NBER WORKING PAPER SERIES

HOW SURVEY DESIGN AFFECTS INFERENCE REGARDING HEALTH PERCEPTIONS AND OUTCOMES

Anneke Exterkate
Robin L. Lumsdaine

Working Paper 17244
http://www.nber.org/papers/w17244

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2011

The authors are grateful to Heather Anderson, James Banks, Martin Evans, Rachel Griffiths, Alastair Hall, Denise Whalen, and seminar participants at the London School of Economics and the University of Manchester for comments on an earlier draft. Financial support for the first author from the Van Beek Fonds is gratefully acknowledged. This paper was written while Lumsdaine served part-time on an Intergovernmental Personnel Agreement (IPA) with the National Institute on Aging. Address correspondence to Robin L. Lumsdaine. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research or the National Institute on Aging.

How Survey Design Affects Inference Regarding Health Perceptions and Outcomes
Anneke Exterkate and Robin L. Lumsdaine
NBER Working Paper No. 17244
July 2011
JEL No. C83,D03,D84,I1

ABSTRACT

This paper considers the role of survey design and question phrasing in evaluating the subjective health assessment responses using the Survey of Health, Ageing and Retirement in Europe (SHARE) dataset. A unique feature of this dataset is that respondents were twice asked during the survey to evaluate their health on a five-point scale, using two different sets of descriptors to define the five points, with the ordering of which set was first given determined randomly. We find no evidence to refute the assertion that the order was determined by random assignment. Yet we document differences in the response distributions between the two questions, as well as differences in inference in comparing the two populations (those that were asked one question first versus those that were asked the other). We then consider determinants of the degree of concordance between the two questions, as well as the determinants of individuals that provide conflicting responses. There appears to be evidence to suggest that individuals' assessments of their health in response to the second question may be influenced by the battery of health questions that were asked following the first assessment. We find that information in self-assessed health responses is useful in examining health outcomes. Our results suggest that adjusting such responses to take into account framing and sequencing of questions may improve inference. In addition, we show that accounting for survey design may be important in models for predicting outcomes of interest, such as the probability of a major health event.

Anneke Exterkate
Erasmus University
Burgemeester Oudlaan 50
3062 PA Rotterdam, Netherlands
a.exterkate@gmail.com

Robin L. Lumsdaine
Kogod School of Business
American University
4400 Massachusetts Avenue NW
Washington, DC 20016
and NBER
robin.lumsdaine@american.edu

Introduction

Questions that rely on self-assessment are now commonplace in health surveys – numerous researchers have documented their usefulness as a proxy for objective health measures when the latter are not readily observed (e.g., Damian, Ruigomez, Pastor, and Martin-Moreno [1999]; Pinquart [2001]; Simon, et al. [2005]). There is less consensus regarding the extent to which they reflect individuals' true health, as there is often substantial heterogeneity in subjective responses among individuals. Separately, subjective responses often provide interesting insights into individuals' preferences and attitudinal biases such as optimism (Van Doorn [1999]), which then can be considered in models of economic decision-making. Used in conjunction with more objective measures, transformations of subjective responses have been shown to alter inference and improve predictability of outcomes in a variety of contexts (Bassett and Lumsdaine [1999, 2000]).

There is also a substantial body of research that has shown that individuals' responses in surveys largely depend on the design of a question and its answer choices. This cognitive phenomenon, known as anchoring, was first theorized by Tversky and Kahneman [1974] who found that when people were first asked about a random subject *"Do you think it will be more or less than X?"*, and then were asked to make an estimation of the value, the value of *X* provided in the first question influenced peoples' response to the estimation question. Anchoring is also discussed in Schwarz [1999], who finds that survey responses are guided by the way in which a question is asked and can depend on the response scale that is offered.

In the context of health surveys, the anchoring of responses was considered in an earlier study using Australian data (Crossley and Kennedy [2002]) that documented response variation even when the same exact self-assessed health question is asked twice – first at the beginning of the health section of the questionnaire, and the second time after a battery of health questions. The study found that 28% of the sample changed their response the second time, where the change in response was partially attributable to the intervening health questions. The authors also argued the change may have been a reflection of individuals' own uncertainty about their actual health, since changes occurred more frequently among those who were older, unemployed, or in the low-income group. Similarly, in a political context using data from the National Election Studies, Zaller and Feldman [1992] argued that the way that individuals respond to survey questions contains much randomness, noting that when the same question is asked twice, only about half of the respondents give the same answer both times. They too argue that individuals' responses are largely dependent on previous questions. Both of these issues, response randomness and dependence on question sequencing of individuals' responses to survey questions related to health, are investigated further in this study.

A unique feature of the Survey of Health, Ageing and Retirement in Europe (SHARE) provides an opportunity to explore the question of anchoring further. In this survey, individuals were twice asked to rate their health using a five-point scale, each time with different adjectives assigned to each point in the scale. In addition, the choice of which scale an individual was offered first was made by random assignment. In between the two

questions, individuals responded to about 20 other questions, including a battery of subjective questions about their health. This unique design enables investigation of a number of interesting questions regarding the extent to which anchoring occurs and whether individuals interpret the scale in absolute or relative terms: (1) Do different characterizations of the points of the scale lead to different responses?, (2) To what extent do respondents focus on the word associated with the scale, rather than the numerical scale itself? That is, do people evaluate the question in absolute terms, responding the same word both times regardless of the placement of that word within a scale, or do people consider the question in relation to a broader population by always choosing, for instance, the middle answer on a response scale?, (3) Does it matter in which order the questions were asked?, (4) How much variation is there in individuals' self-assessed health and how does it relate to actual health?, (5) If there is predictable variation in responses to the self-assessed health questions, can that information be used to improve predictions of future health outcomes?

An important study by Jürges, Avendano, and Mackenbach [2008] first documented differences in individuals' responses to both versions of the self-assessed health question using a subset of five countries from the SHARE dataset (described in more detail below); they concluded that the two versions of self-assessed health are comparable, that is, despite some differences, overall they measure the same underlying general health and hence either version can be used interchangeably for subsequent analysis. They also considered two types of concordance between the two responses: (1) literal concordance, defined as the event where "an individual's response to both versions is verbally consistent regardless of the self-rated health version" (e.g., a respondent answers *"Very good"* to both versions of the self-assessed health question, regardless of where on the response scale this adjective lies), and (2) relative concordance, which they defined as the event where "an individual's responses to both versions are consistent in terms of their position on the self-rated health scale" (e.g., an individual answers *"2"*, regardless of the descriptor given for that number). Overall, they found evidence of literal concordance for 69.0% of their sample and relative concordance for 30.1%. In their framework, however, the two definitions are not mutually exclusive; that is, some individuals can be both literally and relatively concordant. We discuss this in more detail below.

The paper proceeds as follows: Section I reviews the literature. Section II describes the data construction and descriptive statistics related to key explanatory variables. Section III considers features of the main questions of interest – the self-assessed health questions. Section IV examines the determinants of differences in responses to the two versions of the questions and concordance type. In section V, models to predict changes in self-assessed health and the probability of a major health event are estimated. The final section concludes.

I. Literature Review

Numerous studies have shown that self-assessed health in surveys is a useful proxy for individuals' true general health, as it is strongly correlated with more objective measures. For example, Damian, Ruigomez, Pastor, and Martin-Moreno [1999], using a sample of community-dwelling elderly in Spain, aged 65 and over, showed that

self-assessed health is largely explained by age, the number of chronic conditions and functional status. This was also documented by Pinquart [2001], who found an age-related decline in self-assessed health as well as strong correlations between physical health, functional health and mental health versus self-assessed health, noting that the association of self-assessed health with physical health was the strongest. Moreover, according to the above-mentioned study by Crossley and Kennedy [2002], the probability of being employed was over 40% lower for individuals who assessed themselves to have poor health versus the ones that indicated their health as being good, suggesting the usefulness of self-assessed health for informing studies of economic outcomes such as labor force attachment.

In addition, however, there is evidence that self-assessed health is more than just an assessment of an individuals' physical health (Simon, et al. [2005], using data from the GLOBE study, a longitudinal study to explain socio-demographic inequalities in health in the Netherlands). During the survey, individuals were first asked to assess their general health on a 5-point scale ranging from *"very good"* to *"poor"*, and after that they were asked to explain the factors that went into their response. In their explanation, respondents referred to many different aspects of health; besides physical health aspects, they also mentioned functional performance, the way they cope with existing illnesses, their wellbeing, and health behavior factors. From this evidence, the study concluded that self-assessed health is a multidimensional concept.

But if self-assessed health reflects more than physical health, how reliable a proxy is it? As noted above, variation in subjective responses may reflect peoples' preferences and/or attitudinal bias, for example their level of optimism. When using a five-point scale, not everyone may interpret the scale in the same way, for example, one individual might think of response option 1 as being more exceptional (i.e., more of a distributional tail) than another would. Forcing homogeneous interpretation of scales may lead to measurement error and biases when using latent variable models to explain underlying health. Using a small sample of individuals aged 65 and over in the New Haven, Connecticut, area, Van Doorn [1999] found evidence of over-optimism among elderly individuals, with many too optimistic and almost no one too pessimistic about his/her health, when comparing self-assessed health to objective health measures. The role of optimism is also explored in Bassett and Lumsdaine [1999] who note that adjusting subjective responses to account for differences in optimism can lead to improved accuracy in predicting economic outcomes of interest. The authors also found evidence of a common component in a series of subjective response questions and demonstrated that controlling for unobserved individual heterogeneity, particularly in samples where the respondent may not fully have understood the question, improves the ability of such questions to predict subsequent economic outcomes (Bassett and Lumsdaine [2000]).

Besides the potentially worrying results that responses depend in predictable ways on question ordering (Zaller and Feldman [1992]; Crossley and Kennedy [2002]) and that individuals' responses may reflect attitudinal biases that are unrelated to the question being asked, individuals' perception of their own health and the way in which they respond to a self-assessed health question also differs according to observable characteristics. Self-reported objective and subjective measures of physical health have been shown to be related to certain aspects of

personality, such as neuroticism (Costa Jr. and McCrae [1985]); another study found that there are cultural differences in reporting self-assessed health (Zimmer, Natividad, Lin, and Chayovan [2000]). This latter study used data from three different countries (the Philippines, Taiwan and Thailand) and showed that although the determinants of self-assessed health are the same across the different countries, the overall distribution of the probabilities of answering each of the response options differs, even after controlling for more objective health measures. This suggests that different individuals have different perceptions or tendencies when reporting their self-assessed health. A similar result is found when comparing self-assessed health across European countries using only one of the two available response scales in SHARE (Jürges [2006]): individuals in Scandinavian countries report themselves to be the healthiest and individuals in Southern Europe assess themselves to be the least healthy. But when controls for more objective health measures are added, Danish and Swedish respondents tend to over-rate their health, while those from Germany tend to under-rate their health. Even after controlling for differences in response styles among countries, such cross-country variations in general health are reduced but not eliminated. In contrast, Moum [1992] considered data from a large sample of Norwegian adults and found that there was little variation in self-assessed health status after controlling for sufficient detailed information on health. The heterogeneity of responses and the extent to which observable differences may play a role in different responses to self-assessed health questions is a key focus of this study.

Others have demonstrated that aspects of question framing, such as the choice of numerical scale or the wording of a question, can influence the distribution of responses in surveys. For example, evidence from a survey in which a question was asked about how successful individuals would say they had been in their life found that when a scale from -5 to 5 was used, 34% of the responses were in the range between -5 and 0, while only 13% ended up between 0 and 5 when the scale was shifted upwards by five points to a (0,10) range (Schwarz [1999]). This result suggests that the interpretation of a symmetrical scale differs from that of a nonsymmetrical one for some individuals. Despite this striking difference, however, others have suggested that inference using such measures may not differ appreciably. For example, the use of three different scales (a five point scale, a seven point scale, and a qualitative scale where respondents were asked to compare their health to that of others) for a question on self-assessed health all led to the same perception regarding underlying health (Eriksson, Undén, and Elofsson [2001]). Similarly, an article by Hernández-Quevedo, Jones, and Rice [2005] suggested that changing the wording of the response options to the self-assessed health question did not seem to have a significant effect on the relationship between various factors and self-assessed health. Using data from the British Household Panel Survey, where the wording of the response options was changed in the ninth wave of the survey but then changed back to their original version from wave ten onwards, they found that, although this change caused a shift in the thresholds respondents used to map underlying health status to the response options of self-assessed health, there was no change in the relationship between various socio-economic characteristics and self-assessed health since the change corresponded more or less to a location shift of the response options along the 5-point scale.

There also exists a wide range of literature about the role of self-assessed health status in predicting economic outcomes of interest. A number of articles have shown that self-assessed health is an important

determinant of the probability of death among the elderly (e.g, Lee [2000]; Zimmer, Natividad, Lin, and Chayovan [2000]; Ford, Spallek, and Dobson [2007]) and of future health changes (Zimmer, Natividad, Lin, and Chayovan [2000]; further, Møller, Kristensen, and Hollnagel [1996] show how self-assessed health predicts coronary heart disease). Given the evidence that a decline in cognitive abilities is associated with greater risk of mortality (Anstey, Luszcz, Giles, and Andrews [2001]; Baker, Wolf, Feinglass, and Thompson [2008]), it is natural, therefore, to consider whether self-reported health responses might reflect cognitive functioning.

Indeed, previous research has already shown that individuals' cognitive function is of great importance concerning subjective responses and decision making. A study of the older US population, using data from the Health and Retirement Study, found that while both numeracy and financial literacy are very low among the entire population, lack of financial literacy is found to be the highest among the most elderly, females, those with the lowest level of educational attainment, African-Americans and Hispanics (Lusardi [2008b]). The study documented that numeracy and financial literacy matter for planning, as the more questions on both characteristics an individual answered correctly during the survey, the larger the probability was that s/he engaged in retirement planning. Another paper showed that lack of information and knowledge (about, for example, Social Security) caused people to make inferior decisions concerning their saving behavior (Lusardi [2008a]). For numerous economic decisions, an individuals' financial literacy and numeracy score (which measure specific aspects of cognitive function) plays an important role.

II. Data and Descriptive Statistics

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a cross-national biennial survey of individuals who were aged 50 or over when first interviewed, and their spouses.[1] Eleven European countries participated in the first wave of SHARE in 2004, representing a balanced mix of European regions: Scandinavia (Denmark and Sweden), Central Europe (Austria, Belgium, France, Germany, the Netherlands and Switzerland) and the Mediterranean (Greece, Italy and Spain); additionally, Israel also contributed to the data in the first wave. Over 31,000 individuals were interviewed on several topics, including social and demographic background, physical and mental health, employment and financial situation (see Börsch-Supan, et al. [2005] for a comprehensive summary of results from the first wave of SHARE). In the second wave three additional countries were added (Czech Republic, Ireland and Poland); all countries except for Ireland and Israel also participated in the third wave in 2008-2009.[2]

[1] As a condition of use of the SHARE dataset, we note that, "This paper uses data from SHARELIFE release 1, as of November 24th 2010 or SHARE release 2.4.0, as of March 17th 2010. The SHARE data collection has been primarily funded by the European Commission through the 5th framework programme (project QLK6-CT-2001- 00360 in the thematic programme Quality of Life), through the 6th framework programme (projects SHARE-I3, RII-CT- 2006-062193, COMPARE, CIT5-CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th framework programme (SHARE-PREP, 211909 and SHARE-LEAP, 227822). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01 and OGHA 04-064, IAG BSR06-11, R21 AG025169) as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions)".

[2] In Ireland the interviews for the second wave only took place in 2008, and therefore the country did not participate in the third wave at that time.

Hence by the writing of this paper, the SHARE dataset consisted of detailed life histories of over 43,000 individuals in fifteen countries. Because of the comprehensive nature of the SHARE dataset, it is ideal for cross-country comparisons of many different factors, including self-assessed health.

