, we would like to estimate the values of the latent variables themselves. Given the MSL setup, this turns out to be fairly straightforward. The conditional density of the latent variables for the n-th individual, given the observed data is

$$f_{\eta|y}(\eta_n \mid y_n) = \frac{f_{y|\eta}(y_n \mid \eta_n)\, f_\eta(\eta_n)}{f_y(y_n)} = \frac{f_{y|\eta}(y_n \mid \eta_n)\, f_\eta(\eta_n)}{\int f_{y|\eta}(y_n \mid \eta_n)\, f_\eta(\eta_n)\ \mathrm{d}\eta_n}$$

or, analogously,

$$f_{\xi|y}(\xi_n \mid y_n) = \frac{f_{y|\xi}(y_n \mid \xi_n)\, f_\xi(\xi_n)}{\int f_{y|\xi}(y_n \mid \xi_n)\, f_\xi(\xi_n)\ \mathrm{d}\xi_n}.$$

Furthermore,

$$\mathrm{E}(\xi_n \mid y_n) = \frac{\int \xi_n f_{y|\xi}(y_n \mid \xi_n)\, f_\xi(\xi_n)\ \mathrm{d}\xi_n}{\int f_{y|\xi}(y_n \mid \xi_n)\, f_\xi(\xi_n)\ \mathrm{d}\xi_n} = \frac{\mathrm{E}_\xi[\xi_n f_{y|\xi}(y_n \mid \xi_n)]}{\mathrm{E}_\xi[f_{y|\xi}(y_n \mid \xi_n)]}$$

and

$$\mathrm{E}(\xi_n\xi_n' \mid y_n) = \frac{\int \xi_n\xi_n' f_{y|\xi}(y_n \mid \xi_n)\, f_\xi(\xi_n)\ \mathrm{d}\xi_n}{\int f_{y|\xi}(y_n \mid \xi_n)\, f_\xi(\xi_n)\ \mathrm{d}\xi_n} = \frac{\mathrm{E}_\xi[\xi_n\xi_n' f_{y|\xi}(y_n \mid \xi_n)]}{\mathrm{E}_\xi[f_{y|\xi}(y_n \mid \xi_n)]},$$

from which $\mathrm{Cov}(\xi_n \mid y_n)$ immediately follows. In these expressions, E_ξ denotes the expectation over the marginal distribution of ξ_n, i.e., the normal distribution with zero mean and identity covariance matrix. All expressions are also implicitly conditional on x_n. The expressions in the denominators are equal to $\exp(L_n)$, where L_n is the loglikelihood contribution introduced earlier. The expressions in the numerators have a similar form, which means that their computation can be done by a simple extension of the simulated likelihood program.

The resulting expressions for the latent variables η_n are now obtained as

$$\begin{aligned}\mathrm{E}(\eta_n \mid y_n) &= \Gamma x_n + C\,\mathrm{E}(\xi_n \mid y_n)\\ \mathrm{Cov}(\eta_n \mid y_n) &= C\,\mathrm{Cov}(\xi_n \mid y_n)C'.\end{aligned}$$

Of course, in practice, we compute these expressions with the parameters replaced by their estimates from the previous model estimation, as given in section 3 above. The estimate $\hat{\eta}_n$ of $\mathrm{E}(\eta_n \mid y_n)$ is our proposed health index. This is computed for each observation.

4.1 Using the results in a subsequent model

An important reason for estimating the latent variables is to use these estimates as dependent or explanatory variables in other models. From a strict statistical viewpoint, if the model is correctly specified, it is preferable to estimate both models jointly, and there is no need to estimate the latent variable separately. In practice, however, there are several reasons why a two-stage approach may be preferred. First, there may be many indicators of the latent variables, and thus a complete joint model may be very large, becoming impractical to estimate. This is even more compelling if the second stage model is a complicated nonlinear model, such as a dynamic programming model. Second, related to the first point, one may wish to do an extensive specification search for a suitable second-stage model, and such a model search procedure is severely hampered by the need to estimate the complete measurement model over and over again.

Initially, assume that the parameters of the measurement model (first-stage model) are known. The differences between parameter estimators and true parameter values are of order $O_p(N^{-1/2})$, whereas the differences between latent variables and their estimators conditional on the observed variables are $O_p(1)$, so that in large samples the results based on this assumption will give good approximations. From the formulas above, we can write

$$\hat{\eta}_n = \Gamma x_n + C\hat{\xi}_n,$$

where $\hat{\xi}_n = \mathrm{E}(\xi_n \mid y_n)$. Hence,

$$\xi_n = \hat{\xi}_n + (\xi_n - \hat{\xi}_n) = \hat{\xi}_n + v_n,$$

where $\mathrm{E}(v_n \mid y_n) = 0$ and $\mathrm{E}(v_n \hat{\xi}_n' \mid y_n) = 0$. Let $w_n = Cv_n$, then

$$\eta_n = \hat{\eta}_n + w_n,$$

with $\mathrm{E}(w_n \mid y_n) = 0$ and $\mathrm{E}(w_n \hat{\eta}_n' \mid y_n) = 0$.

To illustrate how estimated latent variables are used in a subsequent model, we discuss a few examples. As a first example, assume that the model of interest is a linear regression model with dependent variable q_n and explanatory variables p_n and η_n:

$$q_n = \beta' p_n + \gamma' \eta_n + u_n,$$

with all the usual assumptions. We can rewrite this as

$$q_n = \beta' p_n + \gamma' \hat{\eta}_n + \hat{u}_n,$$

where $\hat{u}_n = u_n + \gamma' w_n$. Given that w_n is uncorrelated with $\hat{\eta}_n$, OLS is consistent, provided that w_n is uncorrelated with p_n as well. This is the *Berkson model* (e.g., Wansbeek & Meijer, 2000, p. 30). If p_n would be correlated with w_n, p_n would contain information about η_n (or ξ_n) that has not been used before. Thus, an improved measurement model would be possible. On the other hand, no model is perfect, so we may wish to allow for such correlation. In that case, IV estimation would be necessary. In general, we may conclude that for a linear regression model, consistent estimation of the parameters is straightforward using estimated latent variables.

A second application is a (binary) probit model, where q_n is replaced by the latent response variable q_n^*, and we now have

$$q_n^* = \beta' p_n + \gamma' \hat{\eta}_n + \hat{u}_n,$$

and we observe $q_n = I(q_n^* > 0)$. Normalizing the variance of u_n to 1 (as is usual in probit models), the variance of $\hat{u}_n$ is $\sigma_{\hat{u}}^2 = 1 + \gamma' \Sigma_w \gamma$, where $\Sigma_w = C \operatorname{Var}(\xi_n \mid y_n) C'$. Clearly, if $\hat{u}_n$ is normally distributed, the probit estimators are consistent estimators of $\beta/\sigma_{\hat{u}}$ and $\gamma/\sigma_{\hat{u}}$. Consistent estimators for β and γ are straightforwardly obtained because Σ_w is known. (In the more general case, it is estimated consistently.) However, normality of w_n is only obtained if all observed indicators are continuous and conditionally normally distributed. The density of $\xi_n \mid y_n$ is

$$f_{\xi|y}(\xi_n \mid y_n) = \frac{f_{y|\xi}(y_n \mid \xi_n)\, f_\xi(\xi_n)}{f_y(y_n)}.$$

Here, $f_\xi(\xi_n)$ is a multivariate normal density and $f_y(y_n)$ is a normalizing constant. But if y contains, e.g., discrete indicators, $f_{y|\xi}(y_n \mid \xi_n)$ is a nonlinear function of ξ_n such that the resulting density $f_{\xi|y}(\xi_n \mid y_n)$ does not reduce to a multivariate normal density. Hence, the probit estimators will in general be inconsistent estimators of the parameters of interest, and cannot be easily corrected. On the other hand, discrete choice models tend to be largely insensitive to the choice of distribution, so that using a simple probit may still be the preferred choice. In section 7 we will only present simple probit results.

The theoretically best solution is to use simulated likelihood for the probit model as well, using the full conditional distribution of q_n. Evidently,

$$\Pr(q_n = 1 \mid \xi_n) = \Phi\Big(\beta' p_n + \gamma'(\Gamma x_n + C\xi_n)\Big),$$

so that

$$\Pr(q_n = 1 \mid y_n) = \int \Phi\Big(\beta' p_n + \gamma'(\Gamma x_n + C\xi_n)\Big)\, f_{\xi|y}(\xi_n \mid y_n)\ \mathrm{d}\xi_n.$$

Because the parameters Γ and C and the density $f_{\xi|y}(\xi_n \mid y_n)$ are known (or more generally, consistently estimated), the parameters of interest can be straightforwardly estimated by applying maximum simulated likelihood to this equation.

As presented here, this method still does not take the variability of the parameter estimators into account. As argued above, this should be fine in large samples, but in moderate samples, the uncertainty about the parameters may have a noticeable effect on the precision of the estimate of

η. One way to incorporate this uncertainty is preceding each draw in the MSL estimation of the probit model by a draw of a parameter vector from a normal distribution with mean equal to the parameter estimates and covariance matrix equal to the estimated asymptotic covariance matrix of the original parameter estimators. Such a procedure, which has a Bayesian flavor, was proposed by Hamilton (1986) in the related context of estimating the variability of the estimated states in a state-space model. Another way to incorporate uncertainty in the parameters into the subsequent estimation procedure would be to use the delta method in some way. We will leave these issues for future research.

5 Estimation results of the measurement model

As mentioned above, we estimate the models separately for each country-gender combination. Here, we only present results with one latent dimension, leaving multidimensional health variables for future research. The variables used in the models have been discussed in section 2 above, but we have linearly transformed some explanatory variables to obtain better scaling and less multicollinearity. See the appendix for a detailed account of variable construction. For each analysis, the fit of the model with the latent health dimension is compared to the fit of the *null model*. The null model is the analog of the constant-only model in linear regression. In our case, the null model is the model without the latent variable η, and thus without Λ, Γ, C, and x. This model is also called the *independence model*, because it implies that all dependent variables are independently distributed of each other and of the explanatory variables. As expected, the model with the latent variable included fits much better than the model without it: The Scaled LR test statistic (see the appendix for an explanation) is always extremely large and significant, and the information criteria AIC and BIC are much smaller for this model than for the null model.

Table 5 presents the corresponding pseudo-R^2 measures. These are defined as $1 - \check{L}_{(1)}/L_{(0)}$, where $\check{L}_{(1)}$ is the maximized simulated log-pseudolikelihood for the target model with one health dimension and $L_{(0)}$ is the maximized log-pseudolikelihood for the null model. Because of the absence of a latent variable, the latter does not involve simulation. The results in Table 5 indicate that the latent variable explains a sizeable amount of the variation in the data, but more experience with the pseudo-R^2 in this type of model is needed to be able to judge whether the values reported here are "good".

Tables 6–9 present the ranges of the estimates and their t-statistics for the intercepts (T), factor loadings (Λ), predictive health equation (Γ and C), and other parameters (Ω, α's), respectively. The values of the intercepts are a bit difficult to interpret, but given the unit residual variances associated with the standard probit specifications, and the generally large t-values, we can conclude that there is considerable cross-country variation. Also, the generally large negative values (combined with the values of the factor loadings and the distribution of latent health) reflect the small number of difficulties that is typically reported, and thus the high threshold for reporting a difficulty, because the intercept can also be interpreted as the negative of a threshold, with the

Table 5: Pseudo-R^2 values for the health model with one latent variable.

Country	Male	Female
Austria	0.23	0.24
Belgium	0.21	0.26
Denmark	0.26	0.25
France	0.26	0.26
Germany	0.23	0.25
Greece	0.26	0.26
Israel	0.31	0.35
Italy	0.25	0.30
The Netherlands	0.18	0.29
Spain	0.28	0.29
Sweden	0.28	0.27
Switzerland	0.17	0.20

intercept being zero.

The factor loadings have the expected sign: negative, meaning that better underlying health gives fewer limitations and better self-reported health. Almost all factor loadings are statistically significant, most of them very strongly. But again, even from this highly condensed table, it can be seen that there are substantial differences across countries and gender.

Table 8 shows the estimation results for the "predictive" equation for the latent health variable η. This has some expected patterns: higher education and higher wealth tend to be associated with better health and being overweight or obese tends to be related with worse health. The coefficients of age and its square and cube are a bit difficult to interpret by themselves, although it's clear that the linear part points at the expected negative relationship between health and age. Plots of the cubic polynomial and pointwise confidence bands around them show that this negative slope generally holds for the complete polynomial, but there are some exceptions at the highest ages. At these ages, however, the confidence bands are very wide.

From Table 9, we can learn that the estimated residual standard deviations for the grip strength equation are fairly similar across countries, even though the health equations are different. We interpret this as an indication that the model is able to capture general health fairly well in a cross-country comparable way and that our assumption that grip strength is not subject to cross-country reporting differences is warranted. The cross-country differences between the threshold parameters for the ordinal indicators (climbing stairs and self-reported health) are larger, comparable to the results for the intercepts. Closer scrutiny of the original estimation results indicates that the cross-country differences in the *differences* between adjacent thresholds are much smaller, which suggests that differential reporting behavior may only be due to a uniform shift.