A. Sample Construction

The construction of our data sample is given in Table 1. We begin by focusing on the first wave of SHARE, because that is the only wave that used random assignment to obtain responses to the two different self-assessment questions. Out of the total number of 31,115 interviewed individuals in this first wave, we first decided to omit Israel from our analysis, for a number of reasons,[3] resulting in a reduction of 2,598 individuals. The sample was also restricted to individuals between the age of 50 and 75, a further reduction of 1,179 individuals under age 50 and 4,543 over age 75, and 17 individuals for whom age was missing.[4] Another 112 individuals are omitted from the sample because they did not have an individual sampling weight. We additionally deleted those individuals that did not have responses to both versions of the self-assessed health questions, resulting in the loss of another 118 individuals. Our baseline sample therefore contains 22,548 individuals.

Because we want to examine the linkages between self-assessed health and other factors, we deleted a further 417 observations that were missing information on key covariates, such as socio-demographic information (marital status and educational attainment – 4 individuals), cognition (word recall, numeracy score and depression score – 314 individuals) or health information (chronic diseases, number of times talked to a medical doctor, hearing ability and eyesight – 99 individuals).[5] After all this, our final wave 1 sample consists of **22,131** individuals.

Unfortunately, not all individuals who were interviewed in the first wave were re-interviewed in the second wave of SHARE. Some of them died, others refused to participate again and others were simply not re-contacted. In total, out of the 31,115 individuals that were interviewed in wave 1, 18,742 individuals (60%) participated again in the second wave.[6] Among our initial wave 1 sample of 22,131 individuals, 15,052 were re-interviewed in wave 2,

[3] Israel is excluded from the analysis because 1) the data in Israel was gathered later than in the other countries (individuals were only interviewed in 2005-2006); 2) weights and imputations for Israel are computed in a different way than for the European countries; 3) later in the paper we additionally consider wave 2 of SHARE which does not include Israel; 4) because they interviewed participants in three different languages (Russian, Arabic and Hebrew), if we split out the sample size by language, the Israeli sample sizes become small; and 5) it is not a European country.

[4] As a robustness check concerning the decision to exclude those above age 75, we re-did much of the analysis with an enlarged sample that included those up to age 85. The main difference between the larger sample and the one we use is that the average health declines, resulting in a larger proportion of individuals for whom we have no basis to distinguish between types of concordance (defined below). As the main focus of our study is on the other types of concordance, and because the main results were not very sensitive to the additional individuals in the sample, we maintain the sample restriction of those between age 50 and 75.

[5] 21 blind individuals are still included in the sample; we treat them as a separate group concerning the self-assessed eyesight questions. We default the eyesight of blind individuals to a sixth category (additional to the five response options), signifying their eyesight is worse than "poor".

[6] This represents a much higher attrition rate (nearly 40%) than in the Health and Retirement Study, where the average re-interview response rate over all waves is 93.5% (Health and Retirement Study [2008]).

representing a slightly higher percentage (68%) than in the overall population.[7] Applying the same filters as in wave 1, our final wave 2 sample consists of **14,768** individuals, which represents slightly more than 33% attrition for our particular wave 1 sample, less than the overall attrition rate.

Except where noted, respondent-level individual weights are used, to reflect sample design and overweighting.

B. Descriptive Statistics

Table 2 provides weighted means and standard deviations of the main explanatory variables used in our analysis for the overall sample, as well as weighted means for each country.[8] Three sets of variables are discussed in turn: socio-demographic, health, and cognition. We focus primarily on the overall sample since many cross-country differences are documented in Börsch-Supan, et al. [2005] and Börsch-Supan, et al. [2008].

1. Socio-demographic variables

Out of the wave 1 sample of 22,131 individuals, the average age is 61.9 years, 52.3% are female, 90% are interviewed in their country of birth and 70.8% report themselves to be married.

Education is measured through questions about highest school degree, from which a corresponding level of educational attainment is derived.[9] This coding ranges from 0 to 6, where 0 denotes pre-primary education (pre-primary school, children aged 3 to 5) and 6 denotes the second stage of tertiary education (obtaining a Ph.D.). There is also a separate category for all "other" education, for example, adult education and special education. Out of the sample of 22,131 individuals, 165 individuals were assigned this special category; they are excluded from the summary statistics in this table since the other categories follow a natural rank order. In subsequent analysis, the "other" category will be included, and codes 5 and 6 will be combined into a single category to avoid the small sample sizes that would otherwise result. Excluding the individuals in the "other" category, the average ISCED-97 level of education is 2.5, which is approximately equal to a high school degree. As might be expected, the level of education varies a lot across countries; it is highest in Denmark with a weighted average of 3.4, and lowest in Spain with a weighted average of 1.6.

[7] This higher percentage is likely due to our age screen, which filters out the oldest individuals. Of the 7,079 that were not re-interviewed in the second wave, 262 are individuals died.

[8] A complete set of summary statistics by gender, marital status, education category, and those born in the country in which they were interviewed is available from the authors on request.

[9] To make the levels of education cross-country comparable, the 1997 International Standard Classification of Education (ISCED-97) as designed by the United Nations Educational, Scientific and Cultural Organization (UNESCO; see United Nations Educational, Scientific and Cultural Organization [2006] for details on ISCED-97 coding) is used as standard coding for education.

Data on each individual's income is constructed from individual-level components in SHARE while net worth is measured at the household level and is therefore assigned to both parties of a married couple rather than attempting to apportion it between the two members.[10] The income and net worth information is in local currency units; to make these variables comparable across countries, their levels were adjusted by country-specific inflation in the year of the survey and converted to euros using exchange rates that are constant across all respondents within a country. In the table the means and standard deviations of the natural logarithms of income and net worth are shown, as these are the variables used in further analysis.[11] In terms of euros, the weighted average adjusted income over the year before the interview is equal to €27,003.59, where the weighted average median is equal to €19,343.93. The weighted average adjusted household net worth is equal to €321,687.20 with a weighted average median of €161,859.80, indicating substantial skewness in both income and net worth.

2. Health

There is a rich variety of health information in SHARE. Individuals were shown a list of fourteen different chronic diseases (for example, heart failure, high blood pressure, diabetes, asthma, rheumatism and Parkinson's disease) and eleven health symptoms (e.g., pain in joints, difficulty breathing, sleeping problems and dizziness) and asked which apply to them; the number of both chronic diseases and health symptoms enter into the analysis below separately. On average, individuals report having 1.4 of each. Individuals were also asked to rate their eyesight and hearing on a five-point scale, ranging from *"1. Excellent"* to *"5. Poor"*, where the average response to each of these lies between the responses *"2. Very good"* and *"3. Good"* (2.7 and 2.6, respectively). For reasons that will become apparent below, the descriptors attached to this five-point scale are referred to as the "US version" (see below). The average self-reported number of times that a respondent has talked to a medical doctor in the past year is seven.

3. Cognition

A number of questions intended to measure cognitive functioning are available in SHARE: (1) Two questions asking respondents to rate their reading skills and writing skills, also on the US scale, where 18 individuals who either answered *"don't know"* or who refused to answer were assumed to be in the poorest category. The average responses are 2.4 and 2.6, respectively, corresponding to an answer in the middle of *"2. Very good"* and *"3. Good"*; (2) Individuals were given a list of ten words and then asked to recall as many as they could, first immediately after they heard the list, and then again after responding to a number of additional questions. Immediately after hearing the list, individuals on average remembered almost five out of the ten words (4.9), where with delayed recall the average declined to less than 3.4; (3) A sequence of numeracy questions (for example, to compute 10% out of 1,000) that increased in difficulty when a question was answered correctly and otherwise

[10] For this reason, an interaction between marital status and net worth is included as a separate variable in the regression analysis.

[11] Only one individual out of the sample has a negative income; for simplicity we set this individual's ln(income) observation equal to zero. There were 692 individuals with a negative household net worth; due to the number of individuals involved, we thought it more appropriate to use minus the logarithm of the absolute value of the household net worth.

decreased. A numeracy score was computed, based both on how many were answered correctly and on the difficulty of the questions, ranging from 1 to 5, where 1 is the worst and 5 is the best score. On average, individuals had a score of 3.3; (4) Individuals were asked which day of the week, which day of the month, which month and which year it is. From their responses, an orientation scale was constructed, defined as the number of questions answered correctly (the maximum is four). The majority of respondents answered all four questions correctly; over our sample the weighted average orientation score is 3.8; (5) Individuals were asked several questions about their feelings, from which an index of depression is constructed.[12] This is a scale ranging from 0 to 12, where a score greater than 3 means that "the respondent has clinically significant symptoms of depression" (Börsch-Supan et al. [2005]). The average for the whole sample is equal to 2.3.

Later on in this paper, several regressions are performed, in which most of the above variables are included as dummy variables. To avoid multicollinearity, it is therefore necessary to omit one category for every set of dummy variables. A list of all suppressed categories, used in all regressions, can be found in Table 3.

III. Self-Assessed Health Questions

The main variables of interest in this paper are the two self-assessed health questions that were asked in SHARE; one at the very beginning of the health section of the interview, and the other immediately after a battery of other health-related questions were asked. For both versions of the self-assessed health question, the phrasing began with *"Would you say your health is…*", after which respondents were asked to choose from five different response options associated with different general descriptors. In one case the response options were given by the scaling used by the World Health Organization: "*1. Very good, 2. Good, 3. Fair, 4. Bad, 5. Very bad"*, hereafter referred to as the "WHO version" of the health question; in the other case, the response options were: *"1. Excellent, 2. Very good, 3. Good, 4. Fair, 5. Poor"*, from now on referred to as the "US version" of the health question.[13] Which question individuals were asked first was chosen by random assignment; the other version of the question was then asked immediately following the end of the health section of the interview. Note that in both cases, the scaling of the response options is such that a lower number means better self-assessed health, while a higher number indicates worse self-assessed health.

The exact wording of the response options differs in the different interview languages that were used across the countries, therefore affecting the number of response options that could exactly correspond across the two scales. In the English example above, WHO-version response options 1, 2 and 3 correspond to US-version options 2, 3 and 4. But in none of the countries was English used as the interview language. For half of the languages used (i.e.,

[12] This index corresponds to the EURO-D scale included as part of SHARE when respondents provided answers to all twelve questions; when fewer than twelve binary responses were available, the average of the available responses was re-normalized by multiplying by twelve.

[13] These latter descriptors have been associated with the five-point scale for self-assessed health in a number of large, longitudinal US-based surveys, in particular the Health and Retirement Study (HRS), on which the SHARE survey instrument is modeled.

Dutch/Flemish, German, Greek and Spanish), four words are the same (WHO-version 1, 2, 3 and 4 correspond to US-version 2, 3, 4 and 5). In the other languages, however, fewer response options correspond. In French, only WHO-version responses 1 and 2 correspond to US-version responses 2 and 3; in Italian, WHO-version 1, 2 and 3 correspond to US-version 2, 3 and 4; in Danish, WHO-version 2, 3 and 4 correspond to US-version 3, 4 and 5. In the Swedish language, there is a skip: WHO-version responses 1, 2 and 4 correspond to US-version responses 2, 3 and 5.

A. Comparing the two versions of self-assessed health

Table 4 contains weighted average responses to the two versions of the self-assessed health question, both for the whole sample and for different subsamples according to socio-demographic characteristics of the respondents. Overall, the weighted average response to the WHO version of this question is equal to 2.34 (between *"Good"* and *"Fair"*), while for the US version of the question it is equal to 3.03 (very close to *"Good"*). This is surprising, since the wording of the response options might suggest a one-point difference between the two versions. The difference of 0.69 suggests that not all individuals select according to the terminology used; they also may pay attention to the position on the response scale that their answer is located. We return to this issue later on in the paper. Among the subsample of 18,601 individuals who could have given the same response to the second health question (i.e., their first response corresponds exactly to a choice on the scale of the second question) [14], the weighted averages are equal to 2.27 for the WHO version and 3.01 for the US version, a difference of 0.74 points.

Comparing the average self-assessed health corresponding to the two questions by a number of different characteristics, the difference between the two versions is close to the sample average of 0.69 points in almost every case. Only a few subsamples show statistically significant differences. The most noticeable are the individuals in Denmark and Sweden, where the differences between the two versions of the self-assessed health question are only 0.41 and 0.31, respectively. Since for most of the socio-demographic subgroups the differences between the two versions are reasonably close to the difference of 0.69, it is instructive to look in more depth at one of the two versions of the question, say the US version, to get a better idea of variations in self-assessed health across subgroups. Across the countries, those in Germany, Italy and Spain report the highest scores on self-assessed health, indicating that they assess themselves to be the least healthy. Those in Denmark, Sweden and Switzerland report being the healthiest. By gender, females assess themselves as being a little bit less healthy than males; around 0.15 points. In the oldest age group (aged 70 to 75), average response to health is 0.66 points higher than in the youngest age group; consistent with the intuition that health seems to be declining with age. A similar result holds for the different education levels: those in the highest educational attainment category report being almost 1 point healthier than those in the lowest educational attainment category. Furthermore, married individuals report being on average

[14] Some individuals did not have the possibility of responding using the same word to the second self-assessed health question as they did in their response to the first; this was either because they responded *"Excellent"* to the first question if the US version was asked first, or *"Very bad"* to the first question if the WHO version was asked first (since these options are not offered in the other version), or as noted above it can be due to linguistic differences in response options so that individuals were forced to choose a different word as a response to the second question.

0.12 points healthier than single individuals and individuals that were born in the country of interview report being on average about 0.15 points more healthy than people with another country of origin.

It is also interesting to consider the distributions of responses to the two self-assessed health questions. The proportion of individuals that answered each of the 25 possible combinations of responses to the two self-assessed health questions can be found in Table 5. Surprisingly, the distribution of responses to the asymmetric US scale is more symmetric than the response to the symmetric WHO scale (see Figure 1). On the US scale, 41.7% selected the middle choice (*"Good"*); the distribution of the other four response options is more or less symmetric around this middle one. Curiously, the option that was most often selected using the WHO scale was also *"Good"* (option 2 on that scale, selected by 46.6%). Further, nearly 30% answered *"Fair"*, the third option on the WHO scale; the other three options were much less frequently chosen.

The proportion of respondents that gave different responses to the two versions of the self-assessed health question, considering only the subsample of those that had the *possibility* of answering the same word the second time (18,601 individuals), is 35.0%. This percentage is larger than the 28% that Crossley and Kennedy [2002] found; recall that in their study, both times the same response options were offered, whereas in SHARE the response options differ across the two questions. However, even within this SHARE subsample, the proportion differs dramatically depending on which version of the question an individual was first asked. Among those who were exposed to the WHO scale first, 41.8% did not provide the same response when subsequently asked the US version, while among those that received the US version first, only 28.6% failed to select the same response the second time around (when asked using the WHO scale).

B. Testing the Assertion of Random Assignment

According to the SHARE documentation (Mannheim Research Institute for the Economics of Aging [2010]), the assignment regarding which version of the health self-assessment an individual was first asked was random. To verify this assertion, Table 6 contains the same summary statistics as above, but now separately for the subsamples of individuals who were first asked the WHO version and those who were first asked the US version. Since the random assignment did not take into account sampling design, to test for randomness in the subgroups, the unweighted sample is used to compute summary statistics.[15] The table verifies that there is little observable difference between those who were first asked the WHO version and those who were first asked the US version. Tests of the equivalence of means indicate there are no statistically significant differences between the two subsamples with respect to any of the key socio-demographic, health, or cognitive variables considered. Therefore, it is reasonable to adopt the assumption that for the most part, the assignment was indeed random. In one dimension, however, there is a statistically significant difference across the two subsamples, an important deviation from the conclusion of truly random assignment: the average response to the WHO version of the question is statistically

[15] In all other parts of this paper, the weighted sample is used.

significantly lower (which denotes people assessing themselves to be healthier) when individuals are asked the US version first, while the average answer to the US version is statistically significantly higher (indicating people assess themselves to be less healthy) for those who were asked the WHO version first. This suggests a possibility that **individuals' responses are influenced by the sequencing of the questions they are given**.

C. Concordances

This section revisits Jürges, Avendano, and Mackenbach's [2008] original observation that despite differences in individuals' responses either scale could be used for inference, in light of the above results documenting the influence of the ordering of the two versions of the self-assessed health question, by investigating the following questions: Do individuals answer the same word to both versions of the self-assessed health question, despite being presented with a different scale? And which individuals are more likely to compare their health to others of their own age and demographic background, using a fixed numerical level as a reference point, regardless of the wording of the response options?

To answer these questions, we divided the sample into four (mutually exclusive) concordance categories:[16] (a) word concordance, defined as the weighted proportion of the sample that answered exactly the same words to both the WHO version and the US version of the self-assessed health question (for example, *"Good"* as a response to both versions), (b) numerical concordance, defined as the weighted proportion of the sample that answered both versions of the question using the same number (for example, response option 3 to both versions, although 3 corresponds to *"Good"* in the US version, and *"Fair"* in the WHO version), (c) discordant responses, and (d) responses where there was no basis to compare concordance. As explained above, because of differences in translations the number of response options used to determine word concordance differs across countries. In general, word concordance is based on the four response options that are the same in both versions of the self-assessed health question, but for a few languages, word concordance can only be based on two or three response options that are the same, due to differences in wording. For this reason, Jürges, Avendano, and Mackenbach [2008] restricted their sample to only the five countries for which all four response options were the same (i.e., Austria, Germany, Greece, the Netherlands, and Spain). In this study, we take a slightly different approach. For example, if someone in France answered *"3. Moyenne"* (English: *"3. Fair"*) to the WHO version and *"4. Acceptable"* (English: *"4. Fair"*) to the US version of the question, we prefer to remain agnostic as to whether this person is word concordant in French, although s/he would have been word concordant in other languages. Therefore, the group of people for whom this situation holds is treated as a separate group, namely that there is no basis to distinguish whether someone is word concordant. Finally, the individuals that gave completely different answers to both versions of the question are categorized as "discordant"; that is, they do not show any evidence of concordance at all.

[16] Jürges, Avendano, and Mackenbach [2008] did something similar, considering literal and relative concordance, where in their construction the two types of concordance are not mutually exclusive.

A slight difficulty arises in attempting to categorize those individuals whose first response is either *"Excellent"* to the US version (8.6% of those who received the US version first, or 4.3% of the overall sample) or *"Very bad"* to the WHO version (1.2% of those who received the WHO version first, or 0.6% of the overall sample), as they do not have the opportunity to select the same response when subsequently asked the second version of the self-assessed health question. It is possible that responding with the same numerical selection (which is treated in our analysis as being "numerically concordant") in fact reflects a desire to respond with the same exact word but the inability to do so. To the extent that this therefore means that some of the individuals whom we classify as numerically concordant are mis-classified (because they are really word concordant), the results should be biased against finding a difference between the two concordance types.

Based on the definitions above, 55.7% of the total sample is word concordant, 30.6% is numerically concordant, 9.1% of the sample is discordant, and in 4.6% of the cases we had no basis to distinguish between word concordance and discordance (these numbers are shown in Table 5). To identify differences in concordance between subgroups, the sample is split out by various socio-demographic characteristics. The distribution of the proportions of individuals that fall into each of the four categories, by these different socio-demographic characteristics, can be found in Table 7. There is a striking difference in concordance patterns in the Scandinavian countries relative to the European continent. Word concordance is much lower in Denmark and Sweden (31.6% and 32.5%, respectively, compared to the overall word concordance of 55.7%), while numerical concordance is much higher in these countries (50.1% and 53.3%, respectively, compared to 30.6% overall). This suggests that Scandinavians tend to rank their own health more in relation to other people than those in other parts of Europe do, suggesting a more relativistic focus. The fraction of people for whom there was no basis to distinguish due to wording differences is by far the largest in France (18.5%) due to only having two responses of word overlap between the two versions.[17] It is important to keep this in mind when comparing word concordances across countries; that a lower word concordance for some countries may reflect fewer available categories for comparison.

Dividing the sample by gender, it seems that females are statistically significantly more word concordant and less numerically concordant than males (with significance probability less than 1%), although the difference is only about two percentage points for both word and numerical concordance. Dividing the sample into age groups, word concordance seems to be first increasing when people turn older, and then decreasing again from age 65. In contrast, numerical concordance appears to decline with age. Consistent with intuition, the percentage of individuals that is discordant increases from 8.7% at age 50-54 to 9.9% at age 70-75, perhaps indicating greater difficulty as individuals age in remembering what they answered the first time. That the proportion of respondents in the "no basis to distinguish" group increases with age reflects the facts that more of the lack of word overlap occurs in the less healthy categories and that health declines with age.

[17] The maximum number of response categories of word overlap is four; therefore, in the table, there are no individuals from Austria, Germany, Greece, Netherlands, and Spain in this concordance group. For all other countries except France, there are three response categories of word overlap. The sample was also split by language (as opposed to country, because in some countries multiple languages were used and some languages were used in multiple countries of interview). Of the four languages in which there is a possibility of being in the group where there is no basis to distinguish (Danish, French, Italian and Swedish), the proportion is only very large in French (18.6%); it is below 7% for the other three languages.

Looking at different levels of educational attainment, we find that the percentage of discordance is the highest among individuals with the lowest educational level. Word concordance increases when the level of education becomes higher, but then decreases again from level 4 onwards (those beyond having a high-school degree). Numerical concordance is around 8% higher for individuals in the highest education levels (levels 5 and 6, university degree and above) relative to those in the lowest (level 0, pre-primary education). For individuals in the group of "other" education, like special education, discordance is similar to the group with education level 0.

According to individuals' self-reported marital status, numerical concordance is a bit higher for married people (30.9% versus 29.6%, $p = 0.054$); discordance is statistically significantly higher for single people (9.7% against 8.8% for married people; $p = 0.034$). Finally, if we divide the sample according to country of origin, we see that word concordance is about three percentage points higher for individuals that were born in the country of interview (a statistically significant difference, $p = 0.007$). In contrast, there is no statistically significant difference in numerical concordance between those that migrated to the country in which they currently live and those that were born there, suggesting that numerical scales may be preferable when surveying non-native populations.

As we have seen in Table 6 above, some important differences regarding the distribution of responses to the self-assessed health questions seem to arise depending on which version of the question was asked first. It is also the case that concordance patterns vary according to which version was asked first: in Table 8, the sample is divided between those individuals who were asked the WHO version first versus individuals who were asked the US version first; the distributions along the 25 possible response combinations are given. Numerical concordance is statistically significantly higher for the subsample that was asked the WHO version first (35.9%, versus 25.1% for the individuals who got the US version of the question first). In contrast, those who received the US version of the health question first and then were asked the WHO version are more word concordant (58.8% versus 52.7% for those that answered the WHO version first; this difference is also statistically significant). Discordance is four percentage points higher in the group that was asked the US version of the self-assessed health question first (11.1% versus 7.1%). Overall, the distribution of response options differs substantially across the two groups.

IV. Determinants of Differences in Self-Assessed Health Responses

Jürges, Avendano, and Mackenbach [2008] use an ordered probit model to identify determinants of responses to each version of the self-assessed health question. Having documented differences in some individuals' responses to the two self-assessed health questions and that those differences might arise from which question was asked first, we now consider whether observable factors might explain those differences. Table 9 contains results of an ordered probit regression where the dependent variable is the difference between the responses to the two versions of the self-assessed health question. To be more specific, the dependent variable is equal to the response to the US version of the question minus the response to the WHO version of the question, resulting in a possible range of responses between -4 and 4, although not all possible values are equally likely. For example, the difference is equal

to zero for numerically concordant individuals (30.6% of the sample) and one for word concordant individuals (55.7% of the sample) and those for whom there was no basis to distinguish (4.6% of the sample).[18] Therefore the differences are consolidated into four numerical ranges: <0, 0, 1 and >1, where <0 and >1 both correspond to discordant individuals. To allow for the possibility of dependence on the ordering of which version was asked first, a dummy variable equal to one if the US version was asked first and zero otherwise was interacted with the full set of explanatory variables. Under the null hypothesis that the ordering of the two versions of the self-assessed health question does not matter, we expect the coefficients of the variables for the two subgroups to be the same, corresponding to a test that the coefficients on all the interacted variables are jointly equal to zero.

The results of the estimation of this ordered probit model can be found in Table 9, along with marginal effects for each stratum. To give an idea of the interpretation of the variables: the significant coefficient of 0.068 for females means that for women, compared to men, the difference in answers on average will be larger, saying that women tend to represent themselves as less healthy when asked the US version, as compared to the WHO version, than men do. However, this effect seems only to apply for women that were asked the WHO version first; the significant coefficient for females that were asked the US version first is of nearly equal and opposite sign, indicating that there is no significant difference between men and women that were asked the US version first. The marginal effects associated with being female indicate that those who were asked the WHO version first are 1.9 percentage points less likely to be numerically concordant (dependent variable = 0) and 1.6 percentage points more likely to be word concordant (dependent variable = 1) than the rest of the sample. Overall, we see that there are large country effects, and also the health and cognition variables help to explain the difference in the responses to the two self-assessed health questions. The latter suggests that the health and cognition questions that were asked in between the two self-assessed health questions have an influence on how people respond to self-assessed health questions. So, the questions asked in the health section of the interview may cause individuals to adjust their response to the second self-assessed health question, if they, for example, conclude that their actual health is not as bad or as good as they thought it was when answering the first self-assessed health question.

Importantly, the results clearly demonstrate that which question was asked first matters, as a Wald test for the joint significance of all coefficients on the interacted variables gives a χ^2-statistic that is equal to 532.720 ($p < 0.001$). The difference in responses to the two versions depends on which of the two versions an individual was asked first. The next section considers the determinants of the responses to the self-assessed health question, first by including both the first- and second-asked questions in the analysis and then by restricting the analysis to the responses to the first question only, prior to individuals being asked the battery of health-related questions.

[18] That is, in terms of the difference in responses between questions, the "no basis to distinguish" appear similar to those that are clearly word concordant.

A. Determinants of Self-Assessed Health

Having documented that the differences in responses depend on which version of the self-assessed health question individuals were first asked, we now consider the implications for inference regarding health-related behavior. Our approach follows, for example, Hernández-Quevedo, Jones, and Rice [2005] and Meijer, Kapteyn, and Andreyeva [2011], where self-assessed health, a categorical variable denoted by y, is assumed to depend on an individual's underlying, unobserved (latent) general "true" health, y^*. True health is assumed to be of the form

$$y^* = X'\beta + \varepsilon,$$

where X is a matrix of explanatory variables, e.g., different health variables, cognitive variables and socio-demographic controls, β is a vector of parameters and ε is a vector of error terms for all individuals. Then, self-assessed health and true health are assumed to be related in the following way:

$$y = \begin{cases} 1 & \text{if } y^* \leq \tau_1 \\ 2 & \text{if } \tau_1 < y^* \leq \tau_2 \\ 3 & \text{if } \tau_2 < y^* \leq \tau_3 \\ 4 & \text{if } \tau_3 < y^* \leq \tau_4 \\ 5 & \text{if } \tau_4 < y^*, \end{cases}$$

e.g., individuals respond option 2 to the self-assessed health question if their underlying true health is between some threshold parameters τ_1 and τ_2.

If the relationship between possible determinants of self-assessed health responses is invariant to the choice of descriptions attached to those responses, then the coefficients on those determinants should be the same whether the US version or the WHO version of the question is used. To consider this invariance, Jürges, Avendano, and Mackenbach [2008] performed cross-equation tests on the coefficients from two ordered probit models (one with the WHO version as dependent variable, and the other with the US version as dependent variable), without distinguishing which version corresponded to the first- and the second-asked question. They found some statistically significant differences in the coefficients between the two models using slightly different (and fewer) variables than in this paper; from their analysis, the coefficients on age, chronic diseases and countries differ significantly (at the 5% level of significance).[19] Because our list of variables and sample differs from theirs, we first replicate their approach including both self-assessed health questions without distinguishing which version was asked first and similarly find statistically significant differences (at the 5% level of significance) between the two equations in the

[19] As noted above, their analysis was limited to five of the eleven wave 1 countries.

coefficients corresponding to a number of variables: gender, country of interview, chronic diseases, eyesight, and hearing (see columns 5 and 6 of Table 10).[20]

The preceding analysis combined all responses to the WHO question (respectively, US question) together, regardless of whether it was the first or second question asked of the respondent. To consider whether the order in which the questions were asked is important, we re-estimated the ordered probit models by separately focusing solely on the responses to the first question individuals were asked and then similarly on the responses to the second question. The sample is therefore split in each regression into those that answered the WHO question first and those that were first asked the US question; coefficients of the determinants of the responses to these questions are then compared. When considering only the first asked health question in the analysis, only the coefficients on the countries and eyesight are statistically significantly different at the 5% level across the two equations (see columns 1 and 2 of Table 10). In contrast, when the dependent variable is instead the responses from the second-asked health question, many more coefficients are statistically significantly different between the two models: gender, health symptoms, self-assessment of ability to see things at close distance, self-assessed hearing, and orientation, in addition to the difference across countries that was present even with the first-asked question. This increase in significance of many covariates when the self-assessment occurs after the battery of health questions suggests that many differences in the responses to the two versions of the self-assessed health question arise during the course of the survey, rather than being apparent at the beginning. We return to this observation later in the paper.

B. A Model to Describe Concordance Type

Having documented determinants of the response choice for the self-assessed health question, we use a probit model to identify factors that influence the likelihood of word or numerical concordance. The dependent variable is again denoted as y, where, for example, in the model that explains word concordance, $y = 0$ means that someone is not word concordant, whereas $y = 1$ means that an individual is word concordant; a similar model and notation is used to explain numerical concordance. It is again assumed that there is some underlying latent variable, y^*, which denotes the (continuous, but bounded between zero and one) probability of being word (respectively, numerically) concordant. This latent variable depends on various covariates that are captured in X:

$$y^* = X'\beta + \varepsilon,$$

where ε is a vector of error terms assumed to be normally distributed. Then, word (numerical) concordance, y, is assumed to be related to the latent variable as follows:

[20] We do not find significant differences in age but note that our age range is smaller than that used by Jürges, Avendano, and Mackenbach [2008]. At the 10% level of significance, they additionally find a significant difference in the coefficients on education while we do not in our sample. At the 10% level of significance we also find a significant difference in the coefficient on word recall; this variable is not included in Jürges, Avendano, and Mackenbach [2008].

$$y = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{otherwise,} \end{cases}$$

so that, because of the assumption of normality, the probability of being word (numerically) concordant is equal to

$$P[y = 1 \mid X] = \Phi(X'\beta).$$

The estimation uses the full set of socio-demographic, health, and cognitive explanatory variables. Since the assignment to which version of the self-assessed health question was asked first may have an influence on being word or numerically concordant, a full set of variables interacted with a dummy variable (equal to one if the US version was asked first and zero otherwise) is included. In the case that there is no difference between those who received the WHO version first and the ones who received the US version first, the coefficients of the interacted variables should all be zero.

1. Word concordance

The estimates from the model for word concordance are given in Table 11, along with associated marginal effects. Women have a significantly higher probability of being word concordant, compared to men, and married individuals are significantly less likely to be word concordant (6.7 percentage points less likely all else held constant), the latter less-so if they have high net worth. The coefficients on the interaction terms associated with these characteristics are significant and of opposite sign, indicating that the effect is more pronounced among those that were given the WHO version of the self-assessed health question first. There are also large differences across the countries (France is the omitted country), consistent with the fact that only two response options correspond in French. Therefore, most countries have a larger probability of being word concordant, compared to France with the two already-noted exceptions of Denmark and Sweden. Surprisingly, those residing in Italy and Spain in addition have a statistically significantly higher probability of word concordance if asked the US version of the self-assessed health question first, around six percentage points higher than the rest of the sample.