As a result of the estimates, we can compute the (unconditional) mean and standard deviation of health for each country-gender combination, where for this computation, the x variables are treated as random variables. Also, we computed the health indexes as discussed in section 4

Table 6: Intercepts (T).

	Male				Female			
	Estimates		t-values		Estimates		t-values	
Indicator	Min	Max	Min	Max	Min	Max	Min	Max
Mobility limitations								
Walk 100m	−4.121	−1.630	−12.0	−5.8	−3.204	−1.429	−16.9	−6.1
Sit 2hrs	−2.713	−1.160	−22.6	−7.5	−2.467	−0.917	−26.8	−7.2
Get up from chair	−2.371	−0.593	−16.8	−3.0	−2.659	−0.443	−19.1	−5.1
Climbing stairs	0	0	n.a.	n.a.	0	0	n.a.	n.a.
Stoop	−2.143	−0.333	−13.1	−1.5	−2.159	0.066	−10.6	0.6
Reach	−3.567	−1.297	−22.8	−7.0	−2.518	−1.033	−27.4	−8.4
Pull	−3.765	−1.441	−15.2	−4.7	−3.027	−0.733	−16.6	−5.9
Lift 5kg	−3.558	−0.958	−16.8	−2.8	−1.639	−0.339	−13.5	−3.2
Pick up coin	−4.717	−1.608	−16.3	−5.5	−3.313	−1.592	−24.5	−8.2
ADLs								
Dress	−4.670	−1.593	−15.0	−5.6	−5.355	−1.838	−18.0	−5.2
Walk room	−9.584	−2.942	−9.4	−2.6	−7.166	−2.382	−9.4	−5.1
Bath	−9.618	−3.063	−10.5	−2.5	−8.020	−2.082	−17.1	−4.4
Eat	−6.869	−2.730	−12.6	−4.1	−5.365	−2.337	−12.3	−2.2
Get out of bed	−6.024	−2.213	−10.4	−3.6	−6.539	−2.177	−13.1	−5.6
Use toilet	−12.308	−2.333	−13.4	−3.0	−9.455	−2.761	−10.2	−4.2
IADLs								
Use map	−4.631	−1.614	−17.3	−6.2	−2.386	−1.082	−23.0	−7.3
Prepare hot meal	−6.326	−1.889	−10.5	−3.3	−7.646	−3.139	−9.7	−2.3
Shop for groceries	−13.922	−3.037	−8.5	−2.4	−6.996	−2.242	−11.1	−4.4
Phone calls	−8.978	−2.291	−17.9	−4.0	−5.041	−2.875	−11.8	−3.3
Take medication	−13.212	−2.721	−11.1	−2.8	−6.074	−3.016	−10.4	−1.3
Work around house	−7.509	−1.606	−10.8	−4.3	−4.550	−1.078	−13.2	−6.4
Manage money	−6.150	−2.411	−14.4	−4.3	−4.196	−2.016	−15.2	−6.1
Self-reported health	0	0	n.a.	n.a.	0	0	n.a.	n.a.
Grip strength resid.	0	0	n.a.	n.a.	0	0	n.a.	n.a.

Table 7: Factor loadings (Λ).

	Male				Female			
	Estimates		*t*-values		Estimates		*t*-values	
Indicator	Min	Max	Min	Max	Min	Max	Min	Max
Mobility limitations								
Walk 100m	−4.572	−1.591	−8.8	−4.5	−4.838	−2.952	−11.6	−4.0
Sit 2hrs	−2.355	−0.961	−7.3	−3.6	−2.402	−1.282	−9.4	−3.8
Get up from chair	−3.120	−1.329	−8.7	−4.1	−4.139	−2.106	−12.5	−5.3
Climbing stairs	−5.251	−1.839	−11.5	−5.4	−4.522	−3.035	−14.3	−5.3
Stoop	−3.597	−1.779	−10.3	−4.3	−4.300	−2.682	−13.9	−3.2
Reach	−3.003	−1.272	−7.5	−3.7	−3.351	−1.597	−10.7	−4.0
Pull	−5.469	−1.710	−8.9	−3.5	−4.169	−2.574	−13.7	−3.8
Lift 5kg	−5.762	−1.828	−9.3	−4.4	−3.643	−2.079	−13.8	−3.3
Pick up coin	−3.490	−1.534	−6.6	−3.1	−2.965	−1.393	−8.3	−3.3
ADLs								
Dress	−4.223	−2.040	−8.5	−3.1	−5.858	−2.731	−10.3	−3.8
Walk room	−11.836	−3.172	−5.8	−2.2	−8.955	−2.815	−6.4	−3.5
Bath	−12.223	−4.004	−6.6	−2.2	−10.215	−3.633	−11.3	−3.9
Eat	−5.072	−1.989	−5.4	−3.0	−6.391	−2.162	−6.6	−1.5
Get out of bed	−4.851	−2.262	−6.4	−2.4	−6.399	−2.405	−7.9	−3.1
Use toilet	−7.429	−1.050	−5.3	−2.1	−8.865	−3.044	−6.5	−2.7
IADLs								
Use map	−3.693	−1.794	−8.3	−3.4	−3.012	−1.545	−11.0	−3.9
Prepare hot meal	−4.342	−2.333	−6.6	−2.0	−10.828	−4.241	−6.9	−1.8
Shop for groceries	−12.631	−3.398	−6.3	−2.4	−11.679	−4.698	−9.1	−3.7
Phone calls	−5.172	−1.760	−5.4	−2.3	−5.041	−2.499	−6.5	−2.2
Take medication	−9.448	−1.419	−5.3	−2.4	−7.605	−2.466	−6.3	−1.1
Work around house	−6.313	−2.834	−7.8	−3.2	−7.436	−3.739	−10.8	−5.1
Manage money	−5.827	−2.110	−6.5	−2.7	−4.234	−2.993	−8.9	−3.2
Self-reported health	−3.525	−0.932	−11.5	−4.7	−3.146	−2.018	−15.6	−4.6
Grip strength resid.	1	1	n.a.	n.a.	1	1	n.a.	n.a.

Table 8: Predictive health equation.

	Male				Female			
	Estimates		*t*-values		Estimates		*t*-values	
Predictor	Min	Max	Min	Max	Min	Max	Min	Max
Aged1	−0.431	−0.104	−8.2	−3.4	−0.211	−0.069	−9.7	−3.2
Aged2	−0.072	0.063	−2.4	2.6	−0.032	0.040	−2.4	1.2
Aged3	−0.045	0.034	−2.4	2.6	−0.024	0.007	−2.9	0.9
Sec. edu	0.035	0.185	0.6	4.1	−0.001	0.176	0.0	4.1
Tert. edu	0.085	0.266	1.1	5.4	−0.003	0.318	−0.1	6.6
Household size	−0.093	0.053	−3.4	2.0	−0.081	0.054	−3.9	2.0
Living w/spouse	−0.038	0.187	−1.0	1.9	−0.055	0.162	−1.6	3.4
IHS networth	−0.003	0.019	−0.7	4.5	0.006	0.017	1.7	5.8
Underweight	−1.221	0.169	−5.5	1.1	−0.290	0.023	−3.7	0.4
Overweight	−0.075	0.029	−1.4	0.6	−0.150	−0.043	−5.5	−1.2
Moderately obese	−0.295	−0.081	−4.0	−1.7	−0.290	−0.153	−7.4	−2.8
Severely obese	−0.563	−0.065	−4.4	−0.6	−0.567	−0.181	−8.2	−2.0
Missing edu	−0.225	0.234	−2.6	2.1	−0.440	0.277	−3.3	3.5
Missing BMI	−0.905	0.482	−2.5	4.1	−0.296	0.112	−4.1	1.1
Constant	−0.544	0.049	−4.9	0.6	−0.220	0.273	−4.3	5.4
Residual s.d.	0.274	0.611	6.3	17.1	0.235	0.375	6.3	23.3

Table 9: Other parameters.

	Male				Female			
	Estimates		*t*-values		Estimates		*t*-values	
Parameter	Min	Max	Min	Max	Min	Max	Min	Max
$\sqrt{\Omega_{GS}}$	0.687	0.945	18.7	44.4	0.545	0.659	19.0	43.2
$\alpha_{stairs,1}$	0.050	2.262	0.2	12.0	−0.252	1.437	−2.2	13.0
$\alpha_{stairs,2}$	1.405	3.648	4.5	17.7	0.921	2.943	7.3	21.4
$\alpha_{SRH,1}$	−2.891	−0.694	−21.9	−4.7	−3.296	−0.403	−26.9	−1.6
$\alpha_{SRH,2}$	−1.516	0.015	−13.4	0.2	−2.026	0.149	−17.7	0.7
$\alpha_{SRH,3}$	0.042	1.528	0.2	20.4	−0.257	1.442	−2.5	21.4
$\alpha_{SRH,4}$	1.482	2.562	6.6	23.3	1.387	2.495	8.1	28.3

Table 10: Estimated distribution of latent (true) health η and of the constructed health index $\hat{\eta}$.

Country	Male				Female			
	latent		index		latent		index	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
Austria	0.22	0.31	0.23	0.27	0.21	0.40	0.22	0.36
Belgium	0.08	0.47	0.10	0.40	0.02	0.41	0.02	0.38
Denmark	0.27	0.53	0.28	0.46	0.04	0.40	0.05	0.36
France	−0.04	0.55	−0.02	0.48	−0.02	0.40	−0.01	0.36
Germany	0.21	0.56	0.21	0.50	0.17	0.39	0.17	0.36
Greece	−0.19	0.53	−0.17	0.44	−0.13	0.31	−0.13	0.28
Israel	−0.39	0.74	−0.35	0.64	−0.31	0.48	−0.30	0.45
Italy	−0.29	0.63	−0.27	0.54	−0.25	0.42	−0.24	0.38
The Netherlands	0.15	0.46	0.17	0.38	0.05	0.43	0.05	0.39
Spain	−0.41	0.57	−0.39	0.49	−0.32	0.47	−0.31	0.43
Sweden	0.04	0.55	0.06	0.47	−0.04	0.41	−0.04	0.37
Switzerland	0.15	0.51	0.15	0.44	0.13	0.40	0.14	0.35

Note. Weighted results.

above, and computed the sample means and standard deviations of this. These results are presented in Table 10. It reflects large cross-country differences in average health, compared to the within-country variation (which is fairly similar across countries): the differences between the countries with the highest and lowest mean health exceed the within-country standard deviation. The patterns in average health are as expected: average health is worse in Southern European countries (Spain, Italy, Greece) and better in Central and Northern Europe, with more affluent countries (Germany, Switzerland, Austria, Denmark and The Netherlands for males). The average position of Sweden was not anticipated, though. The sample means of the health index track the estimated means of true health quite closely. The sample standard deviations of the health index are somewhat smaller than the standard deviations of true health. This is always the case when a conditional mean is used as the best estimate of a random variable.

Table 11 shows the precision with which individual true health is estimated. It presents the R^2 of the predictive health equation derived from the estimates, which is a measure of how well health is estimated from only the explanatory variables and the model parameters, and the reliability of the health index, which is the squared correlation between the health index and true health, derived similarly (see Appendix C). Although the covariates clearly provide some information about true health, the resulting R^2s are too low to use the resulting index functions (i.e., $\hat{\Gamma}x$) as a health index. In contrast, our proposed health index achieves a satisfactory reliability of about 0.80. We conclude that the health indexes have a satisfactory reliability for measuring (this dimension of) health, which should make them useful in subsequent modeling, but they are not entirely without measurement error.

Figures 1 and 2 plot the mean of the health index (aggregated across countries with weights

Table 11: Squared correlation (R^2, reliability) between health measure and true latent health.

Country	Male		Female	
	Covariates only	Health index	Covariates only	Health index
Austria	0.22	0.78	0.35	0.84
Belgium	0.17	0.76	0.35	0.85
Denmark	0.29	0.78	0.32	0.83
France	0.30	0.79	0.39	0.83
Germany	0.32	0.81	0.42	0.86
Greece	0.29	0.79	0.43	0.85
Israel	0.32	0.81	0.38	0.88
Italy	0.30	0.79	0.36	0.86
The Netherlands	0.19	0.73	0.32	0.83
Spain	0.22	0.81	0.40	0.88
Sweden	0.34	0.78	0.41	0.84
Switzerland	0.38	0.74	0.29	0.80

Note. Derived from parameter estimates; weighted results.

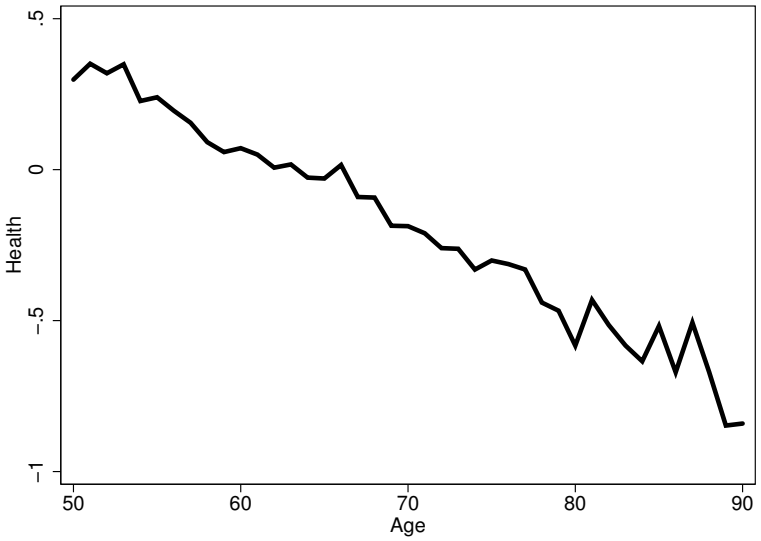

Figure 1: Mean health index by age, males (aggregated across countries; weighted).

proportional to population size) against age. From these figures, we see that health deteriorates linearly over the age range studied.