Regarding health, the more symptoms one has, the less likely one is to be word concordant; there is no significant difference between those that were asked the US version first and those that received the WHO version. Chronic diseases statistically significantly lower the probability of being word concordant among the subgroup that is asked the US version of the self-assessed health question first; they do not appear to influence the probability of word concordance among the WHO-first group. In all cases, better eyesight and hearing (relative to the omitted category of *"Good"*) are associated with a statistically significantly lower probability of being word concordant. What may be surprising, however, is that in some cases, having worse eyesight and hearing is also associated with a statistically significantly lower probability of being word concordant. This may reflect response-anchoring; that

respondents are often drawn to the *"Good"* category (note that in both the US and WHO versions, over 40% of the sample selected this response).[21]

Turning to the cognitive assessments, individuals who assess themselves to have *"Excellent"* self-rated writing skills are statistically significantly less likely to be word concordant, if they were given the WHO version of the health question first; the effect of *"Excellent"* self-rated writing skills is not significant if the US version was asked first. Word recall is shown to be related to word concordance; if an individual is only able to recall fewer than four words out of a list of ten directly after the words have been said, s/he is statistically significantly less likely to be word concordant on the self-assessed health questions, regardless of which version of the question was given first. Also, a high numeracy score is associated with a higher likelihood of word concordance, suggesting cognitive complementarity between numeracy and literacy; in contrast, depression is strongly significantly negatively related to word concordance, particularly if one is asked the WHO version of the question first. To the extent that depressed individuals are biased towards pessimism in their outlook, it is possible that this pattern reflects their ability to select the least healthy WHO category initially, for which there is no corresponding US-version category.[22]

Across all of the covariates, many interacted coefficients are not significant, suggesting that the order in which the two versions of the self-assessed health question were asked does not matter in the probability of word concordance, although a Wald test of the joint significance of all the interacted variables is equal to 191.380 with a corresponding p-value that is smaller than 0.001 using a χ^2 distribution. This means that statistically speaking, being asked the US version first does in fact matter for the probability of being word concordant.

2. Numerical concordance

We also estimated coefficients for the probit model for the probability of being numerically concordant, again with the same explanatory variables (Table 12). Females are statistically significantly less likely to be numerically concordant compared to males, although this difference is primarily apparent among those who are given the WHO version of the question first. In contrast to the factors affecting the probability of word concordance, marital status has no significant relationship to the probability of numerical concordance. Not surprisingly, individuals interviewed in the Scandinavian countries (Denmark and Sweden) have a statistically significantly higher probability of being numerically concordant than those in France (the omitted category), even more so those in Denmark who were given the US version of the question first. Those interviewed in Belgium also have a statistically significantly higher probability of being numerically concordant than those in France; this increase does not depend

[21] To consider this possibility, we computed concordances between the US-version of the self-assessed health question and the explanatory variables that also use this scale for eliciting responses. Among those that responded *"Good"* to the self-assessed health question, about half also responded *"Good"* to the questions on eyesight and/or hearing and 40% answered *"Good"* in evaluating their reading and/or writing skills. The number of individuals in our sample that responded *"Good"* to all five variables that use the US version of the descriptors is 1,188, or 5.4% of the sample.

[22] Out of all depressed individuals that were asked the WHO version first, a weighted proportion of 4.6% chose *"Very bad"*, as compared to 0.6% among the non-depressed, corresponding to approximately 112 individuals.

on which version of the question was first asked. There is no evidence of statistically significant influences of either education level or income/wealth levels on the probability of being numerically concordant. Having no chronic diseases and no other symptoms also makes people more likely to be numerically concordant; the healthier one is, the higher the probability of answering the same position on the five-point scale to both self-assessed health questions. As with word concordance, eyesight and hearing seem to be important determinants: better eyesight and hearing increases the probability that one is numerically concordant, in some cases more so when one is first asked the US version of the question. When the US version of the health question is asked first, individuals with excellent reading skills are statistically significantly *more* likely to be numerically concordant and individuals with excellent writing skills are statistically significantly *less* likely to be numerically concordant, than if the WHO version was asked first. Regarding the word recall, those people who could not recall more than four out of ten words directly after the words were said are more likely to be numerically concordant (recall they were less likely to be word concordant in the previous regression), compared to the people who recalled five or more words; this increased probability is statistically significant for those that were given the WHO version of the self-assessed health question first. Perhaps counterintuitively, individuals who first responded to the WHO question and had a bad numeracy score (score 1 out of 5) also have a statistically significantly higher probability of being numerically concordant. Finally, those that score highly on the depression scale have a statistically significantly higher probability of being numerically concordant on self-assessed health; this effect does not appear to depend on which version of the question was asked first.

Although some of the added interactions with the included dummy variable that controls for when the US version of the health question was asked first are significant, most of them have an opposite sign so that the overall effect of the factor is reduced. In many cases, this suggests that being asked the WHO version of the question first results in statistically significantly different probabilities of numerical concordance than when the US version is administered first. The Wald test of joint significance for all interacted variables is equal to 261.980 ($p < 0.001$) indicating that the probability of numerical concordance is also affected by which version of the health question is asked first. This is consistent with what we already observed in the descriptive statistics, that a larger fraction of the sample is numerically concordant if the WHO version was asked first, compared to those that were asked the US version first.

V. Subsequent Health Assessments and Outcomes

Although the finding that there are differences in self-assessed health depending on the wording of the survey question and the order in which the questions are asked is novel and somewhat fascinating, the results are primarily of interest to the extent that such differences influence inference with respect to behavioral outcomes. We therefore investigate whether information from SHARE wave 1 can be used to predict outcomes in wave 2 and if concordance information might improve such predictions.

Self-assessed health in SHARE wave 2

In the second wave of SHARE only the US version of the self-assessed health question was asked.[23] For the sample of 14,768 individuals for which key explanatory variables are present, changes in response between wave 1 and wave 2 are compared; summary statistics are shown in columns 1 to 3 of Table 13. Out of this sample, 48.4% answered exactly the same number to the self-assessed health question in both waves; 20.9% report themselves to be healthier (i.e., gave a lower number) in the second wave and 30.7% indicate that they became less healthy (i.e., gave a higher number) in the second wave. Across countries, those in Greece had the largest persistence in self-assessed health (55.6%) between the two waves, and a greater proportion of those in Sweden had a decline in self-assessed health than in other countries: 40.8% indicated poorer health in the second wave. Another interesting result is that there seems to be a U-shaped relationship between the change in self-assessed health and age. In the middle age group (60 to 64 years), most people reported better health in the second wave and the proportion of people that reported worse health was the smallest in this group, compared to the other age groups; the youngest (50 to 54 years) and oldest (70 to 75 years) age groups reported that they became less healthy compared to the other age groups. Generally between the two waves, there seems to be little variation in changes in self-assessed health along demographic lines.

In the second wave, individuals were also asked the question: *"Compared with your health when we talked with you in {month and year previous interview}, would you say that your health is better now, about the same, or worse?"* The distribution of each of the three answers can be found in columns 4 to 6 of Table 13. Comparing the responses to this question to the change in self-assessed health over the waves on an individual basis, only 48.7% of the individuals gave consistent responses to the two questions (column 7). On aggregate, when asked to assess changes in health in this way, individuals are much less likely to indicate health improvement than one would infer by comparing responses to the five-point self-assessment scale between the two waves. Specifically, over the whole sample, only 5.6% responded *"Better"* to this question, while 20.9% appeared healthier when comparing the wave 2 response to the self-assessed health question to their earlier wave 1 response. There are also differences in response patterns to the two questions by country. Only 1.2% of respondents in Greece indicated that their health had gotten *"Better"*, despite 20.2% selecting a number from the five point scale that was lower (meaning healthier) than they had in wave 1. The country that had the largest number of respondents select *"Better"*, Sweden, had the largest proportion of individuals (40.8%) select a higher number (indicating worse health) using the five-point scale.

Interestingly, in all cases, both overall and with respect to each of the covariates considered, a much higher proportion of individuals (67.0%) reported that their health was *"About the same"* as two years earlier when the question was phrased in that way than actually answered the same number in response to the self-assessed health question using the five point scale (48.4%). This is perhaps not surprising; individuals were not reminded what their wave 1 response was when asked the wave 2 question and it is hard to imagine anyone would remember what they

[23] In wave 2, individuals were also asked to rate their health on a scale of 0 (worst) to 10 (best).

had responded two years prior. The distribution of responses by age for the question asking how health compares to wave 1 is more consistent with intuition: the percentage that answered *"About the same"* is monotonically decreasing with age (from 72.8% at age 50-54 to 59.2% at age 70-75), while *"Worse"* is monotonically increasing with age (from 20.9% in the youngest age group up to 37.4% among the oldest). So it appears that there exists a substantial difference between the inference that would arise when comparing responses to self-assessed health questions between waves versus an analysis of the question in which individuals are actually asked to compare their current health to their health in the previous wave. Such differences likely also affect predictions of associated behavior. The next section considers how self-assessed health changes over time and whether these changes relate to individuals' initial state of health, cognition and concordance, as measured in the first wave.

A. Changes in Self-Assessed Health

Because the WHO version of self-assessed health was not asked in SHARE wave 2, this section focuses on the change in the US version of self-assessed health between wave 1 and wave 2 and the incremental contribution that specific sets of variables make to this change, using the wave 2 sample of 14,768 individuals. Because only 194 individuals have an absolute change that is greater than two (changes in self-assessed health range from -4 to 4), these individuals were grouped with other categories. Therefore, the possible values associated with a change in self-assessed health between waves are <-1, -1, 0, 1 or >1, where a positive change means that individuals assessed themselves to be less healthy in the second wave and a negative change denoting an improved health assessment in wave 2. These five possible outcomes are used as the dependent variable in an ordered probit regression, to find out which variables influence changes in self-assessed health. First, only socio-demographic variables (age, gender, country of birth, marital status, country of interview, education and income) were included as explanatory variables, then changes in all health and cognition variables between the two waves were added (number of chronic diseases and symptoms, self-assessed eyesight and hearing, number of times seen a medical doctor, self-assessed reading and writing skills, word recall, numeracy score, orientation to the date and depression score), and in the third regression the dummy variables to indicate word concordance, numerical concordance and discordance in wave 1 were additionally included. The first two of these regressions were also performed for the four different concordance subgroups: the individuals that were word concordant in wave 1, the ones that were numerically concordant, the ones where we had no basis to distinguish between word and non-concordance, and finally the ones that were discordant in wave 1. The pseudo-R^2s of all these regressions can be found in Table 14. [24]

Although earlier many socio-demographic variables were significant determinants of the responses to the self-assessed health question, they have virtually no explanatory power regarding *changes* in self-assessed health between waves 1 and 2. After controlling for changes in health and cognition between the two waves, however, the pseudo-R^2s rise substantially (from 0.002 to 0.049 in the overall sample). Interestingly, if we split the sample of

[24] In the interest of space, because the regressions contain a large number of covariates and our main interest is on the incremental explanatory power of the groups of variables, the regression results are omitted and only the pseudo-R^2 is reported. Complete sets of results for all regressions are available from the authors on request.

individuals by concordance type and consider the explanatory power of the changes in health and cognition measures, the pseudo-R^2s are more than 33% higher than when the samples are combined (the lowest is 0.065 for the word-concordant subsample). Word concordance, numerical concordance and discordance are similar in terms of the pseudo-R^2 from the ordered probit regressions including both demographic variables and changes in health and cognition measures but for the regression on the sample for whom there was no basis to distinguish in wave 1, the fraction of variation that is accounted for is much higher (the pseudo-R^2 is 0.169).[25] In the overall sample, adding the three dummy variables for concordance to the existing set of over 60 variables (mostly binary) for the whole sample, the pseudo-R^2 increases by over 85% from 0.049 to 0.092. Thus there appears to be evidence that knowing an individual's concordance type in wave 1 provides information on how that individual will respond to the self-assessed health question in wave 2.

In addition, the results from these regressions can be used to generate a predicted distribution of responses to self-assessed health in wave 2. Using the parameters of each probit model, the predicted distribution over the five possible changes in self-assessed health was computed for each individual; each individual is then assigned to the group associated with his/her highest probability among the five changes. By calculating weighted proportions over all individuals in the sample, a predicted distribution of responses can be obtained for all three regressions (first using only socio-demographic variables, then adding changes in health and cognition, and then including measures of concordance in wave 1); actual and predicted distributions can be found in Table 15. Despite the relatively low explanatory power of the variables in the regressions, for all three the actual and predicted distributions are fairly close.[26] Another metric that can be used to evaluate the incremental contribution of the added covariates is the proportion of individuals for whom the predicted change (defined as above, where an individual is assigned to the change given the highest probability among the five possibilities) is equal to the actual change. For the regression model that includes only socio-demographic factors, a weighted percentage of 48.4% of the sample is predicted correctly; this increases to 50.4% if all variables, including concordances, are included, representing a 4% increase (i.e., 236 additional individuals). Nonetheless, only about half of the individuals in the sample are predicted correctly, despite the overall predicted distribution being quite close to the actual distribution.

B. Probability of a major health event

As noted above, the primary reason the differences we have documented may be important to researchers is because failure to account for them may lead to incorrect inference. This section explores this possibility by considering a specific outcome, the probability of a major health event in the two years after the first interview.

[25] This may reflect the smaller sample size associated with the population where there was no basis to distinguish between types of concordance.

[26] The predicted proportions to the five responses do not add up to one for the two most expanded regressions, because in these cases some people were predicted to respond outside of the range (for example, someone who answered *"5. Poor"* in wave 1 and has a predicted change of 1 is therefore predicted to answer a "sixth" response option now). The proportion that falls outside the range is not very large; it is equal to zero for the first model, 0.002 (32 individuals) for the second model and 0.005 (52 individuals) for the third model.

Controlling for individuals' state of health and cognition in the first wave, we consider whether the inclusion of concordance improves prediction of these events.

In order to investigate the effect that concordance information and survey design might have on inference regarding health outcomes of interest, the occurrence of five major health "shocks" in between the two waves of interview are considered: (a) heart attack, (b) stroke/being diagnosed with cerebral vascular disease, (c) being diagnosed with cancer, (d) having suffered a hip fracture, and (e) death. Due to the small sample sizes associated with each of these events, a single dichotomous variable is constructed, equal to one if an individual reports having experienced at least one of these five major health events between the waves, and zero otherwise. Next, a sequence of five probit regressions is estimated, each with this binary variable as dependent variable. In the first regression, only socio-demographic explanatory variables were included in the regression. Next, baseline variables on health and cognition from wave 1 were included. In the third regression both versions of the self-assessed health question in wave 1 were added, while for the fourth regression, a full set of interacted variables (i.e., all included variables interacted with a dummy variable equal to one if the individual was asked the US version of the self-assessed health question first and zero otherwise) was included. Finally, the three dummy variables for concordance in wave 1 were added. These five regressions were performed for the whole sample, as well as estimating the first four using the four concordance subsamples.

Results in terms of pseudo-R^2s can be found in Table 16.[27] Again, explanatory power is improved by dividing the sample into four groups according to concordance in wave 1, although the sample sizes may be too small for the groups that were discordant and where there was no basis to distinguish since by the fourth regression over 120 variables are included; the pseudo-R^2s here are very large compared to the other groups. We focus the discussion, therefore, on the results from the overall sample; results from the subsample regressions are qualitatively similar and if anything more significant. There is a large increase (from 0.044 to 0.074) in pseudo-R^2 when health and cognition information is included in the regression. Additionally including both responses to self-assessed health in wave 1 improves the pseudo-R^2 another 1.6 percentage points. Controlling for whether the US version of the question was asked first increases the pseudo-R^2 by a similar amount. Finally, even after controlling for the self-assessed health questions and which was asked first, concordance in wave 1 has some additional explanatory power (pseudo-R^2 is 0.109).