Figures 3–6 plot the average of the health index versus log household income (PPP adjusted, Euros) and the inverse hyperbolic sine of household net worth (ditto) for males and females, using weighted nonparametric regression. Generally, this shows the well-known health-SES gradient:

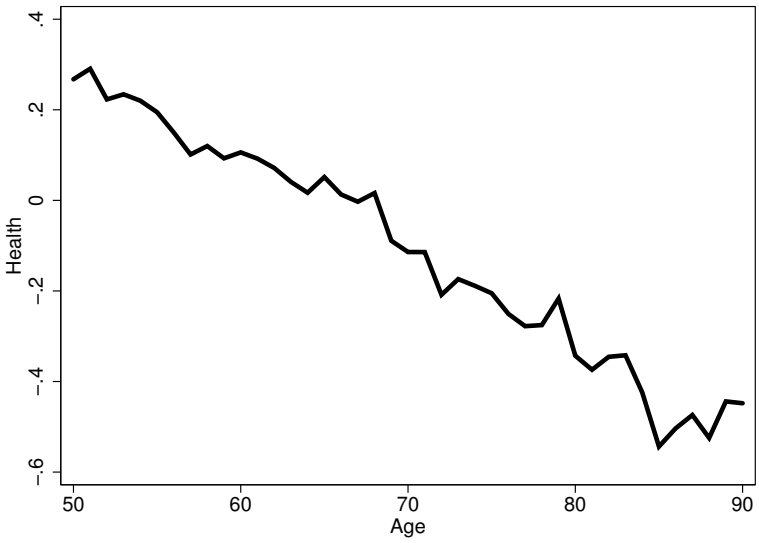

Figure 2: Mean health index by age, females (aggregated across countries; weighted).

health is better for the more affluent. However, at very low income levels, the relationship becomes more erratic. In fact, there is a small number of observations with even lower income and wealth levels, and for these, the relationships are even more erratic (not shown). Presumably this can be attributed to the small number of observations and possibly significant measurement errors at these levels.

6 A descriptive analysis of retirement and some of its determinants

In section 7, we will present some simple retirement models to assess the potential usefulness of the health index in retirement modeling. This section discusses the criteria for selecting an analysis sample and their consequences. Furthermore, the variables to be used in the retirement models are selected and constructed here, and some descriptive statistics are presented as well.

Selection of the analysis sample

Two criteria are used to select our analysis sample. The first is that the observation must have been selected for the health model as well, which means that age and gender must not be missing and age must be at least 50. The second criterion is that the respondent must have been working at age 50 or over, i.e., he or she was in the labor force at age 50 (or later). The latter includes respondents who currently have labor force status "(Self-)employed" and respondents with other current labor force statuses as long as they were working in the year they turned 50 or later.

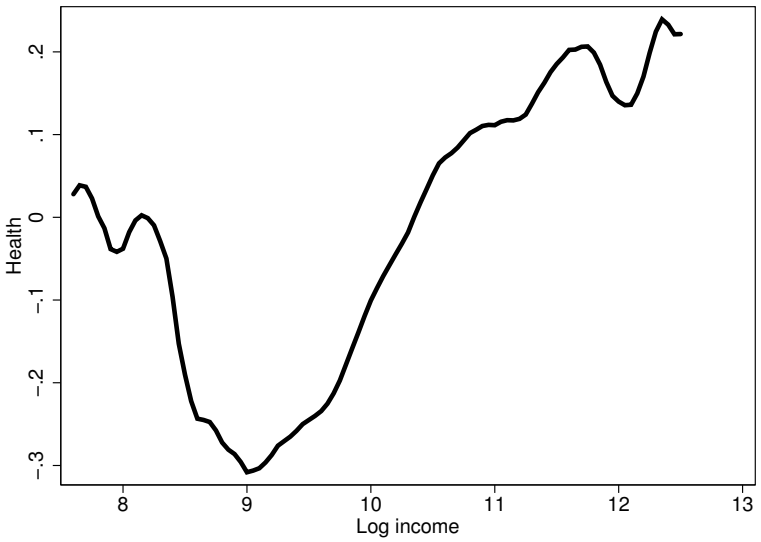

Figure 3: Mean health index by log household income, males (aggregated across countries,weighted).

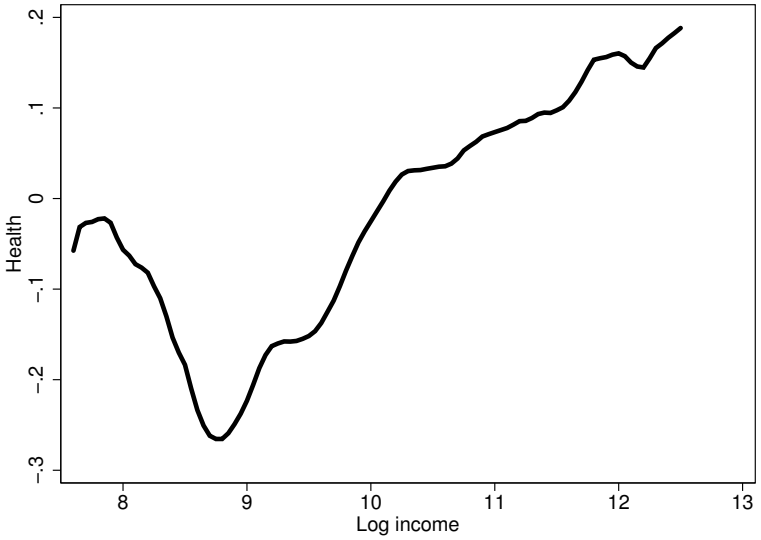

Figure 4: Mean health index by log household income, females (aggregated across countries, weighted).

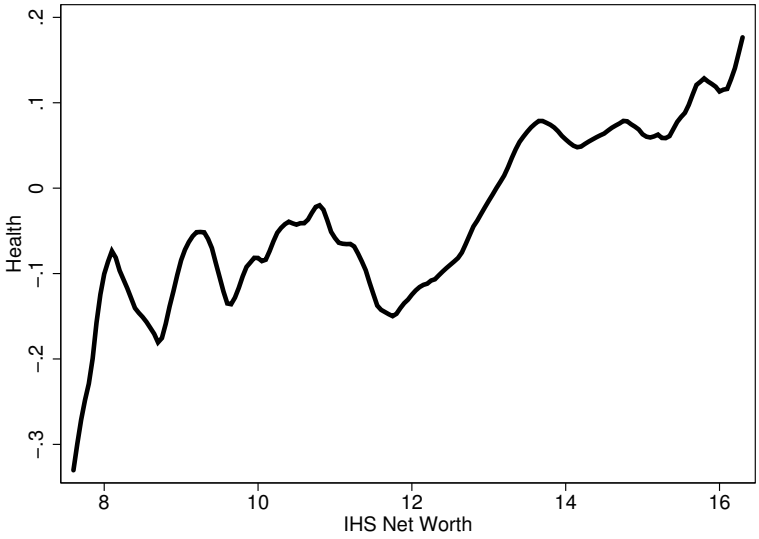

Figure 5: Mean health index by inverse hyperbolic sine of household net worth, males (aggregated across countries,weighted).

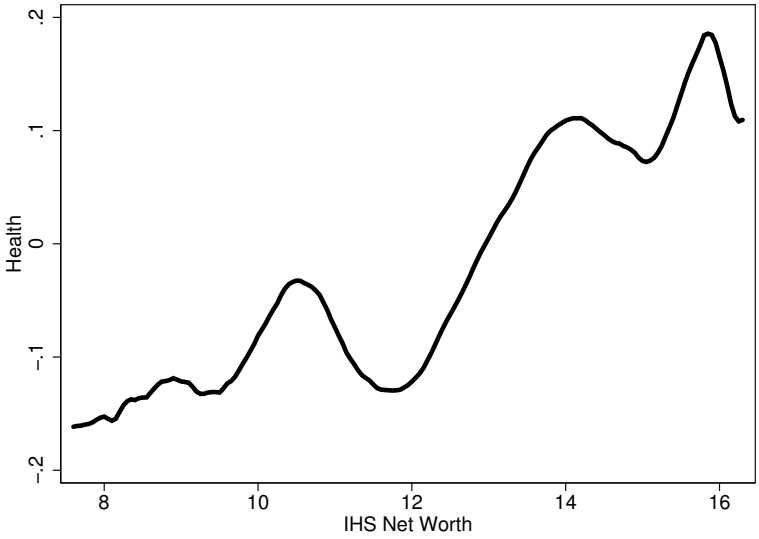

Figure 6: Mean health index by inverse hyperbolic sine of household net worth, females (aggregated across countries,weighted).

Table 12: Sample composition by country and gender.

Country	Male	Female	Total
Austria	727	715	1,442
Belgium	1,538	960	2,498
Denmark	695	710	1,405
France	1,236	1,111	2,347
Germany	1,278	1,093	2,371
Greece	1,145	618	1,763
Israel	987	878	1,865
Italy	1,009	597	1,606
The Netherlands	1,209	743	1,952
Spain	870	488	1,358
Sweden	1,340	1,410	2,750
Switzerland	419	347	766
Total	12,453	9,670	22,123

Note. Unweighted number of respondents.

The reason for the second criterion is that a model relating current health to current retirement status makes no sense for people who have never worked or who have stopped working at a very young age, e.g., 25. Ideally, we would use health condition at the age of retirement, but evidently this is not available. Furthermore, decisions to never enter the labor force or to exit at a very early age are likely to be very different from health-related retirement decisions later in life. Because 50 is also the age at which one becomes eligible for inclusion in the health model (and for inclusion in the sample in the first place), this seems to be a reasonable choice of the cutoff age. However, this criterion leads to a considerable reduction in the sample size, and the respondents selected for the analysis sample differ substantially from those who are not. Therefore, below we present several tables comparing the analysis sample with the whole sample including those who are not selected for the retirement model. The resulting sample size of the analysis sample is 22,123. Table 12 gives a breakdown by country and gender.

Table 13 shows the percentage of respondents included in the health model (i.e., aged 50 and over) who were already out of the labor force at age 50 and thus were excluded from the retirement model. (These results and the ones presented further below are all weighted.) It follows that about 6% of males was already out of the labor force at age 50, which is noticeable but not extremely high. For females, the percentages are much higher, and the differences across countries are also more significant. These large percentages are mainly due to female homemakers who never worked or exited the labor force at a very early age. Female labor force participation is generally higher for later cohorts, so that this percentage will be reduced over time.

Table 13: Percentage out of the labor force at age 50.

Country	Male	Female
Austria	5.6	30.5
Belgium	8.1	47.1
Denmark	6.6	13.1
France	4.1	29.6
Germany	5.4	27.4
Greece	5.2	46.9
Israel	3.9	11.7
Italy	8.6	50.5
The Netherlands	8.5	46.4
Spain	7.0	58.7
Sweden	3.9	8.8
Switzerland	3.4	24.5
Total	6.1	37.6

Note. Weighted results.

Current labor force status

Table 14 presents the distribution of current labor force status by country and gender for the analysis sample, i.e., for those who work(ed) after age 50. For males, the corresponding table for all respondents is very similar, but for females, the table for all respondents has much higher percentages of homemakers and correspondingly lower percentages for retired and employed. Table 14 shows some substantial differences across countries. The most striking phenomenon is that the percentage of females who are homemakers is much higher in The Netherlands and Spain than in the other countries. Apparently, becoming a homemaker is an important exit route out of the labor force in these countries. For our analysis, we consider this group as retired. Similarly, we consider persons who worked after 50 years of age, but are now permanently sick or disabled as being retired. Unemployment could be an exit route out of the labor force. Therefore, we consider retirement to be unknown (missing) for this group. An exception to the preceding rules is when respondents report being temporarily away from work. In that case, they are considered not retired, whatever their current self-reported labor force status.

With this definition of retirement, Table 15 gives the percentage of retired in the analysis sample (i.e., those who worked after 50) by country. Additionally, Figure 7 shows the percent retired as a function of age (in 5-year categories to smooth the figures somewhat), for males and females separately. There are sizeable differences across countries. When looking at the underlying country-level data (not depicted), we see that, for example, individuals in Switzerland tend to retire relatively late, whereas the French on average retire early (both conditional on being in the labor force at age 50). Understanding cross-country and cross-gender differences will be an important goal in the study of retirement.

Table 14: Current labor force status (analysis sample).

Country	Retired	(Self-) Employed	Unemployed	Perm. sick/ Disabled	Homemaker
Males					
Austria	64.5	31.3	2.4	1.5	0.3
Belgium	62.8	31.8	2.6	2.7	0.1
Denmark	47.8	46.8	4.4	0.9	0.2
France	61.2	33.9	3.6	0.9	0.4
Germany	55.3	37.8	4.8	2.1	0.0
Greece	58.0	40.2	1.3	0.4	0.0
Israel	37.5	55.0	3.2	2.3	1.0
Italy	66.7	30.9	2.0	0.4	0.0
The Netherlands	48.2	45.0	2.2	3.5	0.5
Spain	58.7	36.2	3.9	1.1	0.0
Sweden	49.7	45.9	2.5	1.5	0.0
Switzerland	42.8	54.3	1.6	1.3	0.0
Total	58.3	36.7	3.4	1.4	0.1
Females					
Austria	72.6	22.9	2.4	0.8	1.4
Belgium	56.2	33.4	4.5	2.1	3.9
Denmark	54.0	40.0	3.7	1.8	0.6
France	58.3	35.4	3.7	0.5	2.0
Germany	59.2	33.0	3.1	0.8	3.8
Greece	62.0	31.7	1.9	0.3	4.1
Israel	48.8	40.0	3.4	3.0	3.9
Italy	64.7	28.8	1.1	0.1	5.3
The Netherlands	28.4	44.8	2.4	6.1	17.7
Spain	31.6	40.5	7.5	3.8	15.9
Sweden	56.0	39.8	2.0	1.6	0.4
Switzerland	46.5	45.5	1.8	0.9	5.2
Total	56.0	34.5	3.2	1.2	5.0

Note. Weighted results.

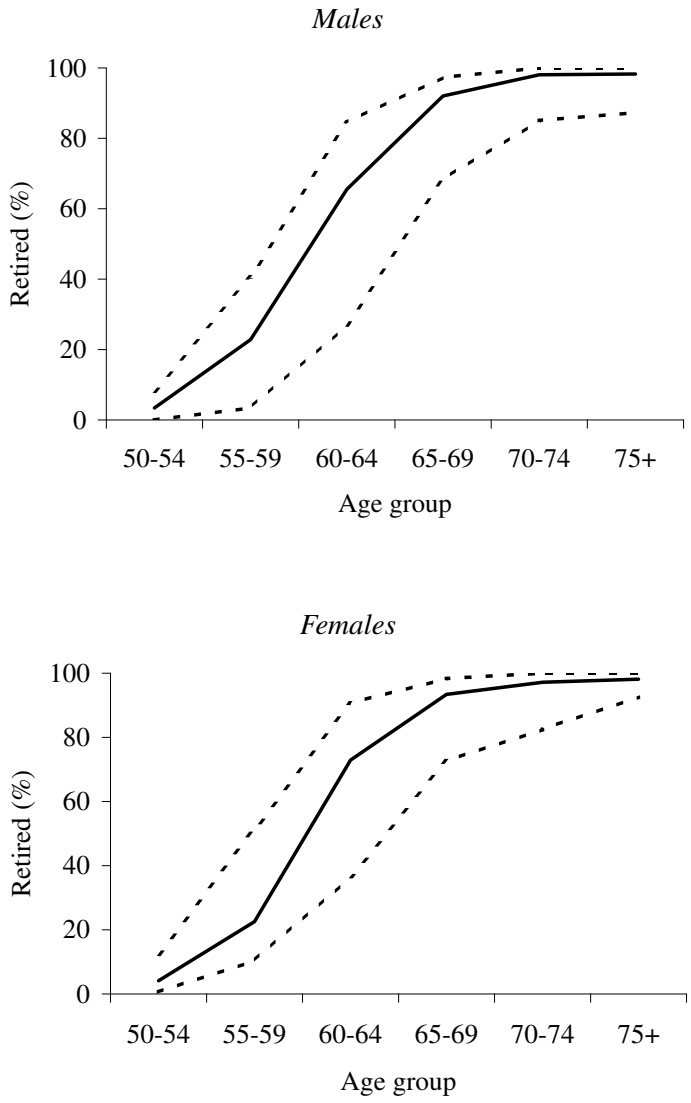

Figure 7: Percent retired by age and gender (analysis sample, weighted results). Solid line: Weighted average of all 12 countries. Dotted lines: minimum and maximum in each age group.

Table 15: Percentage who have exited the labor force (analysis sample).

Country	Male	Female
Austria	65.4	72.0
Belgium	66.9	64.6
Denmark	50.7	58.2
France	64.8	63.1
Germany	60.1	65.9
Greece	56.5	63.8
Israel	42.4	57.0
Italy	67.4	69.8
The Netherlands	53.6	53.6
Spain	60.7	53.6
Sweden	52.6	59.3
Switzerland	40.0	47.8
Total	61.2	63.5

Note. Weighted results.

Stated reasons for exiting the labor force

Table 16 shows the reasons that were given for exiting the labor force, by retirees and homemakers, for both the whole sample and the analysis sample. Eligibility for a public or private pension of some sort is the dominant reason to retire. Health is also important, mentioned by about 15% of the respondents. It is sometimes observed that ill health of the partner is an important reason to retire for women, but this does not appear to be the case here, unless a large proportion of the "other reasons" for homemakers falls into this category. The same holds for retiring at the same time as the partner, which is mentioned by less than 2% of the retired.

The differences between the whole sample and those who were still in the labor force at age 50 are not very large for those who are currently retired. For (current) homemakers, however, health is a much more important reason to stop working after 50, whereas taking care of children is much more important before 50 (compare the columns "All respondents" and "Analysis sample").

Combining all exit routes, 19.9% of all those who stopped working report a health reason, whereas for the analysis sample, this is 16.7%. In computing these numbers, we have included the "permanently sick or disabled" among the health reasons, and for the homemakers we have included "too tiring" among the health reasons.

Table 17 breaks down this number by country and gender for the analysis sample (the whole sample shows similar patterns). There are some clear differences, both across countries and between men and women. In Spain, Sweden, and The Netherlands, a relatively large number of females stopped working because of a health reason. Spain and The Netherlands are also the countries with many females exiting the labor force to become homemakers. It remains to be seen whether this indicates that health is worse for females from these countries, whether institutions (tax systems, social security, pension systems) are such that it is easier to leave the labor force,

Table 16: Reasons for exiting the labor force.

	All respondents	Analysis sample
Reported labor force status = Retired		
Eligible for public pension	48.4	51.2
Eligible for private occupational pension	6.8	7.6
Eligible for a private pension	5.9	6.8
Was offered early retirement	11.4	13.0
Made redundant	5.3	5.7
Own ill health	15.6	14.2
Ill health of relative or friend	2.5	2.4
Retire at same time as spouse or partner	1.6	1.8
Spend more time with family	6.4	3.6
Enjoy life	5.1	5.4
Other	4.6	3.4
Reported labor force status = Homemaker		
Health problems	12.2	23.0
Too tiring	6.0	14.0
Too expensive to hire someone to look after home or family	3.8	1.6
Wanted to take care of (grand)children	41.2	14.5
Other reason	42.5	53.4

Note. Weighted results.

Table 17: Percentage who stopped working because of a health reason, all exit routes combined (analysis sample).

Country	Male	Female
Austria	22.8	15.4
Belgium	15.9	17.7
Denmark	25.6	23.0
France	13.2	11.8
Germany	25.8	16.0
Greece	7.8	10.7
Israel	26.5	17.4
Italy	7.8	11.4
The Netherlands	17.0	28.2
Spain	17.2	37.8
Sweden	22.7	29.6
Switzerland	10.6	10.9
Total	16.8	16.7

Note. Weighted results.

and/or whether this reflects the often mentioned justification bias.

Characteristics of current or last job and eligibility for pensions

Table 18 shows the percentage of people who are (were) self-employed or work(ed) in the public sector in their current or last job. Furthermore, it shows the percentage of people receiving public or private pensions or who expect to be eligible for this in the future. Again, there are large differences across countries and, to a lesser extent, by gender. The large differences in eligibility for private pensions are likely to have an impact on retirement, and in addition to that on poverty in old age.

The retirement decision is, of course, strongly related to the amount of pension wealth that has been accrued. There is ample information about this in the SHARE data, but properly modeling the effect of wealth on retirement is beyond the scope of this paper. As a very crude proxy for pension accrual, we can use the number of years worked in the last job (for retired) or in the current job (for currently employed). However, this variable suffers from the problem that it depends on the retirement decision itself: if someone decides to retire, no more years are added, whereas if someone keeps working, additional years are added. Hence, including this as an explanatory variable in a model for retirement would likely lead to a coefficient with the opposite sign (more work years are related to nonretirement), which is correct, but not the effect we are aiming to estimate. Therefore, we have instead constructed a variable that contains the number of years worked (in the current or last job) at age 50. Given that all respondents in the analysis sample are 50 or over, this does not suffer from the problem mentioned. However, it has other imperfections.

Table 18: Percentage self-employed and working in public sector in current or last job and percentage eligible for public and private pensions, now or in the future (analysis sample).

	Current or last job		(Current or future) Eligibility for pension	
Country	Self-employed	Public sector	Public	Private
Males				
Austria	12.5	30.9	91.5	8.9
Belgium	15.0	31.7	91.2	9.8
Denmark	15.7	31.0	96.8	57.6
France	19.0	26.5	94.0	68.0
Germany	12.2	24.2	93.3	26.8
Greece	40.9	24.0	77.8	1.5
Israel	25.5	29.6	80.8	37.9
Italy	29.8	25.1	81.6	15.1
The Netherlands	12.7	44.2	97.5	75.6
Spain	27.5	20.0	85.3	4.3
Sweden	17.0	27.6	93.7	30.4
Switzerland	24.6	25.0	96.2	66.4
Total	20.6	25.9	89.7	31.4
Females				
Austria	14.0	25.9	91.8	6.7
Belgium	19.7	36.9	90.0	6.2
Denmark	11.0	55.9	97.1	50.9
France	15.4	32.7	94.3	63.2
Germany	7.4	29.5	94.0	19.7
Greece	39.2	18.9	78.9	0.7
Israel	12.0	42.5	85.8	33.2
Italy	28.2	29.4	84.3	10.3
The Netherlands	14.9	48.9	98.4	63.7
Spain	26.4	24.1	77.1	3.3
Sweden	6.8	59.1	94.5	42.3
Switzerland	21.5	29.9	97.9	47.5
Total	15.9	32.4	90.8	29.1

Note. Weighted results.

For example, it may not capture pension accrual at the retirement age of, say, 62 very well, and some respondents changed jobs after age 50. The latter have been assigned the number zero on this variable. For this variable, starting (and ending) ages below age 15 have been recoded as missing. The resulting variable ranges from 0 to 35, with a mean of 15.9 and a standard deviation of 11.1 for those who are not retired and a mean of 17.0 and a standard deviation of 11.5 for those who are. At first sight, these differences do not seem very large, but whether this variable is able to explain a significant part of retirement is investigated further in the next section.

7 A simple retirement model

There are several types of retirement models. Some retirement models are structural models, i.e., they embed the retirement decision in an economic model that is intended to capture the causal mechanisms. Many of these are dynamic programming models, in which a lifetime utility function based on preferences for consumption, leisure, and sometimes other variables, is assumed to be maximized by economic agents. The decision whether or not to retire influences the budget available for consumption and the amount of time available for leisure, and the choice whether or not to retire (at each age) thus influences current and future utility through the budget and time available. Examples of dynamic programming models for retirement are Gustman and Steinmeier (1986a, 1986b, 2000), Berkovec and Stern (1991), Rust and Phelan (1997), and French (2005). The option value model of Stock and Wise (1990) is a variation on this, but it simplifies the decision process in the model so that the model becomes more tractable.

Dynamic programming models are often difficult to estimate, for a number of reasons. One reason is the computational complexity of optimizing an intertemporal utility function over a large number of periods, with many choice options, and stochastic shocks to many of the variables of interest. Unrealistic simplifying assumptions often have to be made in order to arrive at a solution. For example, uncertainty about the future is sometimes eliminated, health is absent or reduced to a single binary variable ("good" or "bad"), and many decisions and other variables are coarsely discretized.

Another difficulty with dynamic programming models (and structural models in general) is that, in order to model the decisions satisfactorily, detailed information about wages, pension accrual, and other assets, for several time periods must be available, which is often not the case.

Therefore, models with a simpler structure have also been estimated. They take the form of relatively simple discrete choice models or hazard models that impose less stringent requirements on the data availability and are computationally simple. On the other hand, they may not accurately account for important aspects of forward-looking behavior and updating expectations about the future over time. Examples of such models are Mitchell and Fields (1984) and Michaud (2005, chap. 2).

A structural dynamic programming model is outside the scope of this paper. Moreover, given that we have data from only the first wave of SHARE, our model will not be able to capture

dynamic decision making and the variables used in the model are necessarily limited to the information available in this single wave. Hence, our model serves as an illustration and a tentative assessment of the usefulness of the health index in retirement modeling.

The dependent variable in our model is whether or not someone is currently retired, according to the definition given in section 6. Thus, we model the current state and not the decision to retire directly. As mentioned in section 6, we restrict our analysis to respondents who were still in the labor force at age 50.