As with the changes in self-assessed health between wave 1 and wave 2, predictions regarding suffering a major health event were made. In the overall sample, 1,075 out of 15,314 individuals (which include all 15,052 individuals of our wave 1 sample that were re-interviewed in wave 2, plus the 262 individuals that died between wave 1 and wave 2) experienced a major health event between wave 1 and wave 2, which is equal to a weighted proportion of 0.074. Therefore, after predicting the probability of having a major health event for each individual,

[27] As with Table 14, in the interest of space, because the regressions contain a large number of covariates and our main interest is on the incremental explanatory power of the group of variables, the regression results are omitted and only the pseudo-R^2 is reported. Complete sets of results for all regressions are available from the authors on request.

those individuals with the largest predicted probabilities (corresponding to 7.4% of the weighted sample) were assigned to have had a "predicted" major health event. This was done for all five regressions (on the whole sample), from which the proportion that was predicted correctly was computed for each regression (Table 17). When only socio-demographic variables were included in the model, 87.7% were predicted correctly, up to 89.3% when all variables were included, up to and including the dummy variables on concordance, an (unweighted) increase of 203 individuals. In the model with only socio-demographic characteristics as explanatory variables, only 17.5% of the individuals who actually had a major health event were predicted to have one; in the model with all variables included the model correctly predicts 27.8% of those events. Overall, it seems that incorporating responses to self-assessed health questions and concordance information in prediction models may lead to improved inference regarding the probability of experiencing major health events.

Conclusion

This paper has considered how responses to self-assessed health questions are formed and whether there is information in such responses for predicting health behaviors and outcomes. Using a unique feature of the first wave of the longitudinal SHARE dataset, where participants were twice asked to rate their health according to a five point scale, immediately preceding and then following a battery of other health-related questions, each time with a different set of words used to describe the points on the scale (the choice of which set was asked first being determined by random assignment), we find strong differences in responses depending on the sequencing of the two questions. In particular, despite verifying the assertion of random assignment across a variety of potential covariates, we find a statistically significant difference between the average response to each version of the question, depending on whether it was asked prior to the battery of health questions or afterwards.

The results expand on the work of Jürges, Avendano, and Mackenbach [2008] who concluded that despite differences in the response patterns, the two questions reflected the same underlying health characteristics and therefore could be used interchangeably for inference. In contrast, our results document that the two scales are not so interchangeable once one considers which version of the self-assessed question was asked first. Notably, we demonstrate that not only does inference regarding an individual's self-assessed health differ according to the sequencing of the question but that these differences depend on observable characteristics.

The paper also considers whether there are differences in the proportion of individuals that correspond to one of four concordance types depending on the question ordering and finds that:

(1) Although those who are word concordant are more likely to be women, single, living outside Scandinavia, healthier, not depressed, and score highly on word recall and numeracy tests, the sequencing of the question versions does not seem to influence the probability of being word concordant;
(2) In contrast, those who are numerically concordant are more likely to be depressed, living in Scandinavia, and have worse memory than the overall sample, particularly among those who were asked the WHO

version of the self-assessed health question first (e.g., men were more likely to be numerically concordant when asked the WHO version first; there was no difference in the numerical concordance proportions of men and women among those who were first-asked the US version).

While the patterns of differences in both the determinants of responses to the self-assessed health question and the responses themselves are interesting, the question that perhaps matters the most is whether it is possible to adjust the responses or control for their differences in a way that leads to better inference regarding health outcomes of interest. To investigate this further, we also considered the second wave of the SHARE dataset, where although only the US version of the self-assessed health question was asked in this wave, an additional subjective health question was asked that had an embedded reference point (i.e., "relative to 2 years ago", as opposed to the more absolute frame suggested by the five-point scales). Comparing the change in self-assessed health between waves 1 and 2 ("transition differences") to the new reference-based question, we found that in comparison to what the transition differences would indicate, the responses to the new question were:

(1) For more than half of the respondents not consistent with the transition difference responses;
(2) Much more likely to indicate that individuals' health was "about the same";
(3) Only one-fourth as likely to indicate health improvements;
(4) Less likely to indicate health declines (about 10% less likely).

These differences raise interesting questions regarding which type of survey instrument is more appropriate for use in studies of transitions – those that use an absolute scale (without reference to an earlier response) or those that are self-referential.

We then compared a sequence of models designed to explain the observed transitions in self-assessed health between the two waves and found that:

(1) Socio-demographic variables alone do little to explain such transitions;
(2) Adding health and cognition information improves the explanatory power of regressions substantially, even more so when population is first divided into concordance subgroups;
(3) Knowing an individual's concordance group results in a large increase in pseudo-R^2, indicating the importance of taking concordance into account when trying to explain transitions;
(4) There is a close fit between the resulting predicted distributions (across the five responses) and the actual distribution of responses;
(5) Although the proportion of respondents whose wave 2 response was predicted correctly is only roughly 50%, there is a 4 percent increase when health/cognition and concordance information is included in the regression (the latter are responsible for 75% of that increase), as compared to a model that includes socio-demographic variables only.

Finally, we considered the role of concordance information, self-assessed health, and the sequencing of survey questions in predicting the probability of a major health event and found that while both self-assessed health and the ordering in which the two versions of the self-assessed health question were asked matters both in terms of pseudo-R^2 and the proportion of events predicted correctly, concordance information does not seem to have much additional explanatory role in the context of outcomes. This may reflect the overall difficulty in predicting low probability events, as well as the possibility that much of the concordance information is captured via other covariates.

In summary, this paper has demonstrated that information in self-assessed health responses is useful in a variety of different contexts. Our results emphasize the role that framing (as a result of variation with respect to the descriptors used to elicit the response) and question-ordering may have in effecting interpretation of these responses and suggest that it may be necessary to adjust responses to take into account such effects in order to better-interpret self-assessed health responses. Through these adjustments, we may be better able to predict health outcomes of interest.

References

Anstey, Kaarin J., Mary A. Luszcz, Lynne C. Giles, and Gary R. Andrews [2001], "Demographic, Health, Cognitive and Sensory Variables as Predictors of Mortality in Very Old Adults," *Psychology and Aging* 16(1), 3-11.

Baker, David B., Michael S. Wolf, Joseph Feinglass, and Jason A. Thompson [2008], "Health Literacy, Cognitive Abilities, and Mortality Among Elderly Persons," *Journal of General Internal Medicine* 23(6), 723-726.

Bassett, William F., and Robin L. Lumsdaine [1999], "Outlook, Outcomes, and Optimism," unpublished manuscript.

Bassett, William F., and Robin L. Lumsdaine [2000], "Probability Limits: Are Subjective Assessments Adequately Accurate?" *The Journal of Human Resources* 36(2), 327-363.

Börsch-Supan, Axel, Agar Brugiavini, Hendrik Jürges, Johan Mackenbach, Johannes Siegrist, and Guglielmo Weber [2005]. *Health, Ageing and Retirement in Europe: First Results from the Survey of Health, Ageing and Retirement in Europe.* Mannheim Research Institute for the Economics of Aging.

Börsch-Supan, Axel, Agar Brugiavini, Hendrik Jürges, Arie Kapteyn, Johan Mackenbach, Johannes Siegrist, and Guglielmo Weber [2008]. *Health, Ageing and Retirement in Europe (2004-2007): Starting the Longitudinal Dimension.* Mannheim Research Institute for the Economics of Aging.

Costa Jr., Paul T., and Robert R. McCrae [1985], "Hypochondriasis, Neuroticism, and Aging," *American Psychologist* 40(1), 19-28.

Crossley, Thomas F., and Steven Kennedy [2002], "The Reliability of Self-Assessed Health Status," *Journal of Health Economics* 21, 643-658.

Damian, Javier, Ana Ruigomez, Vicente Pastor, and Jose M. Martin-Moreno [1999], "Determinants of Self Assessed Health Among Spanish Older People Living at Home," *Journal of Epidemiology and Community Health* 53, 412-416.

Doorn, Carol van [1999], "A Qualitative Approach to Studying Health Optimism, Realism and Pessimism," *Research on Aging* 21(3), 440-457.

Eriksson, Ingeborg, Anna-Lena Undén, and Stig Elofsson [2001], "Self-Rated Health. Comparisons Between Three Different Measures. Results From a Population Study," *International Journal of Epidemiology* 30, 326-333.

Ford, Jessica, Melanie Spallek, and Annette Dobson [2007], "Self-Rated Health and a Healthy Lifestyle Are the Most Important Predictors of Survival in Elderly Women," *Age and Ageing* 37, 194-200.

Health and Retirement Study [2008], *Sample Sizes and Response Rates (2002 and beyond)*. Downloaded from http://hrsonline.isr.umich.edu/sitedocs/sampleresponse.pdf, accessed April 11, 2011.

Hernández-Quevedo, Cristina, Andrew M. Jones, and Nigel Rice [2005], "Reporting Bias and Heterogeneity in Self-Assessed Health. Evidence From the British Household Panel Survey," *Working paper Health, Econometrics and Data Group (HEDG) of the University of York*.

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

Jürges, Hendrik, Mauricio Avendano, and Johan P. Mackenbach [2008], "Are Different Measures of Self-Rated Health Comparable? An Assessment in Five European Countries," *European Journal of Epidemiology* 23(12), 773-781.

Lee, Yunhwan [2000], "The Predictive Value of Self Assessed General, Physical, and Mental Health on Functional Decline and Mortality in Older Adults," *Journal of Epidemiology and Community Health* 54(2), 123-129.

Lusardi, Annamaria [2008a], "Household Savings Behavior: The Role of Financial Literacy, Information, and Financial Education Programs," NBER Working Paper # 13824.

Lusardi, Annamaria [2008b], "Financial Literacy: An Essential Tool for Informed Consumer Choice?" NBER Working Paper # 14084.

Mannheim Research Institute for the Economics of Aging [2010], *SHARE: Release Guide 2.4.0, Waves 1 & 2*. Downloaded from http://www.share-project.org/, accessed April 28, 2011.

Meijer, Erik, Arie Kapteyn, and Tatiana Andreyeva [2011], "Internationally Comparable Health Indices," *Health Economics* 20, 600-619.

Møller, Lars, Tage S. Kristensen, and Hanne Hollnagel [1996], "Self-Rated Health as a Predictor of Coronary Heart Disease in Copenhagen, Denmark," *Journal of Epidemiology and Community Health* 50, 423-428.

Moum, Torbjørn [1992], "Self-Assessed Health Among Norwegian Adults," *Social Science & Medicine* 35(7), 935-947.

Pinquart, Martin [2001], "Correlates of Subjective Health in Older Adults: A Meta-Analysis," *Psychology and Aging* 16(3), 414-426.

Schwarz, Norbert [1999], "How the Questions Shape the Answers," *American Psychologist* 54(2), 93-105.

Simon, J.G., J.B. de Boer, I.M.A. Joung, H. Bosma, and J.P. Mackenbach [2005], "How is Your Health in General? A Qualitative Study on Self-Assessed Health," *European Journal of Public Health* 15(2), 200-208.

Tversky, Amos, and Daniel Kahneman [1974], "Judgment under Uncertainty: Heuristics and Biases," *Science* 185, 1124-1131.

United Nations Educational, Scientific and Cultural Organization [2006], *ISCED 1997*. Downloaded from http://www.uis.unesco.org/TEMPLATE/pdf/isced/ISCED_A.pdf, accessed April 7, 2011.

Zaller, John, and Stanley Feldman [1992], "A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences," *American Journal of Political Science* 36(3), 579-616.

Zimmer, Zachary, Josefina Natividad, Hui-Sheng Lin, and Napaporn Chayovan [2000], "A Cross-National Examination of the Determinants of Self-Assessed Health," *Journal of Health and Social Behavior* 41, 465-481.

Table 1: Sample selection

	Wave 1	Wave 2
Total SHARE wave 1 (2004)	31,115	
Usable sample last wave		22,131
- Not interviewed in wave 2 (2006)		-7,079
- Israel	-2,598	
- Not between age 50 and 75 at baseline	-5,739	
- No individual weight	-112	-31
- Not answered both health questions	-118	-56
Total sample	**22,548**	**14,965**
Other independent variables		
Demographic		
- No marital status	-1	
- No level of education	-3	
Cognition		
- No number of words recalled (first, second, or both times missing)	-247	-91
- No numeracy score	-44	-32
- Less than 6 out of 12 questions on depression answered	-23	-11
Health		
- No number of times talked to medical doctor last year	-57	-42
- No self-reported eyesight distance (except for blind individuals)	-23	-8
- No number of chronic diseases	-13	-4
- No self-reported eyesight close (except for blind individuals)	-3	-8
- No self-reported hearing	-3	-1
Total sample without missing explanatory variables	**22,131**	**14,768**
Missing health behavior	22,131	
No obese behavior	258	
No drinking behavior	19	
No smoking behavior	1	

Table 2: Descriptive statistics, overall sample and by country

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	Whole sample		Means per country										
	Mean	Std.D.	Austria	Belgium	Denmark	France	Germany	Greece	Italy	Netherl.	Spain	Sweden	Switzerl.
Socio-demographic													
Age	61.876	7.351	61.754	62.000	60.785	61.574	61.872	62.147	62.405	60.926	62.146	61.259	61.139
Female	0.523	0.499	0.529	0.519	0.520	0.531	0.519	0.523	0.530	0.518	0.520	0.510	0.514
Born in country of interview	0.900	0.300	0.913	0.929	0.955	0.844	0.819	0.981	0.985	0.936	0.973	0.906	0.829
Married	0.708	0.455	0.664	0.756	0.677	0.716	0.694	0.730	0.715	0.727	0.719	0.627	0.730
Level of education	2.540	1.516	3.045	2.787	3.381	2.401	3.367	2.162	1.903	2.704	1.623	2.795	2.688
ln(Income)	8.954	2.950	9.329	8.906	9.932	9.350	9.490	8.347	8.199	9.336	7.927	10.023	9.619
ln(Net worth)	10.851	4.179	10.524	11.710	9.293	11.184	10.546	11.256	10.853	10.590	11.290	9.547	11.542
Health													
Health WHO	2.337	0.876	2.233	2.141	2.109	2.246	2.399	2.123	2.491	2.140	2.450	2.091	1.864
Health US	3.027	1.015	2.839	2.824	2.524	2.992	3.116	2.838	3.162	2.818	3.176	2.402	2.552
Number of chronic diseases	1.417	1.369	1.154	1.556	1.434	1.409	1.332	1.315	1.582	1.184	1.578	1.329	0.967
Number of symptoms	1.414	1.542	1.195	1.402	1.336	1.438	1.391	1.184	1.494	1.108	1.626	1.424	0.908
Self-rated eyesight	2.653	0.945	2.315	2.586	2.297	2.518	2.563	2.463	2.927	2.613	2.945	2.261	2.286
Self-rated hearing	2.561	1.018	2.290	2.504	2.371	2.548	2.560	2.194	2.636	2.631	2.742	2.264	2.279
Times to medical doctor	7.004	10.372	5.789	7.483	4.200	6.587	7.228	5.143	8.184	4.323	8.600	2.802	4.587
Cognition													
Reading skills	2.447	1.127	1.970	2.100	1.910	2.168	2.323	2.618	2.752	2.443	3.060	1.623	2.035
Writing skills	2.594	1.165	2.080	2.313	2.076	2.340	2.473	2.754	2.900	2.601	3.164	1.747	2.205
Word recall (first time)	4.862	1.807	5.293	5.058	5.647	4.697	5.558	4.958	4.277	5.296	3.789	5.512	5.589
Word recall (second time)	3.366	1.942	3.670	3.431	4.396	3.303	3.852	3.434	2.771	3.898	2.539	4.237	4.258
Numeracy score (1-5)	3.317	1.135	3.740	3.371	3.579	3.241	3.712	3.483	2.976	3.670	2.607	3.731	3.825
Orientation date (0-4)	3.848	0.467	3.863	3.819	3.861	3.853	3.880	3.917	3.848	3.832	3.748	3.905	3.853
Depression score (0-12)	2.328	2.261	1.821	2.275	1.786	2.712	1.861	2.009	2.727	1.970	2.797	1.935	1.810
Number of observations	22,131		1,513	2,965	1,275	2,294	2,465	2,165	2,131	2,363	1,796	2,411	753

Note: Weighted means and standard deviations of the key variables used in this paper over the used wave 1 sample of 22,131 individuals; also weighted means per country.