The explanatory variables that we consider are primarily the ones described in section 6. Interactions between current or future eligibility for public or private pensions and dummies for reaching early or normal retirement age (from Coe & Zamarro, 2008, Table 1, with data for Israel from OECD, 2005, pp. 29–30) are added, each interacted with being self-employed or in the public sector as well. Furthermore, a third-degree polynomial in age is included, as well as some of the socio-demographic variables used in the health model: education, household size, and living with a spouse/partner.

To assess the usefulness of the health index, we estimate models without any health measure, models with the health index, and models with some alternative health variables. The latter include grip strength, self-reported health (used as a continuous variable or a set of dummies), the numbers of mobility limitations, ADLs, and IADLs. In addition, we estimate models that use entirely different health variables, namely the number of doctor-diagnosed chronic conditions and the number of symptoms (i.e., directly noticeable physical health problems). We also estimate models with multiple health measures.

For education and the health variables, we use the approach described earlier to deal with missing data: we added dummies for missingness and assigned arbitrary values to these variables (zero, except for self-reported health, where the middle category 3 = "good" was used) if they were missing. We have made no such attempts for the other explanatory variables, which means that cases that had corresponding missings were automatically removed by `Stata`.

To identify variables with the highest explanatory power, we ran forward stepwise probits with all variables eligible for inclusion in principle. Variables that were logically connected were declared as such, except for the powers of age. Table 19 indicates which health measures were selected by this procedure. With the exception of Greek Males, all models include at least one health variable. Unfortunately, there is not a clear "best" health measure, and even the number of health measures varies considerably (from 0 to 4).

The stepwise procedure selects a different set of explanatory variables for each country-gender combination, including the non-health variables. Most likely, correlations between the health variables and the non-health variables imply that the set of non-health variables selected also influences which health variables are selected. This makes it difficult to compare the results for different country-gender combinations.

Therefore, we have also estimated models with a fixed set of explanatory variables that we considered most interesting, often including ones that were removed from the stepwise procedure and vice versa. The variables that we included are all three powers of age, living

Table 19: Health variables selected for the retirement model by the stepwise probit procedure. (forward, with p-value for entering = 0.05 and p-value for removal = 0.10).

Country	Grip strength resid.	SRH (dummies)	Mob (#)	ADL, IADL (#)	Chronic (#)	Symptoms (#)	$\hat{\eta}$
Males							
Austria	×						
Belgium	×	×					
Denmark		×		×			
France							×
Germany			×				×
Greece							
Israel			×	×			×
Italy						×	
The Netherlands		×		×			
Spain			×		×		
Sweden	×	×		×	×		
Switzerland				×			
Females							
Austria		×					
Belgium		×					×
Denmark	×	×		×			
France	×			×			
Germany						×	×
Greece	×						
Israel		×	×				
Italy					×		
The Netherlands				×	×		×
Spain	×	×			×		
Sweden		×					×
Switzerland				×	×		

Note. Grip strength resid. = the residual grip strength as used in the health measurement model and as constructed in the appendix; SRH = self-reported health; Mob, ADL, IADL = number of mobility, ADL, and IADL limitations, respectively; Chronic = number of chronic conditions reported; Symptoms = number of health symptoms reported; $\hat{\eta}$ = the healthindex constructed above.

with a spouse/partner, household size, education dummies, work years at age 50, self-employed, employment in the public sector, reached normal retirement age, reached early retirement age × entitled to a private pension, reached normal retirement age × entitled to a private pension, reached normal retirement age × self-employed, and reached normal retirement age × entitled to a private pension × self-employed. These were present in all subsequent models, unless they led to perfect collinearity among the regressors, in which case some of the regressors were automatically removed by Stata. In addition to this list, we varied the health variables included in the model.

Just like the stepwise results in Table 19, the country-specific results of this model show different "preferred" models for different countries, with no discernable overriding pattern. Tables 20 and 21 give the fit statistics for each of these models after aggregating over countries. That is, all coefficients in the model are still country-specific, but sample sizes, loglikelihoods, and degrees of freedom are summed across countries, and then Akaike's Information Criterion (AIC) and Schwarz' Bayesian Information criterion (BIC) were computed for the combined model. Hence, this indicates how well each model does "on average".

From these tables, we can observe the following: First, models with some kind of health measure tend to fit (much) better than the model without health measures when looking at log (pseudo)likelihood values and AIC. The higher penalty for additional parameters in the BIC formula, compared to AIC, implies that many models with health measures have a higher BIC and thus worse fit if BIC is used as a fit criterion. However, whichever fit statistic is preferred, the best fitting model is one with at least one health measure. According to the BIC criterion, the model with the health index $\hat{\eta}$ as only health measure fits best, both for males and for females. When looking at the log pseudolikelihood or AIC, the model with the largest number of variables (and thus the largest number of parameters) fits best, which is a model without $\hat{\eta}$, but with all components that are used in its construction included separately, with their own parameters (Model 18 in the tables). When restricting attention to models with only a single health component (Models 2–9), the model with $\hat{\eta}$ as health measure fits best according to the AIC as well, and in terms of log pseudolikelihood, it is only surpassed by the model with self-reported health included as dummies, at the expense of almost 40 additional parameters.

Hence, the health index $\hat{\eta}$ appears to capture the most important dimension of health in its relation to retirement. However, we interpret these findings also as indications that more than one dimension of health may be relevant for retirement modeling. We believe that the best way to do this is to include a second (and perhaps third) latent health dimension to the measurement model and construct a corresponding multidimensional health index. We leave this for further research.

The models that include the number of chronic conditions and/or the number of health symptoms in the model, in addition to the health index, do not seem to fit noticeably better than the model with only the health index. We find this somewhat surprising, because these additional variables have not been used in the construction of the health index, and we had expected these to add a dimension not captured by the health index. Especially for the chronic conditions, it seems obvious that these must have an impact on health and retirement. But apparently, the effects of these conditions on health are largely captured by the health indicators that are used in our

Table 20: Fit statistics for selected models with different health variables, males.

Model	Health vars	N	LL0	LL	df	AIC	BIC
1	No health var	11,391	−7637.8	−2682.5	188	5740.9	7120.9
2	$\hat{\eta}$	11,391	−7637.8	−2564.4	200	5528.9	6997.0
3	GS	11,391	−7637.8	−2646.4	212	5716.8	7273.0
4	SRH-lin	11,389	−7636.6	−2580.8	202	5565.7	7048.5
5	SRH-dum	11,389	−7636.6	−2532.7	238	5541.3	7288.3
6	Mob	11,390	−7637.1	−2592.0	201	5586.0	7061.4
7	ADL, IADL	11,389	−7636.6	−2582.2	212	5588.4	7144.5
8	Chron	11,388	−7635.8	−2645.0	202	5694.0	7176.7
9	Sympt	11,389	−7635.4	−2621.8	200	5643.7	7111.8
10	SRH-lin, $\hat{\eta}$	11,389	−7636.6	−2538.5	214	5505.1	7075.9
11	Chron, $\hat{\eta}$	11,388	−7635.8	−2557.8	214	5543.6	7114.4
12	Sympt, $\hat{\eta}$	11,389	−7635.4	−2551.2	212	5526.4	7082.5
13	Chron, Sympt, $\hat{\eta}$	11,382	−7631.4	−2544.6	224	5537.3	7181.4
14	GS, SRH-lin	11,389	−7636.6	−2554.6	226	5561.1	7220.1
15	SRH-lin, Mob	11,388	−7635.9	−2536.2	215	5502.5	7080.6
16	SRH-lin, ADL, IADL	11,387	−7635.4	−2512.8	226	5477.6	7136.5
17	Mob, ADL, IADL	11,389	−7636.6	−2544.1	225	5538.3	7189.9
18	GS, SRH-lin, Mob, ADL, IADL	11,387	−7635.4	−2467.1	263	5460.3	7390.8
19	Chron, Sympt	11,382	−7631.4	−2607.0	212	5637.9	7194.0

Notes. N = sample size; LL0 = log pseudolikelihood of null model (constant-only model per country, or equivalently, only country dummies); LL = log pseudolikelihood of the target model; df = degrees of freedom; AIC = Akaike's Information Criterion; BIC = Schwarz' Bayesian Information Criterion.

GS = the residual grip strength as used in the health measurement model and as constructed in the appendix; SRH = self-reported health (lin = as a linear continuous variable, dum = as 4 dummy variables); Mob, ADL, IADL = number of mobility, ADL, and IADL limitations, respectively; Chron = number of chronic conditions reported; Sympt = number of health symptoms reported; $\hat{\eta}$ = the healthindex constructed above.

Table 21: Fit statistics for selected models with different health variables, females.

Model	Health vars	N	LL0	LL	df	AIC	BIC
1	No health var	8,920	−5910.8	−2127.4	180	4614.9	5892.2
2	$\hat{\eta}$	8,920	−5910.8	−2020.3	192	4424.5	5787.0
3	GS	8,920	−5910.8	−2086.9	204	4581.7	6029.3
4	SRH-lin	8,920	−5910.8	−2026.5	192	4436.9	5799.4
5	SRH-dum	8,920	−5910.8	−1989.9	228	4435.7	6053.6
6	Mob	8,919	−5910.2	−2024.9	192	4433.9	5796.3
7	ADL, IADL	8,919	−5909.7	−2034.5	204	4477.1	5924.7
8	Chron	8,919	−5910.3	−2074.4	193	4534.8	5904.3
9	Sympt	8,918	−5909.9	−2081.9	192	4547.8	5910.2
10	SRH-lin, $\hat{\eta}$	8,920	−5910.8	−1996.6	204	4401.3	5848.9
11	Chron, $\hat{\eta}$	8,919	−5910.3	−2007.8	205	4425.6	5880.3
12	Sympt, $\hat{\eta}$	8,918	−5909.9	−2009.2	204	4426.3	5873.9
13	Chron, Sympt, $\hat{\eta}$	8,918	−5909.9	−1996.8	217	4427.5	5967.3
14	GS, SRH-lin	8,920	−5910.8	−1998.4	216	4428.9	5961.6
15	SRH-lin, Mob	8,919	−5910.2	−1986.7	204	4381.4	5829.0
16	SRH-lin, ADL, IADL	8,919	−5909.7	−1974.9	216	4381.7	5914.5
17	Mob, ADL, IADL	8,918	−5909.1	−1991.8	216	4415.5	5948.2
18	GS, SRH-lin, Mob, ADL, IADL	8,918	−5909.1	−1927.7	252	4359.3	6147.5
19	Chron, Sympt	8,918	−5909.9	−2055.4	205	4520.7	5975.3

Note. See the notes for Table 20 for an explanation of the abbreviations.

measurement model, and thus after accounting for these in the form of our health index, the chronic conditions do not add much predictive power. However, in some country-gender combinations, chronic conditions have additional explanatory power, and in Table 19, they are included in the preferred model several times as well.

To get an impression of the size of the health effect, we have computed the actual and predicted retirement percentages for the analysis sample, and computed the predicted retirement percentages for the analysis sample after an improvement of health by one (gender and country-specific) standard deviation for all observations. The model used for these calculations is the one with $\hat{\eta}$ as the only health variable in the model. The predicted retirement percentages are computed as the means of the predicted retirement probabilities. We have computed the retirement percentages for each of the 5-year age groups 50–54, 55–59, 60–64, 65–69, 70–74, and 75 and over.

The predicted retirement percentages show a somewhat smoother relation with age group than the observed ones. For example, in the observed data, retirement percentages do not always increase monotonically with age group, whereas in the predicted percentages they do. So there are some small differences between observed and predicted percentages. In order not to confound the effect of a health improvement with the effect between observed and predicted for the same health, we confine ourselves to comparisons between predicted percentages with and without health improvement. Not surprisingly, these effects are typically largest in the age group 60–64, in which retirement percentages increase rapidly. Figure 8 shows the results for this age group.

The effects vary widely across countries. In some cases, health improvement has almost no effect, whereas in others, it leads to an almost 20 percentage points lower predicted retirement rate. The effects are largest in Germany, Spain, and Denmark for males, and Germany, The Netherlands, and Sweden for females. Although strong causal interpretations of these simple reduced form models are unwarranted, it appears to indicate that the effects of health on retirement are noticeable and relevant.

8 Discussion

In this paper we have estimated a measurement model for health, with health as a latent (unobserved) variable. The indicators that are assumed to depend on health are mobility, arm function, and fine motor function limitations, limitations of activities of daily living (ADL), limitations of instrumental activities of daily living (IADL), self-reported health, and grip strength. The latter is also allowed to depend on a second-degree polynomial in height and weight directly. In addition to this submodel, the health model contains a predictive health equation, in which latent health is regressed on a standard set of socio-demographic covariates: a third-degree polynomial in age, living with a spouse/partner, household size, education dummies, and (the inverse hyperbolic sine transform of) household wealth, as well as dummies for being underweight, overweight, moderately obese, or severely obese. The model is a special case of the LISCOMP model, with a linear structure in the latent domain and threshold relations between latent response variables

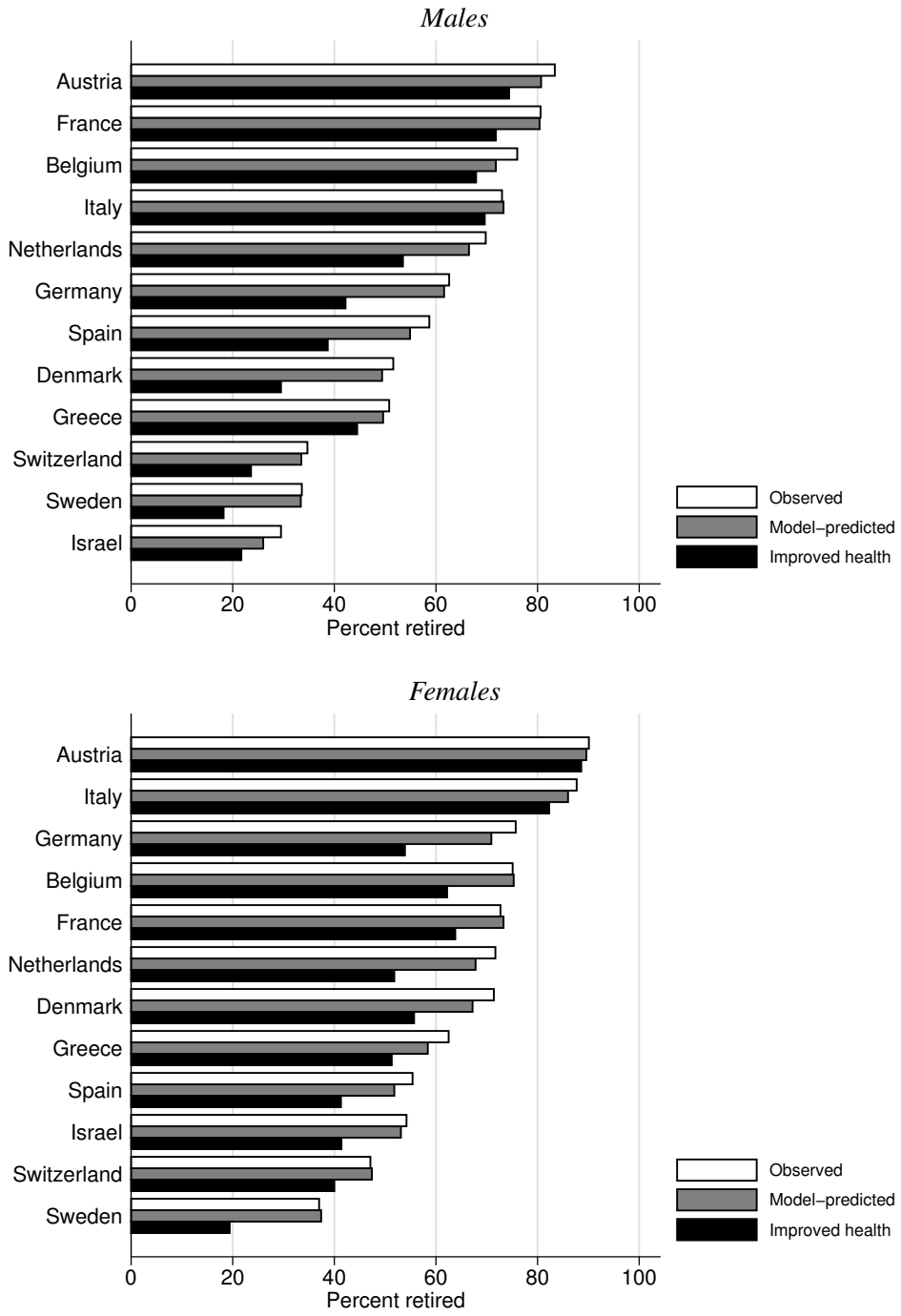

Figure 8: Effect of a 1 s.d. health improvement on percent retired for 60–64 year olds (analysis sample, weighted results).

and categorical indicators. All analyses have been done with the SHARE wave 1 data (Survey of Health, Ageing and Retirement in Europe).

Using the estimation results of the model, we computed a health index for each observation. This health index is the best possible estimate of the latent health variable in the model. Overall, the model fit seems satisfactory, although more experience with the pseudo-R^2 in this type of model and comparison with competing models are necessary to be able to make firmer statements about model fit. The reliabilities of the constructed health indexed are approximately 80%, which is satisfactory.

In order to assure cross-country comparability of health, the coefficient of latent health in the grip strength equation was normalized to 1, and the intercept in this equation was normalized to 0. Using this objective indicator for normalization led to a health variable that is comparable across countries.

Our primary objective for estimating the health model and the health index is to use the health index in retirement modeling. We have presented some descriptive statistics on retirement and its determinants in our data, from which we can already conclude that health plays an important role in retirement decisions. To assess the usefulness of our health index in retirement modeling, we have estimated some simple probit models for a subsample of individuals who worked until at least 50 years of age. "Being retired" at the time of the interview was the dependent variable, and explanatory variables were mainly socio-demographic variables, in addition to some limited economic incentives proxies, and health. We varied the health variable(s) included in these models. It appears that the health index is an adequate predictor of retirement. When judged by Schwarz' Bayesian Information Criterion (BIC), the model with the health index as only health variable is the best fitting model among the wide range of models considered. However, judged by Akaike's Information Criterion (or log pseudolikelihood by itself), the model with five health components, which are used in constructing the health index, separately included instead of the health index, is the best fitting model overall. We interpret this as an indication that extending our health model to include a second health dimension may lead to further improvements. It must be noted, however, that when looking at different country-gender subsamples separately, there is considerable variation as to which model fits best.

For understanding different retirement patterns in different countries, and the role of that different institutions in different countries pley in this, retirement models with different health variables for different countries are less useful. Moreover, even models with the same health variables, other than a cross-country comparable health index like the one constructed here, are more difficult to compare, because of the cultural and linguistic differences in the response patterns to most of the variables.

Using the retirement model with the health index as only health variable, we have assessed the strength of the relation between health and retirement by hypothetically improving each individual's health by one (country-gender specific) standard deviation, and comparing the resulting predicted percent retired with the predicted percent retired using actual health. This suggests that in some countries, improved health may lead to considerably lower retirement

fractions (up to 20 percentage points), especially in the 60–64 age group. However, because our model is a very simple reduced form model, causal interpretations are unwarranted and result like these should be viewed as tentative.

Further research is needed to improve the measurement of health and its use in retirement modeling. As mentioned above, the health model with more than one health dimension is expected to better reflect the multidimensional nature of health and to explain retirement. Furthermore, sensitivity analyses using different specifications of the predictive health equation are necessary.

An improvement of the retirement modeling is to take the estimated uncertainty in the health index into account. This requires simulated likelihood estimation, but it is a viable option. Further improvements and/or alternative approaches can be obtained by using health in structural dynamic programming models of health. Our data are not sufficient to estimate these models now, but with forthcoming subsequent waves of SHARE, this will become possible.

Acknowledgements

We would like to thank Meena Fernandes for helping with data preparation, and Susann Rohwedder, Pierre-Carl Michaud, Jeff Dominitz, Giovanni Mastrobuoni, Gema Zamarro, and seminar participants at McGill University, RAND Corporation, and the Workshop on the Economics of Ageing (Torino) for stimulating discussions and constructive comments from which we have greatly benefited.

We thank the US National Institute on Aging for funding under research grants P01 AG022481-01 and R01 AG030824-01. Additional funding was provided by the US Department of Labor, contract number J-9-P-2-0033.

This paper uses data from Release 2 of SHARE 2004. 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). Additional funding came from the US National Institute on Ageing (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01 and OGHA 04-064). Data collection in Austria (through the Austrian Science Foundation, FWF), Belgium (through the Belgian Science Policy Office) and Switzerland (through BBW/OFES/UFES) was nationally funded. The SHARE data collection in Israel was funded by the US National Institute on Aging (R21 AG025169), by the German-Israeli Foundation for Scientific Research and Development (G.I.F.), and by the National Insurance Institute of Israel. Further support by the European Commission through the 6th framework program (projects SHARE-I3, RII-CT-2006- 062193, and COMPARE, 028857) is gratefully acknowledged. The SHARE data set is introduced in Börsch-Supan et al. (2005); methodological details are contained in Börsch-Supan and Jürges (2005).

References

Al-Snih, S., Markides, K., Ray, L., Ostir, G., & Goodwin, J. (2002). Handgrip strength and mortality in older Mexican Americans. *Journal of the American Geriatric Society*, *50*, 1250–1256.

Anderson, K. H., & Burkhauser, R. V. (1985). The retirement-health nexus: A new measure of an old puzzle. *Journal of Human Resources*, *20*, 315–330.

Asparouhov, T., & Muthén, B. O. (2005). Multivariate statistical modeling with survey data. In *Proceedings of the Federal Committee on Statistical Methodology (FCSM) research conference.* Available from `http://www.fcsm.gov/05papers/Asparouhov_Muthen_IIA.pdf`

Berkovec, J., & Stern, S. (1991). Job exit behavior of older men. *Econometrica*, *59*, 189–210.

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

Börsch-Supan, A., & Jürges, H. (Eds.). (2005). *The Survey of Health, Aging, and Retirement in Europe — methodology*. Mannheim, Germany: Mannheim Research Institute for the Economics of Aging (MEA).

Börsch-Supan, A., McFadden, D. L., & Reinhold, S. (1996). Living arrangements: Health and wealth effects. In D. A. Wise (Ed.), *Advances in the economics of aging* (pp. 193–218). Chicago: University of Chicago Press. (with a comment by S. F. Venti)

Bound, J. (1991). Self-reported versus objective measures of health in retirement models. *Journal of Human Resources*, *26*, 106–138.

Bound, J., Schoenbaum, M., Stinebrickner, T., & Waidmann, T. (1999). The dynamic effects of health on the labor force transitions of older workers. *Labour Economics*, *6*, 179–202.

Bound, J., Stinebrickner, T., & Waidmann, T. (2008). *Health, economic resources and the work decisions of older men* (Working Paper No. 13657). Cambridge, MA: National Bureau of Economic Research.

Burkhauser, R. V., & Cawley, J. (2006). *The importance of objective health measures in predicting early receipt of social security benefits: The case of fatness* (Working Paper No. WP 2006-148). Ann Arbor, MI: Michigan Retirement Research Center, University of Michigan.

Carlson, P. (1998). Self-perceived health in East and West Europe: Another European health divide. *Social Science and Medicine*, *46*, 1355–1366.

Christensen, K., McGue, M., Yashin, A. I., Iachine, I. A., Holm, N. V., & Vaupel, J. W. (2000). Genetic and environmental influences on functional abilities among Danish twins aged 75 years and older. *Journal of Gerontology: Medical Sciences*, *55A*, M446–M452.

Coe, N. B., & Zamarro, G. (2008). *Retirement effects on health in Europe* (Working Paper). Santa Monica, CA: RAND Corporation.

Coile, C. (2004). *Health shocks and couples' labor supply decisions* (NBER Working Paper No. 10810). Cambridge, MA: National Bureau of Economic Research.

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

Currie, J., & Madrian, B. C. (1999). Health, health insurance and the labor market. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3, pp. 3309–3416). Amsterdam: Elsevier.

Drukker, D. M., & Gates, R. (2006). Generating Halton sequences using Mata. *Stata Journal*, *6*, 214–228.

Dwyer, D. S., & Mitchell, O. S. (1999). Health problems as determinants of retirement: Are self-rated measures endogenous? *Journal of Health Economics*, *18*, 173–193.

Erickson, P. (1998). Evaluation of a population-based measure of quality of life: The Health and Activity Limitation Index (HALex). *Quality of Life Research*, *7*, 101–114.

Erickson, P., Wilson, R., & Shannon, I. (1995). *Years of healthy life* (CDC/NCHS, Healthy people, Statistical Notes No. 7). Hyattsville, MD: US Department of Health and Human Services, National Center for Health Statistics.

Field, A. E., Coakley, E. H., Must, A., Spadano, J. L., Laird, N., Dietz, W. H., et al. (2001). Impact of overweight on the risk of developing common chronic diseases during a 10-year period. *Archives of Internal Medicine*, *161*, 1581–1586.

French, E. (2005). The effects of health, wealth, and wages on labour supply and retirement behaviour. *Review of Economic Studies*, *72*, 397–427.

Gordon, R. H., & Blinder, A. S. (1980). Market wages, reservation wages, and retirement decisions. *Journal of Public Economics*, *14*, 277–308.

Gouriéroux, C., & Monfort, A. (1991). Simulation based inference in models with heterogeneity. *Annales d'Économie et de Statistique*, *20/21*, 69–107.

Gruber, J., & Wise, D. A. (Eds.). (1999). *Social security and retirement around the world*. Chicago: University of Chicago Press.

Gruber, J., & Wise, D. A. (Eds.). (2004). *Social security programs and retirement around the world: Micro estimation*. Chicago: University of Chicago Press.

Gruber, J., & Wise, D. A. (2005). *Social security programs and retirement around the world: Fiscal implications: Introduction and summary* (Working Paper No. 11290). Cambridge, MA: National Bureau of Economic Research.

Gruber, J., & Wise, D. A. (Eds.). (2007). *Social security programs and retirement around the world: Fiscal implications of reform*. Chicago: University of Chicago Press.

Gustman, A. L., & Steinmeier, T. L. (1986a). A disaggregated structural analysis of retirement by race, difficulty of work and health. *Review of Economics and Statistics*, *68*, 509–513.