Table 3: Omitted categories in regressions

Variable	Omitted category
Socio-demographic	
Gender	Male
Country of origin	Not born in country of interview
Marital status	Unmarried
Country	France
Education level	Other education
Health	
Chronic diseases	2 chronic diseases
Symptoms	2 symptoms
Eyesight	*"3. Good"*
Hearing	*"3. Good"*
Times to medical doctor	0-5 times
Cognition	
Reading skills	*"3. Good"*
Writing skills	*"3. Good"*
Word recall	5 words recalled
Numeracy score (1-5)	3 (in the middle of *Good* and *Bad*)
Orientation date (0-4)	4 (*Good*)
Depression score (0-12)	0-3 (not depressed)

Note: In all regressions, information on the above variables is included in the form of dummy variables; these are the omitted categories per variable.

Table 4: Self-assessed health in wave 1

	(1)	(2)	(3)	(4)
Subsample	Mean WHO	Mean US	Difference	P-value vs. 0.690
Total	2.337	3.027	0.690	
Country				
Austria	2.233	2.839	0.607	0.016**
Belgium	2.141	2.824	0.682	0.755
Denmark	2.109	2.524	0.415	0.000***
France	2.246	2.992	0.746	0.040**
Germany	2.399	3.116	0.717	0.291
Greece	2.123	2.838	0.715	0.370
Italy	2.491	3.162	0.670	0.498
Netherlands	2.140	2.818	0.678	0.656
Spain	2.450	3.176	0.725	0.259
Sweden	2.091	2.402	0.311	0.000***
Switzerland	1.864	2.552	0.689	0.981
Gender				
Male	2.287	2.950	0.664	0.052*
Female	2.383	3.096	0.713	0.051*
Test for equality of means	*0.000*	*0.000*	*0.000*	
Age group				
Age 50-54	2.093	2.730	0.637	0.005***
Age 55-59	2.226	2.886	0.660	0.117
Age 60-64	2.348	3.060	0.713	0.238
Age 65-69	2.469	3.177	0.708	0.368
Age 70-75	2.632	3.381	0.749	0.005***
Level of education				
Education level 0	2.798	3.539	0.740	0.234
Education level 1	2.523	3.236	0.713	0.180
Education level 2	2.351	3.032	0.681	0.675
Education level 3	2.295	2.999	0.705	0.353
Education level 4	2.053	2.710	0.657	0.581
Education level 5 and 6	2.013	2.640	0.628	0.001***
Other education	2.400	3.009	0.609	0.444
Marital status				
Married	2.310	2.993	0.683	0.532
Single	2.403	3.108	0.705	0.408
Test for equality of means	*0.000*	*0.000*	*0.032*	
Born in country of interview				
Yes	2.322	3.011	0.689	0.923
No	2.467	3.165	0.698	0.806
Test for equality of means	*0.000*	*0.000*	*0.559*	

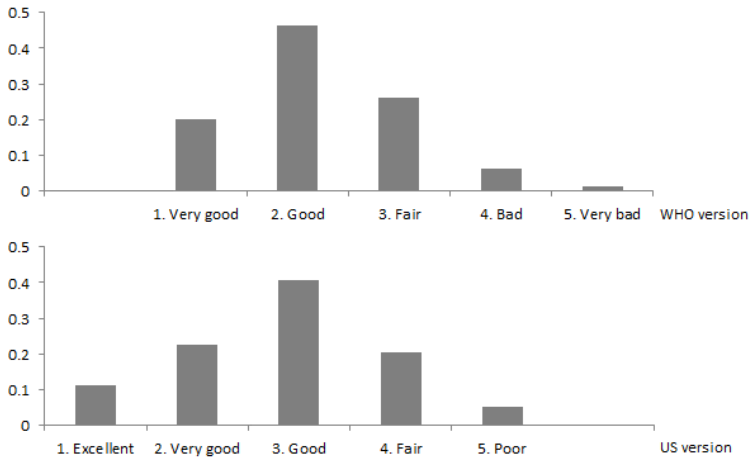
Note: This table shows average responses to both (1) the WHO version and (2) the US version of the self-assessed health question; column (3) shows the difference between the two averages and column (4) contains the p-value corresponding to a test of whether the difference is statistically significant from the overall difference of 0.690. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Distribution of combinations of responses, total sample

	US version					
WHO version	*1. Excellent*	*2. Very good*	*3.Good*	*4. Fair*	*5. Poor*	Total
1. Very good	0.062	0.072	0.017	0.000	0.000	0.151
2. Good	0.018	0.111	0.309	0.028	0.001	0.466
3. Fair	0.001	0.007	0.087	0.183	0.013	0.292
4. Bad	0.000	0.000	0.003	0.032	0.040	0.076
5. Very bad	0.000	0.000	0.000	0.001	0.013	0.015
Total	0.082	0.189	0.417	0.245	0.067	1.000
Word concordance		0.557				
Numerical concordance		0.306				
Discordance		0.091				
No basis to distinguish		0.046				

Note: By adding up the proportions along the diagonal, numerical concordance is obtained. By adding up the proportions where the wording is the same for both responses and subtracting the "No basis to distinguish" group, word concordance is obtained. In French, "No basis to distinguish" is the sum of the cells (*3.Fair*,*4.Fair*) and (*4.Bad*,*5.Poor*); in Italian it is equal to the cell (*4.Bad*,*5.Poor*); in Danish (*1.Very good*,*2.Very good*); in Swedish (*3.Fair*,*4.Fair*). Discordance is obtained by adding up all remaining proportions.

Figure 1: Distribution of responses for the two self-assessed health questions



Note: This figure shows the distributions of responses to the two versions of the self-assessed health question. The vertical axis denotes the proportion of individuals.

Table 6: Descriptive statistics, by wording scale

	(1)	(2)	(3)	(4)	(5)	(6)
	WHO first		US first		Test on	Test on
	Mean	Std.Dev.	Mean	Std.Dev.	means (p)	vars (p)
Socio-demographic						
Age	61.688	7.206	61.806	7.190	0.224	0.814
Female	0.537	0.499	0.532	0.499	0.392	0.933
Born in country of interview	0.918	0.275	0.921	0.270	0.357	0.045
Married	0.758	0.428	0.756	0.430	0.650	0.701
Level of education	2.612	1.515	2.579	1.508	0.109	0.646
ln(Income)	9.002	2.946	8.982	2.959	0.622	0.652
ln(Net worth)	10.942	4.237	10.955	4.221	0.824	0.698
Health						
Number of chronic diseases	1.378	1.343	1.389	1.338	0.525	0.724
Number of symptoms	1.347	1.501	1.333	1.473	0.470	0.048
Self-rated eyesight	2.546	0.945	2.537	0.958	0.494	0.157
Self-rated hearing	2.482	1.022	2.468	1.019	0.320	0.727
Times to medical doctor	5.941	8.998	6.087	9.538	0.242	0.000
Cognition						
Reading skills	2.276	1.104	2.294	1.111	0.225	0.485
Writing skills	2.430	1.164	2.434	1.154	0.805	0.372
Word recall (first time)	5.047	1.752	5.040	1.733	0.754	0.268
Word recall (second time)	3.568	1.965	3.585	1.933	0.498	0.087
Numeracy score (1-5)	3.425	1.101	3.421	1.095	0.744	0.518
Orientation date (0-4)	3.855	0.428	3.850	0.449	0.464	0.000
Depression score (0-12)	2.161	2.150	2.166	2.166	0.865	0.435
Self-assessed health						
Health WHO	2.276	0.896	2.169	0.861	0.000	0.000
Health US	2.814	1.005	2.906	1.060	0.000	0.000
Number of observations	11,051		11,080			

Note: Unweighted means and standard deviations separately for the individuals that were asked the WHO version first, and the ones that were asked the US version first. Columns (5) and (6) provide p-values of tests on equality of means and variances, respectively.

Table 7: Variations in concordance, by country and demographic characteristics

	(1)	(2)	(3)	(4)	(5)
	Sample size	Word conc.	Numerical conc.	Discordant	No basis
Total	22,131	0.557	0.306	0.091	0.046
Country					
Austria	1,513	0.553	0.376	0.071	–
Belgium	2,965	0.508	0.324	0.099	0.069
Denmark	1,275	0.316	0.501	0.115	0.069
France	2,294	0.424	0.287	0.104	0.185
Germany	2,465	0.649	0.283	0.067	–
Greece	2,165	0.624	0.303	0.072	–
Italy	2,131	0.539	0.312	0.101	0.048
Netherlands	2,363	0.603	0.298	0.099	–
Spain	1,796	0.621	0.274	0.105	–
Sweden	2,411	0.325	0.533	0.115	0.027
Switzerland	753	0.581	0.316	0.079	0.024
Gender					
Male	10,301	0.546	0.317	0.094	0.044
Female	11,830	0.568	0.296	0.088	0.048
Tests for equality of means		*0.001*	*0.001*	*0.150*	*0.089*
Age group					
Age 50-54	4,880	0.534	0.345	0.087	0.034
Age 55-59	5,009	0.552	0.323	0.089	0.036
Age 60-64	4,490	0.583	0.289	0.090	0.039
Age 65-69	3,972	0.568	0.290	0.091	0.051
Age 70-75	3,780	0.555	0.269	0.099	0.077
Level of education					
Education level 0	1,006	0.439	0.273	0.122	0.166
Education level 1	5,533	0.538	0.292	0.105	0.064
Education level 2	4,125	0.578	0.302	0.093	0.027
Education level 3	6,433	0.595	0.293	0.079	0.033
Education level 4	534	0.564	0.350	0.075	0.011
Education level 5 and 6	4,335	0.541	0.355	0.080	0.024
Other education	165	0.437	0.374	0.115	0.074
Marital status					
Married	16,753	0.559	0.309	0.088	0.044
Single	5,378	0.555	0.296	0.097	0.052
Tests for equality of means		*0.610*	*0.054*	*0.034*	*0.017*
Born in country of interview					
Yes	20,346	0.561	0.304	0.092	0.043
No	1,785	0.530	0.319	0.079	0.071
Tests for equality of means		*0.007*	*0.163*	*0.041*	*0.000*

Note: In countries where the wording of four of the five response options overlaps, there are, by construction, no individuals who fall in the "No basis to distinguish" group.

Table 8: Distribution of combinations of responses, depending on which version was asked first

a. WHO version asked first

	US version					
WHO version	*1. Excellent*	*2. Very good*	*3.Good*	*4. Fair*	*5. Poor*	Total
1. Very good	0.059	0.069	0.014	0.001	0.000	0.144
2. Good	0.016	0.127	0.291	0.013	0.000	0.447
3. Fair	0.002	0.011	0.119	0.175	0.005	0.312
4. Bad	0.000	0.000	0.006	0.041	0.035	0.082
5. Very bad	0.000	0.000	0.000	0.003	0.012	0.016
Total	0.078	0.207	0.431	0.232	0.052	1.000

Word concordance	0.527
Numerical concordance	0.359
Discordance	0.071
No basis to distinguish	0.043

b. US version asked first

	US version					
WHO version	*1. Excellent*	*2. Very good*	*3.Good*	*4. Fair*	*5. Poor*	Total
1. Very good	0.065	0.074	0.019	0.000	0.000	0.159
2. Good	0.020	0.094	0.328	0.043	0.002	0.487
3. Fair	0.001	0.003	0.055	0.191	0.022	0.271
4. Bad	0.000	0.000	0.001	0.023	0.045	0.069
5. Very bad	0.000	0.000	0.000	0.000	0.014	0.014
Total	0.086	0.171	0.402	0.258	0.083	1.000

Word concordance	0.588
Numerical concordance	0.251
Discordance	0.111
No basis to distinguish	0.049

Note: The distribution of responses to the self-assessed health questions is shown separately for the sample that was asked the WHO version first and the sample that was asked the US version first.

Table 9: Probit; dependent variable = Health change US-WHO (1 of 4)

	(1)	(2)	(3)	(4)	(5)	(6)
Covariate	Coef.	Std.Err.	Marg. <0	Marg. 0	Marg. 1	Marg. >1
Age	0.001	0.002	-0.000	-0.000	0.000	0.000
Female	0.068***	0.025	-0.004	-0.019	0.016	0.008
Born in country of interview	-0.018	0.038	0.001	0.005	-0.004	-0.002
Married	-0.082	0.062	0.005	0.023	-0.018	-0.009
Austria	-0.135*	0.076	0.010	0.037	-0.033	-0.014
Belgium	-0.215***	0.067	0.017	0.059	-0.055	-0.021
Denmark	-0.486***	0.087	0.047	0.129	-0.138	-0.038
Germany	-0.047	0.038	0.003	0.013	-0.011	-0.005
Greece	0.004	0.066	-0.000	-0.001	0.001	0.000
Italy	-0.215***	0.039	0.016	0.059	-0.053	-0.022
Netherlands	-0.051	0.059	0.003	0.014	-0.012	-0.005
Spain	-0.105**	0.043	0.007	0.029	-0.025	-0.011
Sweden	-0.701***	0.069	0.079	0.177	-0.209	-0.047
Switzerland	-0.021	0.079	0.001	0.006	-0.005	-0.002
Education level 0	0.042	0.143	-0.003	-0.011	0.009	0.005
Education level 1	0.197	0.137	-0.012	-0.054	0.042	0.024
Education level 2	0.226*	0.137	-0.013	-0.061	0.046	0.028
Education level 3-4	0.179	0.136	-0.011	-0.049	0.039	0.021
Education level 5-6	0.216	0.137	-0.013	-0.058	0.044	0.027
ln(Income)	-0.002	0.004	0.000	0.000	-0.000	-0.000
ln(HH net worth)	-0.003	0.004	0.000	0.001	-0.001	-0.000
ln(HH worth)*Married	0.003	0.005	-0.000	-0.001	0.001	0.000
Chronic diseases: 0	-0.054	0.036	0.004	0.015	-0.013	-0.006
Chronic diseases: 1	-0.027	0.032	0.002	0.007	-0.006	-0.003
Chronic diseases: 3 or more	-0.044	0.037	0.003	0.012	-0.010	-0.005
Symptoms: 0	-0.094***	0.035	0.006	0.026	-0.022	-0.010
Symptoms: 1	-0.053	0.034	0.004	0.014	-0.012	-0.006
Symptoms: 3 or more	-0.111***	0.039	0.008	0.030	-0.027	-0.012
Eyesight 1 (Excellent)	-0.194***	0.050	0.014	0.054	-0.049	-0.019
Eyesight 2 (Very good)	-0.075**	0.035	0.005	0.021	-0.018	-0.008
Eyesight 4 (Fair)	0.107***	0.041	-0.007	-0.029	0.023	0.013
Eyesight 5 (Poor/blind)	0.085	0.092	-0.005	-0.023	0.018	0.010
Eyesight distance 1 (Excellent)	-0.246***	0.044	0.018	0.068	-0.062	-0.024
Eyesight distance 2 (Very good)	-0.146***	0.034	0.010	0.040	-0.035	-0.016
Eyesight distance 4 (Fair)	-0.101**	0.049	0.007	0.028	-0.024	-0.011
Eyesight distance 5 (Poor/blind)	-0.080	0.078	0.006	0.022	-0.019	-0.008
Eyesight close 1 (Excellent)	0.027	0.050	-0.002	-0.007	0.006	0.003
Eyesight close 2 (Very good)	-0.009	0.035	0.001	0.002	-0.002	-0.001
Eyesight close 4 (Fair)	-0.011	0.036	0.001	0.003	-0.003	-0.001
Eyesight close 5 (Poor/blind)	0.118**	0.046	-0.007	-0.032	0.025	0.014

To be continued on next page

Table 9: Probit; dependent variable = Health change US-WHO (2 of 4)