Gustman, A. L., & Steinmeier, T. L. (1986b). A structural retirement model. *Econometrica*, *54*, 555–584.

Gustman, A. L., & Steinmeier, T. L. (2000). Retirement in dual-career families: A structural model. *Journal of Labor Economics*, *18*, 503–545.

Hamilton, J. D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, *33*, 387–397.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–161.

Hillier, T., & Pedula, K. (2001). Characteristics of an adult population with newly diagnosed Type 2 diabetes: The relation of obesity and age of onset. *Diabetes Care*, *24*, 1522–1527.

Hurd, M. D. (1990). Research on the elderly: Economic status, retirement, and consumption and saving. *Journal of Economic Literature*, *28*, 565–637.

Jamshidian, M., & Bentler, P. M. (1999). ML estimation of mean and covariance structures with missing data using complete data routines. *Journal of Educational and Behavioral Statistics*, *24*, 21–41.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 user's reference guide*. Chicago: Scientific Software International.

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

Kopp, M. S., Skrabski, A., & Szedmak, S. (2000). Self-rated health and social transitions. In P. Nilsson & K. Orth-Gomer (Eds.), *Self-rated health in a European perspective* (pp. 85–102). Uppsala, Sweden: Swedish Council for Planning and Coordination of Research, Ord & Form AB.

Kunst, A. E., Bos, V., Lahelma, E., Bartley, M., Lissau, I., Regidor, E., et al. (2005). Trends in socioeconomic inequalities in self-assessed health in 10 European countries. *International Journal of Epidemiology*, *34*, 295–305.

Lee, L.-F. (1995). Asymptotic bias in simulated maximum likelihood estimation of discrete choice models. *Econometric Theory*, *11*, 437–483.

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

Little, R. J. A., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39–75). New York: Plenum Press.

Lumsdaine, R. L., & Mitchell, O. S. (1999). New developments in the economic analysis of retirement. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3, pp. 3261–3307). Amsterdam: Elsevier.

Macinko, J. A., Shi, L., Starfield, B., & Wulu, J. T. (2003). Income inequality and health: A critical review of the literature. *Medical Care Research and Review*, *60*, 407–452.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge, UK: Cambridge University Press.

Meijer, E., & Wansbeek, T. (2007). The sample selection model from a method of moments perspective. *Econometric Reviews*, *26*, 25–51.

Michaud, P.-C. (2005). *Dynamic panel data models and causality: Applications to labor supply, health and insurance*. Unpublished Ph.D. thesis, Tilburg University, Tilburg, The Netherlands.

Mitchell, O. S., & Fields, G. S. (1984). The economics of retirement behavior. *Journal of Labor Economics*, *2*, 84–105.

Murray, C. J. L., Salomon, J. A., Mathers, C. D., & Lopez, A. D. (Eds.). (2002). *Summary measures of population health: Concepts, ethics, measurement and applications*. Geneva, Switzerland: World Health Organization.

Must, A., Spadano, J., Coakley, E. H., Field, A. E., Colditz, G., & Dietz, W. H. (1999). The disease burden associated with overweight and obesity. *Journal of the American Medical Association*, *282*, 1523–1529.

Muthén, B. O. (1998–2004). *M*plus *technical appendices*. Los Angeles: Muthén & Muthén.

Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431–462.

Muthén, B. O., & Satorra, A. (1995). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*, *60*, 489–503.

OECD. (2005). *Pensions at a glance: Public policies across OECD countries*. Paris: OECD.

Rantanen, T., Guralnik, J. M., Foley, D., Masaki, K., Leveille, S. G., Curb, J. D., et al. (1999). Midlife hand grip strength as a predictor of old age disability. *Journal of the American Medical Association*, *281*, 558–560.

Rantanen, T., Harris, T., Leveille, S. G., Visser, M., Foley, D., Masaki, K., et al. (2000). Muscle strength and Body Mass Index as long-term predictors of mortality in initially healthy men. *Journal of Gerontology: Medical Sciences*, *55A*, M168–M173.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592. (with discussion)

Rust, J., & Phelan, C. (1997). How social security and medicare affect retirement behavior in a world of incomplete markets. *Econometrica*, *65*, 781–831.

Sammartino, F. J. (1987). The effect of health on retirement. *Social Security Bulletin*, *50*(2), 31–47.

Soldo, B. J., Mitchell, O. S., Tfaily, R., & McCabe, J. F. (2006). *Cross-cohort differences in health on the verge of retirement* (NBER Working Paper No. 12762). Cambridge, MA: National Bureau of Economic Research.

Sondik, E. (2002). Summary measures of population health: Applications and issues in the United States. In C. J. L. Murray, J. A. Salomon, C. D. Mathers, & A. D. Lopez (Eds.), *Summary measures of population health: Concepts, ethics, measurement and applications* (pp. 75–81). Geneva, Switzerland: World Health Organization.

Stewart, S., Woodward, R. M., & Cutler, D. M. (2005). *A proposed method for monitoring US population health: Linking symptoms, impairments, chronic conditions, and health ratings* (NBER Working Paper No. 11358). Cambridge, MA: National Bureau of Economic Research.

Stock, J. H., & Wise, D. A. (1990). Pensions, the option value of work, and retirement. *Econometrica*, *58*, 1151–1180.

Train, K. E. (2003). *Discrete choice methods with simulation*. Cambridge, UK: Cambridge

University Press.

van Doorslaer, E., Wagstaff, A., Bleichrodt, H., Calonge, S., Gerdtham, U.-G., Gerfin, M., et al. (1997). Income-related inequalities in health: Some international comparisons. *Journal of Health Economics*, *16*, 93–112.

Wansbeek, T., & Meijer, E. (2000). *Measurement error and latent variables in econometrics*. Amsterdam: North-Holland.

World Bank. (1993). *World development report: Investing in health*. New York: Oxford University Press.

Appendix

A Variables used

A.1 Health indicators

Mobility

Limitations with mobility, arm function & fine motor function. The table below gives the detailed content of these variables. They are all yes/no variables ("difficulty with ...").

Walking 100 metres
Sitting for about two hours
Getting up from a chair after sitting for long periods
Climbing several flights of stairs without resting
Climbing one flight of stairs without resting
Stooping, kneeling, or crouching
Reaching or extending your arms above shoulder level
Pulling or pushing large objects like a living-room chair
Lifting or carrying weights over 10 pounds/5 kilos, like a heavy bag of groceries
Picking up a small coin from a table

For the analysis, we have combined "Climbing several flights of stairs without resting" and "Climbing one flight of stairs without resting" into one ordinal variable with three categories: no difficulties, difficulties with several flights only, and difficulties with both. A small number of respondents reported difficulties with one flight but not with several flights. These have been included in the "both" category. We assume that in these respondents' view, the several flights category is implied by the one flight category. They were not asked explicitly about each category, but rather had to pick the ones that applied to them from a showcard.

ADL

Limitations with activities of daily living. They are also yes/no variables.

Dressing, including putting on shoes and socks
Walking across a room
Bathing or showering
Eating, such as cutting up your food
Getting in or out of bed
Using the toilet, including getting up or down

IADL

Limitations with instrumental activities of daily living. Again, they are yes/no variables.

Using a map to figure out how to get around in a strange place
Preparing a hot meal
Shopping for groceries
Making telephone calls
Taking medication
Doing work around the house or garden
Managing money, such as paying bills and keeping track of expenses

Self-reported health

We used self-reported health according to the U.S. categorization from "excellent" to "poor". This has a more symmetrical distribution than the European version, which uses a categorization from "very good" to "very bad". Half of the sample answered the U.S. version before the other health variables and the European version after, the other half of the sample answered in the reverse order. Unlike Crossley and Kennedy (2002), we have not found a systematic effect of the order of presentation on the distribution of this variable, and thus we combine the two half-samples in one variable.

Grip strength

This is the maximum of up to four measurements: two on the left hand and two on the right hand. This variable is missing if the original does not have two measurements on at least one hand or if these differ by more than 20 kg or had implausible values. In the analyses, we have divided the original variable by 10 to obtain better scaling.

A.2 Selection variables, covariates, and sampling weights

- *Country*, *Gender*, and *household size* are taken directly from the data.
- *Age* is taken from the imputations file. Age does not vary across imputations, so this is uniquely defined. Israel is not in the imputations file, so for Israel we computed age as interview year − birth year, where the birth year variable used is `dn003_` from the original data, which is the one answered by the individual respondents themselves. In the analyses, we used aged1 = (age − 65)/10, aged2 = aged1^2, and aged3 = aged1^3.
- *Living with a spouse/partner* is a constructed variable, using the variable `mstat` (answered by the household respondent), and the variable `dn014_`, as answered by the respondent and by the alleged spouse/partner, taking the information about household size into account, as well as whether either the respondent or the alleged partner was the household respondent

answering the `mstat` question. The end result is a variable that is almost the same as `mstat` recoded to 1 = living with a spouse or partner and 0 = living as a single, but with the "other" and "don't know" categories assigned to either 0 or 1, and four observations recoded from "living with a partner" to "single".

- *Education* is taken from the imputations file. This variable implements the 1997 International Standard Classification of Education (ISCED-97), giving a variable ranging from 0 (none) to 6 (second stage of tertiary education), and separate categories for "still in school" (95) and "other, abroad" (97). However, because in most countries, some of the categories are almost or completely empty, we combined categories 0 (none), 1 (primary education), 2 (lower secondary education), and 95 (still in school) into one category ("primary"), categories 3 (upper secondary education) and 4 (post-secondary, non-tertiary education) into the second category ("secondary"), and categories 5 (first stage of tertiary education) and 6 (second stage of tertiary education) into the third category ("tertiary"). Category 97 (other, abroad) was considered "missing". We have little hope that the resulting education variable is truly comparable across countries, but this recoding at least allows us to use the same model structure for different countries. The different coefficients for different countries are then (partly) a reflection of the differences in educational systems across countries. For observations in which this variable was imputed (rather than just derived from the original variables), we have used the first imputation.

- *Log household income* is the log of purchasing power parity (PPP) adjusted, before tax household income, in Euros. This variable was supplied as a generated variable in the supplementary dataset `gv_inc1`. This variable is based on numerous original variables. For many observations, one or more of these are missing, where a so-called unfolding bracket, leading to a range of possible values, also counts as a partial missing. Therefore, the data set contains five imputations for observations with missings. The variable used is (the log of) the average of the five imputed income values. From the resulting variable, we have replaced zero values and values over €1 million by missings.

- *IHS net worth* is the inverse hyperbolic sine transformation of the average of the five imputations of PPP adjusted household net worth in Euros, taken from the imputations file.

- *Body mass*. We use height and weight and a categorization of BMI into underweight, overweight, moderately obese, and severely obese dummies, with normal weight being the reference category. See the main text for the thresholds used for these dummies. We have replaced implausible values (weight less than 10 kg, height less than 110 cm) by missings. In the grip strength model, we use heightd1 = height/100, heightd2 = heightd1^2, weightd1 = weight/10, weightd2 = weightd1^2, and htwt = heightd1 $\times$ weightd1.

- *Sampling weight*. This is the individual, calibrated sampling weight variable (`wgtaci`). For the model estimations, we used a rescaled version, which for each country-gender

combination sums to the number of observations used in the health model, instead of the population size. This gives the same estimates as using the original weights, but the loglikelihood is now of the same order of magnitude as an unweighted (proper) loglikelihood. In the grip strength model, we have experimented with both, as well as unweighted analyses, as described in appendix B. In these models, the results from the analyses using differently scaled weight variables can be different, because population sizes are different and hence the rescaling per country-gender combination results in different weighting of countries. For Figures 1–6, we used the original `wgtaci` variable, so that the figures represent averages for a representative sample from the 12 countries jointly.

A.3 Reasons for retirement, eligibility for pensions, and other employment and pension variables

These variables have mostly been taken directly from the original data, although in some cases the original data had different variables for respondents who report different labor force statuses. In such cases, we have combined information from different original variables. The main text describes most of the choices we have made.

The most important issue to note here is that respondents from The Netherlands were not given the public pension option for future eligibility for pensions, because everyone who lived in The Netherlands for some time before age 65 is entitled to a public pension at age 65 (with the benefit amount reduced if one lived fewer than 50 years in The Netherlands between ages 15 and 65). Hence, we have imputed this future eligibility for each respondent younger than 65. We have not imputed current receipt of public pensions for respondents 65 or older, because from the description above, it follows that an individual who never lived in The Netherlands before age 65 is not entitled to public pensions after age 65. However, presumably most respondents over 65 who do not report receiving public pensions must misstate their situation, typically because the respondent's public pension is deposited in the spouse's bank account.

Further details of the variable constructions are available from the authors upon request.

B Grip strength correction for height and weight

Grip strength has been shown to be a good predictor of future medical problems (Christensen et al., 2000; Rantanen et al., 1999, 2000; Al-Snih et al., 2002). Therefore, this is a useful indicator of health. However, grip strength is also related to overall body size, with larger individuals being stronger on average than smaller people. Hence, we expect to obtain a better indicator of health if we develop a corrected measure that partials out the size effect. The conceptual model we use for this is

$$\mathrm{GS}_{cn} = \lambda' \eta_{cn} + \tau' p_{cn} + \varepsilon_{cn}, \tag{6}$$

where η_{cn} is "true health" of the n-th individual in the c-th country, which may be multidimensional; λ is a vector of coefficients; $\tau' p_{cn}$ is a polynomial in height and weight, with

p_{cn} consisting of height, weight, their squares, and their product, and τ a vector of coefficients to be estimated. Finally, ε_{cn} is a random error term. We do not assume that parameters are equivalent for males and females and thus, equations, variables, coefficients, and so forth, are gender-specific throughout, but this will be suppressed in the notation.