	(1)	(2)	(3)	(4)	(5)	(6)
Covariate	Coef.	Std.Err.	Marg. <0	Marg. 0	Marg. 1	Marg. >1
Hearing 1 (Excellent)	-0.352***	0.034	0.028	0.097	-0.092	-0.033
Hearing 2 (Very good)	-0.167***	0.029	0.012	0.046	-0.040	-0.017
Hearing 4 (Fair)	0.042	0.035	-0.003	-0.011	0.009	0.005
Hearing 5 (Poor)	0.212***	0.071	-0.012	-0.057	0.042	0.027
Medical doctor: 6 or more times	-0.008	0.026	0.001	0.002	-0.002	-0.001
Reading skills 1 (Excellent)	-0.060	0.055	0.004	0.017	-0.014	-0.007
Reading skills 2 (Very good)	-0.058	0.041	0.004	0.016	-0.013	-0.006
Reading skills 4 (Fair)	-0.145***	0.051	0.011	0.040	-0.035	-0.015
Reading skills 5 (Poor/DK/RF)	-0.113	0.085	0.008	0.031	-0.028	-0.012
Writing skills 1 (Excellent)	-0.111**	0.056	0.008	0.031	-0.027	-0.012
Writing skills 2 (Very good)	0.007	0.042	-0.000	-0.002	0.002	0.001
Writing skills 4 (Fair)	0.023	0.046	-0.001	-0.006	0.005	0.003
Writing skills 5 (Poor/DK/RF)	0.115	0.074	-0.007	-0.031	0.025	0.014
Word recall - first time: 0-4	-0.055*	0.030	0.004	0.015	-0.013	-0.006
Word recall - first time: 6-10	0.095***	0.032	-0.006	-0.026	0.021	0.011
Word recall - second time: 0-4	0.044	0.035	-0.003	-0.012	0.010	0.005
Word recall - second time: 6-10	-0.018	0.042	0.001	0.005	-0.004	-0.002
Numeracy score 1 (Bad)	-0.112**	0.050	0.008	0.031	-0.027	-0.012
Numeracy score 2	0.019	0.034	-0.001	-0.005	0.004	0.002
Numeracy score 4	0.042	0.030	-0.003	-0.012	0.010	0.005
Numeracy score 5 (Good)	0.001	0.036	-0.000	-0.000	0.000	0.000
Orientation date 0-3 (Bad)	0.076**	0.035	-0.005	-0.021	0.017	0.009
Depression scale >3 (depressed)	-0.004	0.029	0.000	0.001	-0.001	-0.000

To be continued on next page

Table 9: Probit; dependent variable = Health change US-WHO (3 of 4)

	(1)	(2)	(3)	(4)	(5)	(6)
Covariate	Coef.	Std.Err.	Marg. <0	Marg. 0	Marg. 1	Marg. >1
Age * US first	-0.001	0.002	0.000	0.000	-0.000	-0.000
Female * US first	-0.067*	0.036	0.005	0.018	-0.015	-0.007
Born in country of interview * US first	0.059	0.054	-0.004	-0.016	0.014	0.007
Married * US first	0.223**	0.087	-0.014	-0.061	0.049	0.026
Austria * US first	-0.184*	0.108	0.014	0.050	-0.047	-0.018
Belgium * US first	0.115	0.095	-0.007	-0.031	0.024	0.014
Denmark * US first	-0.127	0.123	0.009	0.035	-0.031	-0.013
Germany * US first	-0.140***	0.054	0.010	0.038	-0.034	-0.015
Greece * US first	-0.066	0.093	0.005	0.018	-0.016	-0.007
Italy * US first	0.057	0.056	-0.004	-0.016	0.013	0.006
Netherlands * US first	-0.190**	0.083	0.015	0.052	-0.048	-0.018
Spain * US first	0.001	0.062	-0.000	-0.000	0.000	0.000
Sweden * US first	0.024	0.097	-0.002	-0.007	0.006	0.003
Switzerland * US first	-0.033	0.113	0.002	0.009	-0.008	-0.004
Education level 0 * US first	0.120	0.213	-0.007	-0.033	0.026	0.014
Education level 1 * US first	-0.135	0.205	0.010	0.037	-0.033	-0.014
Education level 2 * US first	-0.260	0.205	0.021	0.070	-0.066	-0.025
Education level 3-4 * US first	-0.072	0.204	0.005	0.020	-0.017	-0.008
Education level 5-6 * US first	-0.238	0.206	0.019	0.065	-0.061	-0.023
ln(Income) * US first	-0.004	0.006	0.000	0.001	-0.001	-0.000
ln(HH net worth) * US first	0.016***	0.006	-0.001	-0.004	0.004	0.002
ln(HH worth)*Married * US first	-0.019**	0.008	0.001	0.005	-0.004	-0.002
Chronic diseases: 0 * US first	-0.350***	0.051	0.029	0.095	-0.091	-0.032
Chronic diseases: 1 * US first	-0.145***	0.046	0.010	0.039	-0.035	-0.015
Chronic diseases: 3 or more * US first	0.106**	0.053	-0.006	-0.029	0.023	0.012
Symptoms: 0 * US first	-0.151***	0.051	0.011	0.042	-0.037	-0.016
Symptoms: 1 * US first	-0.020	0.048	0.001	0.005	-0.005	-0.002
Symptoms: 3 or more * US first	0.060	0.056	-0.004	-0.016	0.013	0.007
Eyesight 1 (Excellent) * US first	0.078	0.072	-0.005	-0.021	0.017	0.009
Eyesight 2 (Very good) * US first	-0.002	0.051	0.000	0.001	-0.001	-0.000
Eyesight 4 (Fair) * US first	0.035	0.059	-0.002	-0.010	0.008	0.004
Eyesight 5 (Poor/blind) * US first	-0.357***	0.130	0.032	0.096	-0.097	-0.031
Eyesight distance 1 (Excellent) * US first	0.273***	0.063	-0.015	-0.073	0.053	0.035
Eyesight distance 2 (Very good) * US first	0.205***	0.049	-0.012	-0.056	0.042	0.025
Eyesight distance 4 (Fair) * US first	-0.085	0.069	0.006	0.023	-0.020	-0.009
Eyesight distance 5 (Poor/blind) * US first	0.105	0.111	-0.006	-0.029	0.022	0.013
Eyesight close 1 (Excellent) * US first	-0.060	0.071	0.004	0.016	-0.014	-0.006
Eyesight close 2 (Very good) * US first	-0.086*	0.051	0.006	0.023	-0.020	-0.009
Eyesight close 4 (Fair) * US first	-0.088*	0.051	0.006	0.024	-0.021	-0.009
Eyesight close 5 (Poor/blind) * US first	-0.127*	0.066	0.009	0.035	-0.031	-0.013

To be continued on next page

Table 9: Probit; dependent variable = Health change US-WHO (4 of 4)

	(1)	(2)	(3)	(4)	(5)	(6)
Covariate	Coef.	Std.Err.	Marg. <0	Marg. 0	Marg. 1	Marg. >1
Hearing 1 (Excellent) * US first	0.340***	0.049	-0.018	-0.090	0.062	0.046
Hearing 2 (Very good) * US first	0.143***	0.042	-0.009	-0.039	0.031	0.017
Hearing 4 (Fair) * US first	-0.062	0.050	0.004	0.017	-0.015	-0.007
Hearing 5 (Poor) * US first	-0.287***	0.103	0.024	0.078	-0.076	-0.026
Medical doctor: 6 or more times * US first	0.145***	0.037	-0.009	-0.040	0.032	0.017
Reading skills 1 (Excellent) * US first	-0.146*	0.078	0.011	0.040	-0.036	-0.015
Reading skills 2 (Very good) * US first	0.066	0.059	-0.004	-0.018	0.015	0.008
Reading skills 4 (Fair) * US first	0.056	0.071	-0.004	-0.015	0.012	0.006
Reading skills 5 (Poor/DK/RF) * US first	-0.093	0.120	0.007	0.025	-0.022	-0.010
Writing skills 1 (Excellent) * US first	0.177**	0.081	-0.010	-0.048	0.037	0.022
Writing skills 2 (Very good) * US first	-0.094	0.060	0.007	0.026	-0.022	-0.010
Writing skills 4 (Fair) * US first	-0.042	0.064	0.003	0.011	-0.010	-0.004
Writing skills 5 (Poor/DK/RF) * US first	-0.272***	0.105	0.022	0.074	-0.071	-0.025
Word recall - first time: 0-4 * US first	0.037	0.043	-0.002	-0.010	0.008	0.004
Word recall - first time: 6-10 * US first	-0.145***	0.046	0.010	0.040	-0.035	-0.015
Word recall - second time: 0-4 * US first	-0.056	0.050	0.004	0.015	-0.013	-0.006
Word recall - second time: 6-10 * US first	0.106*	0.060	-0.006	-0.029	0.023	0.013
Numeracy score 1 (Bad) * US first	0.102	0.072	-0.006	-0.028	0.022	0.012
Numeracy score 2 * US first	0.041	0.049	-0.003	-0.011	0.009	0.005
Numeracy score 4 * US first	-0.005	0.043	0.000	0.001	-0.001	-0.001
Numeracy score 5 (Good) * US first	-0.064	0.052	0.004	0.018	-0.015	-0.007
Orientation date 0-3 (Bad) * US first	-0.064	0.049	0.004	0.017	-0.015	-0.007
Depression scale >3 (depressed) * US first	-0.007	0.041	0.000	0.002	-0.002	-0.001
Constant * US first	0.481*	0.275	-0.031	-0.133	0.111	0.053
τ_1	-2.045***	0.188				
τ_2	-0.497***	0.188				
τ_3	1.622***	0.188				
Number of observations	22,131					
Log-likelihood	-19,680.273					
Pseudo R^2	0.057					
Wald test on interacted variables (χ^2)	532.720***					

Note: *** p<0.01, ** p<0.05, * p<0.1.

Table 10: Cross-equation tests for different probit models (WHO versus US)

	(1)	(2)	(3)	(4)	(5)	(6)
	First health question		Second health question		Both health questions	
Covariates	χ^2	P-value	χ^2	P-value	χ^2	P-value
Age	0.765	0.382	0.795	0.373	0.082	0.775
Gender	0.051	0.820	4.496	0.034**	4.650	0.031**
Born in country of interview	0.994	0.319	0.627	0.428	0.146	0.702
Marital status	0.002	0.967	0.006	0.938	0.007	0.932
Countries	88.151	0.000***	88.398	0.000***	413.731	0.000***
Education	5.724	0.334	8.080	0.152	5.250	0.386
Income/wealth	1.362	0.715	1.153	0.764	1.623	0.654
Chronic diseases	6.419	0.093*	0.710	0.871	8.658	0.034**
Symptoms	5.472	0.140	10.250	0.017**	1.015	0.798
Eyesight	14.020	0.007***	4.027	0.402	9.878	0.043**
Eyesight distance	4.009	0.405	7.231	0.124	6.571	0.160
Eyesight close	2.823	0.588	9.657	0.047**	7.554	0.109
Hearing	5.710	0.222	11.141	0.025**	14.010	0.007***
Medical doctor	2.657	0.103	1.288	0.257	0.761	0.383
Reading skills	4.459	0.347	0.700	0.951	7.116	0.130
Writing skills	5.513	0.239	1.504	0.826	1.778	0.776
Word recall	5.825	0.213	7.005	0.136	8.749	0.068*
Numeracy score	4.955	0.292	3.265	0.515	6.471	0.167
Orientation date	0.218	0.641	3.891	0.049**	2.033	0.154
Depression score	0.173	0.677	1.094	0.296	0.251	0.616

Note: This table includes cross-equation tests on coefficients from probit models with the WHO and US versions of the self-assessed health question as dependent variables; in columns (1) and (2) only the first-asked question is considered, in columns (3) and (4) only the second-asked question and in columns (5) and (6) both asked health questions are considered.
*** p<0.01, ** p<0.05, * p<0.1.

Table 11: Probit; dependent variable = P(word concordance)

	(1)	(2)	(3)	(4)	(5)	(6)
	Variables whole sample			Interacted variables US first		
Covariate	Coef.	Std.Err.	Marg. 1	Coef.	Std.Err.	Marg. 1
Age	0.002	0.002	0.001	-0.002	0.003	-0.001
Female	0.140***	0.028	0.051	-0.093**	0.040	-0.034
Born in country of interview	-0.009	0.043	-0.003	0.117*	0.061	0.043
Married	-0.184***	0.071	-0.067	0.215**	0.099	0.079
Austria	0.271***	0.085	0.097	-0.018	0.121	-0.007
Belgium	0.070	0.075	0.026	0.143	0.105	0.052
Denmark	-0.351***	0.105	-0.130	0.021	0.145	0.008
Germany	0.390***	0.042	0.142	0.097	0.060	0.036
Greece	0.447***	0.073	0.156	0.106	0.104	0.039
Italy	0.176***	0.044	0.064	0.170***	0.062	0.062
Netherlands	0.336***	0.066	0.119	0.000	0.093	0.000
Spain	0.422***	0.049	0.149	0.153**	0.069	0.056
Sweden	-0.449***	0.083	-0.167	0.315***	0.114	0.111
Switzerland	0.280***	0.088	0.100	0.117	0.125	0.042
Education level 0	0.051	0.164	0.019	0.021	0.242	0.008
Education level 1	0.220	0.157	0.079	-0.074	0.233	-0.027
Education level 2	0.287*	0.157	0.103	-0.182	0.233	-0.067
Education level 3-4	0.257*	0.156	0.094	-0.013	0.231	-0.005
Education level 5-6	0.252	0.158	0.091	-0.105	0.234	-0.039
ln(Income)	0.001	0.005	0.000	-0.006	0.007	-0.002
ln(HH net worth)	-0.002	0.005	-0.001	0.003	0.007	0.001
ln(HH worth)*Married	0.013**	0.006	0.005	-0.012	0.009	-0.004
Chronic diseases: 0	-0.035	0.040	-0.013	-0.268***	0.057	-0.099
Chronic diseases: 1	0.018	0.036	0.007	-0.165***	0.051	-0.061
Chronic diseases: 3 or more	0.026	0.042	0.010	-0.131**	0.059	-0.048
Symptoms: 0	-0.059	0.040	-0.022	-0.040	0.057	-0.015
Symptoms: 1	0.084**	0.038	0.031	-0.060	0.054	-0.022
Symptoms: 3 or more	-0.104**	0.044	-0.038	-0.024	0.062	-0.009
Eyesight 1 (Excellent)	-0.123**	0.057	-0.045	0.009	0.081	0.003
Eyesight 2 (Very good)	-0.081**	0.039	-0.030	0.039	0.057	0.014
Eyesight 4 (Fair)	0.002	0.046	0.001	0.014	0.066	0.005
Eyesight 5 (Poor/blind)	-0.389***	0.104	-0.144	0.262*	0.146	0.093
Eyesight distance 1 (Excellent)	-0.222***	0.049	-0.083	0.151**	0.071	0.055
Eyesight distance 2 (Very good)	-0.113***	0.038	-0.042	0.071	0.055	0.026
Eyesight distance 4 (Fair)	-0.092*	0.054	-0.034	-0.032	0.077	-0.012
Eyesight distance 5 (Poor/blind)	-0.104	0.088	-0.039	-0.131	0.123	-0.049
Eyesight close 1 (Excellent)	-0.074	0.057	-0.027	0.091	0.080	0.033
Eyesight close 2 (Very good)	-0.083**	0.040	-0.030	0.092	0.057	0.034
Eyesight close 4 (Fair)	-0.022	0.041	-0.008	-0.135**	0.057	-0.050
Eyesight close 5 (Poor/blind)	0.144***	0.052	0.052	-0.099	0.074	-0.037

To be continued on next page

Table 11: Probit; dependent variable = P(word concordance)

	(1)	(2)	(3)	(4)	(5)	(6)
	Variables whole sample			Interacted variables US first		
Covariate	Coef.	Std.Err.	Marg. 1	Coef.	Std.Err.	Marg. 1
Hearing 1 (Excellent)	-0.363***	0.039	-0.135	0.216***	0.055	0.078
Hearing 2 (Very good)	-0.199***	0.033	-0.073	0.115**	0.047	0.042
Hearing 4 (Fair)	-0.007	0.040	-0.003	-0.015	0.056	-0.005
Hearing 5 (Poor)	-0.229***	0.079	-0.085	0.063	0.114	0.023
Medical doctor: 6 or more times	-0.056*	0.029	-0.021	-0.007	0.041	-0.003
Reading skills 1 (Excellent)	0.057	0.062	0.021	-0.292***	0.088	-0.108
Reading skills 2 (Very good)	0.025	0.047	0.009	0.006	0.066	0.002
Reading skills 4 (Fair)	-0.107*	0.057	-0.039	0.061	0.079	0.022
Reading skills 5 (Poor/DK/RF)	-0.056	0.096	-0.021	-0.109	0.134	-0.040
Writing skills 1 (Excellent)	-0.216***	0.063	-0.080	0.306***	0.090	0.109
Writing skills 2 (Very good)	-0.048	0.048	-0.018	-0.032	0.068	-0.012
Writing skills 4 (Fair)	-0.022	0.051	-0.008	-0.030	0.072	-0.011
Writing skills 5 (Poor/DK/RF)	0.019	0.083	0.007	-0.142	0.118	-0.053
Word recall - first time: 0-4	-0.097***	0.034	-0.036	0.049	0.049	0.018
Word recall - first time: 6-10	0.048	0.036	0.018	-0.020	0.051	-0.007
Word recall - second time: 0-4	0.037	0.040	0.014	-0.049	0.056	-0.018
Word recall - second time: 6-10	-0.032	0.048	-0.012	0.057	0.068	0.021
Numeracy score 1 (Bad)	-0.144**	0.056	-0.053	0.060	0.080	0.022
Numeracy score 2	-0.010	0.039	-0.004	-0.035	0.055	-0.013
Numeracy score 4	0.087***	0.033	0.032	0.015	0.048	0.006
Numeracy score 5 (Good)	0.024	0.041	0.009	-0.053	0.059	-0.019
Orientation date 0-3 (Bad)	0.061	0.039	0.022	-0.049	0.055	-0.018
Depression scale >3 (depressed)	-0.165***	0.032	-0.061	0.080*	0.046	0.029
Constant	0.178	0.213		0.289	0.310	0.106
Number of observations	22,131					
Log-likelihood	-14,232.854					
Pseudo R^2	0.063					
Wald test on interacted variables (χ^2)				191.380***		

Note: *** p<0.01, ** p<0.05, * p<0.1.