Our aim is to estimate τ and then subtract the height-weight polynomial $\tau' p_{cn}$ from the observed grip strength, thereby obtaining a better indicator of true health. However, true health is unknown, and thus, η_{cn} is a latent variable and we cannot estimate (6) directly as a linear regression model. Furthermore, p_{cn} may be correlated with true health, so we cannot confidently estimate τ by regressing grip strength on p_{cn} as well.

In our model, we have a predictive equation for true health, of the form

$$\eta_{cn} = \Gamma_c x_{cn} + \zeta_{cn}, \tag{7}$$

where x_{cn} is a vector of explanatory variables, Γ_c is a matrix of regression coefficients, and ζ_{cn} is a vector of random errors. Note that we do not assume that the regression coefficients in this equation are the same for different countries. Combining (6) and (7), we obtain

$$\begin{aligned} \mathrm{GS}_{cn} &= \lambda' \Gamma_c x_{cn} + \tau' p_{cn} + \varepsilon_{cn} + \lambda' \zeta_{cn} \\ &= \beta_c' x_{cn} + \tau' p_{cn} + u_{cn}, \end{aligned} \tag{8}$$

with β_c and u_{cn} implicitly defined. Clearly, this equation can be estimated by a regression of grip strength on p_{cn} and the explanatory variables in x_{cn}, with the latter interacted with country dummies.

However, in the SHARE data, there is a fairly large proportion of individuals for which grip strength is missing, and these tend to be individuals in worse health (according to other health indicators that are observed). Hence, estimating (8) by linear regression may suffer from selectivity bias. To alleviate this potential problem, we can estimate a Heckman selection model (Heckman, 1979; Meijer & Wansbeek, 2007). Because observing grip strength is related to health, the selection equation of this two-equation model includes a large number of other health indicators, interacted with the country dummies, in addition to the variables that are already in the equation of interest, (8).

Another potential issue is whether to include economic variables in x_{cn}. Health is related to socio-economic status, so in terms of more precise measurement of health, it may be advantageous to include a measure of income and/or wealth in its predictive equation. On the other hand, the measurement model consisting of (6), (7), and a large number of additional equations for other health indicators (mobility and functional limitations, ADLs and IADLs, self-reported health), will be used to compute a health index for usage as an explanatory variable in retirement models, and inclusion of income and/or wealth in the construction of the health index may give some endogeneity problems, especially with income depending on the retirement decision.

Two final considerations are the weights and the functional form of the relation between grip strength and height and weight. The weights that are supplied in the SHARE data sum to

population size within each country. If the model is correct, more efficient estimates are obtained with rescaled weights that sum to the sample size within each country. We employ both, as well as an unweighted analysis. Sensitivity to functional form is assessed by nonparametric regression using lowess estimators. This is done sequentially, inspired by the Frisch-Waugh theorem of linear regression (Wansbeek & Meijer, 2000, p. 352): We regress the dependent variable (grip strength) on one of the explanatory variables (height) and also regress the other explanatory variable (weight) on this explanatory variable. Both of these are done by lowess. Then the residuals of grip strength are regressed on the residuals of weight, and the corrected health indicator is the residual from this equation.

Thus, we have the following choices: (1) Three types of weighting: unweighted, with weights that sum to population size, and with rescaled weights that sum to sample size; (2) with or without selection equation; and (3) different model specifications. For the latter, we have a *base* set of explanatory variables, which includes the elements p_{cn} of the height-weight polynomial (consisting of height, weight, height squared, weight squared, height × weight), and the following additional explanatory variables: a cubic polynomial in age, education dummies, a dummy for living with a spouse or partner, household size, and body mass index categorized into five categories, plus dummies to capture missingness of various explanatory variables. These additional explanatory variables are all interacted with country dummies. If present, the corresponding selection equation contains the same set of variables plus a set of over 20 binary health indicators (difficulties with...) and self-reported health, which are also interacted with the country dummies. The set *base + log income* adds log household income plus its interactions with the country dummies, to the set of explanatory variables. The income variable used is the average of the 5 "imputations" provided. If this is zero, then log income is set to zero. A "zero income" dummy is added to capture the average effect of this group. The set *base + IHS net worth* adds the inverse hyperbolic sine of household net worth to the base set, plus its interactions with the country dummies. We use the inverse hyperbolic sine function rather than the log, because a nonnegligible fraction of households have negative net worth, which indicates less access to funds for investing in health, so it is meaningful to take this into account. The fourth set of explanatory variables, *only height-weight polynomial*, only includes the elements p_{cn} of the height-weight polynomial, and the fifth set, *Only height-weight, nonparametric*, estimates the relation between grip strength and height and weight nonparametrically, as indicated above. These two sets do not include the additional explanatory variables or any interactions with the country dummies.

For practical reasons, we do not compute all combinations of possibilities implied by the classification above, but a subset that should reflect the sensitivity of the resulting indicator to the choices made. The models estimated are listed in Table 22.

Apart from the arbitrary location and scale differences, which do not influence the usage as an indicator of health, the residuals of these models are extremely insensitive to the choices made in specification and estimation. Weights and presence or absence of a selection equation have no impact whatsoever, and neither does the presence or absence of the income or wealth variables in the model. The only discernible influence is whether the analysis is restricted to only height and

Table 22: Estimation settings.

No.	Weights	Selection eq.	Variables[a]
1	sum to sample size	yes	Base
2	sum to sample size	yes	Base + log income
3	sum to sample size	yes	Base + IHS net worth
4	sum to population size	yes	Base
5	sum to population size	yes	Base + log income
6	sum to population size	yes	Base + IHS net worth
7	unweighted	yes	Base
8	unweighted	yes	Base + log income
9	unweighted	yes	Base + IHS net worth
10	sum to sample size	no	Base
11	sum to sample size	no	Base + log income
12	sum to sample size	no	Base + IHS net worth
13	sum to sample size	no	Only height-weight polynomial
14	unweighted	no	Base
15	unweighted	no	Base + log income
16	unweighted	no	Base + IHS net worth
17	unweighted	no	Only height-weight polynomial
18	unweighted	no	Only height-weight, nonparametric

[a] See text for description.

weight (models 13, 17, and 18, group I, say) or the model includes other explanatory variables as well (models 1–12 and 14–16, group II, say). The correlations between residuals from group I on the one hand and residuals from group II on the other hand are between 0.98 and 0.992 for both males and females, whereas the correlations among residuals in group I are all above 0.998 for both males and females, and the correlations among residuals in group II are all above 0.999 for both males and females. This insensitivity is not (necessarily) due to lack of relation between grip strength and height/weight: For males, the absolute values of individual t-values of coefficients of the height-weight polynomial are typically between 4 and 7 and the coefficients are sizeable. For females, though, many coefficients are nonsignificant, although typically one or two are significant, with t-values between 2 and 3. Also, the inverse Mills' ratio is highly significant (absolute t-values between 8 and 25), which clearly points at a selection effect.

Concluding, we can confidently select one of the residuals as a health indicator and do not need to worry about misspecification of this equation. On the basis of theoretical considerations, we selected the residual from model 3, i.e., with a correction for selectivity bias, with all the base explanatory variables included, as well as wealth, and estimated with weights that sum to the sample size within country. Table 23 presents the means and standard deviations of grip strength before and after this correction for height and weight, by country and gender. This shows that there are some shifts in relative positions of some countries, e.g., Sweden drops two places for both males and females, but these shifts are minor. They do, however, illustrate that the corrections may be important at the individual level.

C Reliabilities and R^2's

The fit of linear regression models is usually assessed by means of the R^2. This can be done for each linear equation in the current model as well. However, because y_n^* and η_n are (typically) not directly observed, computation of the R^2 cannot be done directly. Instead, the R^2 measures are derived from the parameters. For a linear regression model $y_n = x_n'\beta + \varepsilon_n$, with ε_n independent of x_n, the R^2 is equal to

$$R^2 = 1 - \frac{\mathrm{Var}(\varepsilon_n)}{\mathrm{Var}(y_n)} = 1 - \frac{\mathrm{Var}(\varepsilon_n)}{\beta' \,\mathrm{Cov}(x_n)\beta + \mathrm{Var}(\varepsilon_n)},$$

where the constant has been removed from x_n and β. Furthermore, x_n is treated as random in this expression, which simplifies the formulas a bit.

The R^2's for the equations in the current model are defined analogously:

$$R^2(\eta_j) = 1 - \frac{\mathrm{Var}(\zeta_{nj})}{\mathrm{Var}(\eta_{nj})}$$
$$R^2(y_i^*) = 1 - \frac{\mathrm{Var}(\varepsilon_{ni})}{\mathrm{Var}(y_{ni}^*)}.$$

Table 23: Means and standard deviations of maximum grip strength by country and gender, before and after correction for height and weight (kg).

	Before correction		After correction	
Country	Mean	s.d.	Mean	s.d.
Males				
Austria	46.1	9.8	−168.5	9.5
Belgium	44.0	10.2	−169.9	9.6
Denmark	46.7	10.5	−168.1	9.6
France	42.4	10.7	−170.9	10.1
Germany	46.0	10.9	−168.6	10.4
Greece	41.2	11.1	−172.3	10.5
Israel	39.4	11.7	−174.0	11.4
Italy	39.7	11.1	−173.3	10.3
The Netherlands	45.5	10.4	−169.3	9.7
Spain	37.4	10.5	−174.4	9.6
Sweden	44.9	10.0	−170.0	9.3
Switzerland	44.3	9.5	−169.4	8.5
Females				
Austria	28.9	7.8	51.5	7.6
Belgium	26.2	7.1	49.4	6.7
Denmark	26.9	7.3	49.5	6.8
France	25.5	7.0	49.0	6.7
Germany	28.3	7.8	50.9	7.5
Greece	24.9	6.9	47.9	6.7
Israel	23.4	7.5	46.4	7.3
Italy	23.3	7.2	46.7	7.1
The Netherlands	27.7	7.6	49.9	7.1
Spain	22.3	7.6	46.3	7.3
Sweden	26.4	7.3	49.0	6.9
Switzerland	27.2	7.2	50.4	6.8

These definitions also mimick the ones in LISREL (Jöreskog & Sörbom, 1993, pp. 26–27).

If y_i^* is a continuous observed variable and there are no observed explanatory variables x in its equation, then $R^2(y_i^*)$ is its *reliability*. Correspondingly, the reliability of the health index $\hat{\eta}$ is the squared correlation between $\hat{\eta}$ and η, as derived from the model estimates.

D Scaled likelihood ratio test

With maximum likelihood estimation, nested models are usually tested against one another by means of the likelihood ratio (LR) test. The test statistic is $T_{\mathrm{LR}} = 2(\hat{L}_u - \hat{L}_r)$, where $\hat{L}_u$ is the maximum of the loglikelihood function for the "unrestricted" (i.e., less restricted) model, and $\hat{L}_r$ is the maximum of the loglikelihood function for the (more) restricted model. Under the null hypothesis, T_{LR} has an asymptotic chi-square distribution with degrees of freedom equal to the number of (effective) restrictions ν that the restricted model imposes on the unrestricted model. Typically, ν is the difference in number of free parameters.

When sampling weights are applied, the maximum values of the log-pseudolikelihood function are inserted in the expression for T_{LR}. But in this case, T_{LR} does not have an asymptotic chi-square distribution. It is asymptotically distributed as a weighted sum of independent chi-square variates with one degree of freedom. The weights can be estimated and quantiles and p-values corresponding to the resulting distribution can be estimated in principle. However, a convenient alternative that tends to work well in practice is to multiply T_{LR} by a scale factor such that the mean of its asymptotic distribution equals ν, i.e., the mean of the chi-square distribution with ν degrees of freedom, and then compare the test statistic with the quantiles of this chi-square distribution. Thus, the Scaled LR test statistic is $T_{\mathrm{SLR}} = cT_{\mathrm{LR}}$, where

$$c = \frac{\nu}{\mathrm{tr}(\hat{V}_u \hat{H}_u) - \mathrm{tr}(\hat{V}_r \hat{H}_r)},$$

in which $\hat{V}_u$ is the estimated asymptotic covariance matrix of the estimator of the parameter vector for the unrestricted model, $\hat{H}_u$ is the Hessian of the log-pseudolikelihood function in the optimum for the unrestricted model, and $\hat{V}_r$ and $\hat{H}_r$ are defined correspondingly for the restricted model. Note that for a proper (unweighted) loglikelihood function, $\hat{V}_u \hat{H}_u$ and $\hat{V}_r \hat{H}_r$ are identity matrices of different dimensions, the dimensions being equal to the number of parameters estimated in these models. Thus, the Scaled LR test reduces to the usual LR test. See Asparouhov and Muthén (2005) for more details and derivations.