Table 12: Probit; dependent variable = P(numerical concordance)

	(1)	(2)	(3)	(4)	(5)	(6)
	Variables whole sample			Interacted variables US first		
Covariate	Coef.	Std.Err.	Marg. 1	Coef.	Std.Err.	Marg. 1
Age	-0.002	0.002	-0.001	0.001	0.003	0.000
Female	-0.062**	0.029	-0.020	0.092**	0.042	0.030
Born in country of interview	-0.011	0.043	-0.004	-0.101	0.064	-0.033
Married	0.029	0.070	0.009	-0.082	0.102	-0.027
Austria	0.124	0.086	0.042	0.230*	0.124	0.079
Belgium	0.226***	0.076	0.077	-0.147	0.111	-0.046
Denmark	0.398***	0.099	0.140	0.285**	0.141	0.099
Germany	0.012	0.043	0.004	0.095	0.063	0.031
Greece	0.034	0.074	0.011	-0.001	0.109	-0.000
Italy	0.201***	0.044	0.067	-0.135**	0.066	-0.043
Netherlands	-0.005	0.068	-0.002	0.163*	0.097	0.055
Spain	0.021	0.050	0.007	0.028	0.073	0.009
Sweden	0.608***	0.079	0.218	0.025	0.111	0.008
Switzerland	-0.014	0.090	-0.004	0.070	0.132	0.023
Education level 0	-0.158	0.161	-0.050	-0.135	0.248	-0.043
Education level 1	-0.207	0.155	-0.066	0.054	0.238	0.018
Education level 2	-0.227	0.155	-0.071	0.143	0.238	0.048
Education level 3-4	-0.182	0.153	-0.059	-0.030	0.236	-0.010
Education level 5-6	-0.144	0.155	-0.046	0.037	0.238	0.012
ln(Income)	0.003	0.005	0.001	0.006	0.007	0.002
ln(HH net worth)	-0.005	0.005	-0.001	-0.005	0.007	-0.002
ln(HH worth)*Married	0.002	0.006	0.001	0.007	0.009	0.002
Chronic diseases: 0	0.116***	0.041	0.038	0.320***	0.060	0.110
Chronic diseases: 1	0.017	0.037	0.006	0.220***	0.055	0.074
Chronic diseases: 3 or more	-0.014	0.043	-0.004	-0.021	0.064	-0.007
Symptoms: 0	0.079**	0.040	0.026	0.186***	0.060	0.063
Symptoms: 1	-0.064*	0.038	-0.021	0.115**	0.058	0.038
Symptoms: 3 or more	0.020	0.045	0.007	0.053	0.067	0.017
Eyesight 1 (Excellent)	0.116**	0.057	0.039	0.019	0.083	0.006
Eyesight 2 (Very good)	0.110***	0.040	0.037	0.001	0.059	0.000
Eyesight 4 (Fair)	-0.062	0.048	-0.020	-0.006	0.071	-0.002
Eyesight 5 (Poor/blind)	0.089	0.105	0.030	0.123	0.152	0.041
Eyesight distance 1 (Excellent)	0.193***	0.049	0.065	-0.074	0.074	-0.024
Eyesight distance 2 (Very good)	0.119***	0.039	0.039	-0.111*	0.058	-0.036
Eyesight distance 4 (Fair)	0.118**	0.056	0.039	0.099	0.081	0.033
Eyesight distance 5 (Poor/blind)	0.075	0.091	0.025	0.088	0.132	0.029
Eyesight close 1 (Excellent)	0.023	0.057	0.008	-0.090	0.083	-0.029
Eyesight close 2 (Very good)	0.076*	0.040	0.025	-0.062	0.060	-0.020
Eyesight close 4 (Fair)	-0.094**	0.042	-0.030	0.210***	0.061	0.071
Eyesight close 5 (Poor/blind)	-0.140***	0.054	-0.045	0.092	0.079	0.031

To be continued on next page

Table 12: Probit; dependent variable = P(numerical concordance)

	(1)	(2)	(3)	(4)	(5)	(6)
	Variables whole sample			Interacted variables US first		
Covariate	Coef.	Std.Err.	Marg. 1	Coef.	Std.Err.	Marg. 1
Hearing 1 (Excellent)	0.296***	0.039	0.101	-0.195***	0.057	-0.061
Hearing 2 (Very good)	0.181***	0.033	0.061	-0.076	0.049	-0.025
Hearing 4 (Fair)	-0.007	0.041	-0.002	0.060	0.060	0.020
Hearing 5 (Poor)	-0.142*	0.084	-0.045	0.327***	0.123	0.114
Medical doctor: 6 or more times	-0.001	0.029	-0.000	-0.056	0.043	-0.018
Reading skills 1 (Excellent)	-0.066	0.063	-0.021	0.307***	0.092	0.105
Reading skills 2 (Very good)	0.013	0.047	0.004	-0.091	0.070	-0.029
Reading skills 4 (Fair)	0.132**	0.058	0.044	-0.020	0.085	-0.006
Reading skills 5 (Poor/DK/RF)	0.147	0.097	0.049	0.272*	0.142	0.094
Writing skills 1 (Excellent)	0.238***	0.064	0.080	-0.343***	0.094	-0.104
Writing skills 2 (Very good)	0.051	0.048	0.017	0.083	0.071	0.027
Writing skills 4 (Fair)	-0.008	0.053	-0.002	-0.030	0.078	-0.010
Writing skills 5 (Poor/DK/RF)	-0.108	0.085	-0.034	0.080	0.126	0.027
Word recall - first time: 0-4	0.109***	0.035	0.036	-0.084*	0.051	-0.027
Word recall - first time: 6-10	-0.064*	0.037	-0.021	0.047	0.054	0.016
Word recall - second time: 0-4	-0.070*	0.040	-0.023	0.086	0.059	0.028
Word recall - second time: 6-10	0.059	0.048	0.019	-0.077	0.070	-0.025
Numeracy score 1 (Bad)	0.161***	0.057	0.054	-0.143*	0.085	-0.045
Numeracy score 2	0.056	0.039	0.019	-0.052	0.058	-0.017
Numeracy score 4	-0.047	0.034	-0.015	-0.015	0.050	-0.005
Numeracy score 5 (Good)	0.016	0.042	0.005	0.059	0.061	0.019
Orientation date 0-3 (Bad)	-0.067*	0.040	-0.022	0.087	0.058	0.029
Depression scale >3 (depressed)	0.075**	0.033	0.025	-0.023	0.048	-0.007
Constant	0.465**	0.213		-0.589*	0.321	-0.191
Number of observations	22,131					
Log-likelihood	-12,758.009					
Pseudo R^2	0.064					
Wald test on interacted variables (χ^2)				261.980***		

Note: *** p<0.01, ** p<0.05, * p<0.1.

Table 13: Self-assessed health, wave 1 and wave 2 compared

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Health US: wave 1 to wave 2			Wave 2 health vs. wave 1			Prop.
Subsample	Healthier	Same	Worse	Better	Same	Worse	consistent
Total	0.209	0.484	0.307	0.056	0.670	0.274	0.487
Austria	0.208	0.452	0.341	0.059	0.675	0.266	0.478
Belgium	0.213	0.486	0.302	0.063	0.717	0.221	0.491
Denmark	0.215	0.477	0.308	0.076	0.656	0.268	0.503
France	0.170	0.521	0.310	0.072	0.727	0.201	0.528
Germany	0.225	0.496	0.280	0.054	0.623	0.323	0.487
Greece	0.202	0.556	0.242	0.012	0.906	0.082	0.548
Italy	0.215	0.471	0.314	0.042	0.631	0.327	0.459
Netherlands	0.199	0.491	0.310	0.073	0.677	0.249	0.495
Spain	0.216	0.436	0.348	0.043	0.662	0.295	0.455
Sweden	0.180	0.413	0.408	0.101	0.667	0.233	0.475
Switzerland	0.232	0.479	0.289	0.098	0.748	0.154	0.504
Male	0.205	0.481	0.313	0.050	0.701	0.249	0.480
Female	0.211	0.486	0.302	0.061	0.643	0.297	0.493
Test for equality of means	*0.366*	*0.529*	*0.140*	*0.006*	*0.000*	*0.000*	*0.053*
Age 50-54	0.207	0.468	0.326	0.063	0.728	0.209	0.501
Age 55-59	0.217	0.481	0.301	0.078	0.703	0.219	0.485
Age 60-64	0.219	0.506	0.275	0.043	0.665	0.292	0.493
Age 65-69	0.198	0.488	0.315	0.055	0.634	0.311	0.484
Age 70-75	0.199	0.480	0.320	0.033	0.592	0.374	0.466
Education level 0	0.209	0.486	0.305	0.053	0.597	0.351	0.466
Education level 1	0.209	0.468	0.323	0.044	0.641	0.315	0.467
Education level 2	0.219	0.459	0.321	0.062	0.669	0.269	0.481
Education level 3	0.211	0.495	0.294	0.055	0.682	0.263	0.497
Education level 4	0.174	0.473	0.353	0.063	0.688	0.248	0.438
Education level 5 and 6	0.200	0.506	0.294	0.066	0.705	0.229	0.510
Other education	0.141	0.495	0.364	0.074	0.546	0.380	0.474
Married	0.204	0.484	0.312	0.051	0.676	0.273	0.493
Single	0.222	0.484	0.294	0.069	0.653	0.277	0.471
Test for equality of means	*0.022*	*1.000*	*0.039*	*0.000*	*0.425*	*0.245*	*0.232*
Born in country of interview	0.209	0.483	0.308	0.056	0.673	0.271	0.487
Not born in country of interview	0.203	0.498	0.299	0.056	0.635	0.308	0.484
Test for equality of means	*0.669*	*0.356*	*0.528*	*0.996*	*0.006*	*0.027*	*0.617*

Note: Column (1), (2) and (3) compare self-assessed health as measured by the US version in wave 2 to wave 1, column (4), (5) and (6) show what proportions rate their health in wave 2 better, the same or worse than 2 years ago, as measured by the question *"Compared with your health when we talked with you in month and year previous interview, would you say that your health is better now, about the same, or worse?"*, and column (7) shows what proportion of the sample gave consistent responses between these two.

Table 14: Pseudo R^2's for probit regressions (dependent variable = change in self-assessed health, wave 2 - wave 1)

	(1)	(2)	(3)	(4)	(5)
	Whole sample	Word conc.	Num. conc.	Discor.	No basis
Number of observations	14,768	7,752	5,064	1,404	548
Added variables			Pseudo R^2		
(a) Socio-demographics	0.002	0.004	0.005	0.015	0.044
(b) Changes in health and cognition	0.049	0.065	0.069	0.070	0.169
(c) Concordance in wave 1	0.092				

Note: Three different ordered probit models were performed for the change in self-assessed health between wave 1 and 2; the table shows levels of pseudo-R^2s for the models as sets of variables were incrementally added: model (a) only contains socio-demographic covariates, model (b) contains socio-demographic covariates as well as changes in health and cognition between wave 1 and wave 2, and model (c) consists of all variables in model (b), with concordance measures in wave 1 added.

Table 15: Actual and predicted distributions, whole sample

	Actual distribution					
	1.Excellent	*2. Very good*	*3. Good*	*4. Fair*	*5. Poor*	
	0.068	0.162	0.413	0.273	0.083	
	Predicted distribution					Proportion
Added variables	*1.Excellent*	*2. Very good*	*3. Good*	*4. Fair*	*5. Poor*	pred. correctly
(a) Socio-demographics	0.086	0.188	0.427	0.238	0.062	0.484
(b) Changes in health and cognition	0.082	0.182	0.421	0.243	0.070	0.489
(c) Concordance in wave 1	0.056	0.181	0.432	0.257	0.068	0.504

Note: The table shows the actual distribution of responses to the self-assessed health question in wave 2, along with the predicted distribution when using probit models with the different variables incrementally added. The last column shows for what proportion of the individuals the response to the self-assessed health question in wave 2 was predicted correctly.

Table 16: Pseudo R^2's for probit regressions (dependent variable = P(major health event))

	(1)	(2)	(3)	(4)	(5)
	Whole sample	Word conc.	Num. conc.	Discor.	No basis
Number of observations	15,314	8,023	5,252	1,443	596
Added variables			Pseudo R^2		
(a) Socio-demographics	0.044	0.052	0.049	0.079	0.083
(b) Health and cognition in wave 1	0.074	0.095	0.108	0.213	0.252
(c) Self-assessed health in wave 1	0.090	0.104	0.126	0.251	0.267
(d) Interaction with "US first"	0.107	0.129	0.174	0.454	0.486
(e) Concordance in wave 1	0.109				

Note: Five different probit models were performed for the probability of a major health event; the table shows levels of pseudo-R^2s for the models were covariates were incrementally added: model (a) only contains socio-demographic variables, in model (b) variables on health and cognition in wave 1 were additionally added to model (a), model (c) additionally contains self-assessed health responses in wave 1, in model (d) a full set of interaction variables for "US asked first" was added to model (c), and in model (e) information on concordance in wave 1 was included.

Table 17: Proportion major health event predicted correctly, whole sample

	(1)	(2)	(3)	(4)
Added variables	Prop. predicted correctly	Number of individuals	Prop. of 1's predicted correctly	Number of individuals
(a) Socio-demographics	0.877	13,526	0.175	179
(b) Health and cognition in wave 1	0.884	13,678	0.217	200
(c) Self-assessed health in wave 1	0.887	13,687	0.237	225
(d) Interaction with "US first"	0.891	13,719	0.270	229
(e) Concordance in wave 1	0.893	13,729	0.278	234

Note: Column (1) of the table shows what proportion of the sample is predicted correctly on having a major health event, when using five probit models were variables were incrementally added to the set of explanatory variables; these proportions are equal to the number of individuals shown in column (2). In column (3) the proportion of individuals that had a major health event that is predicted correctly is shown for each of the five probit regressions, along with the actual number of individuals in column (4).