HEDG
HEALTH, ECONOMETRICS AND DATA GROUP
THE UNIVERSITY of York

WP 12/10

A comparison of parametric and non-parametric adjustments using vignettes for self-reported data

Andrew M. Jones, Nigel Rice, Silvana Robone

May 2012

york.ac.uk/res/herc/hedgwp

A comparison of parametric and non-parametric adjustments using vignettes for self-reported data

Andrew M. Jones, Nigel Rice, Silvana Robone

16 March 2011

Abstract

This paper compares the use of parametric and non-parametric approaches to adjust for heterogeneity in self-reported data. Despite the growing popularity of the HOPIT model to account for reporting heterogeneity when dealing with self-reported categorical data, recent evidence has questioned the validity of this heavily parametric approach. We compare the performance of the HOPIT model with the non-parametric estimators put forward by King et al. (2004) and King and Wand (2007). Using data relating to the health domains of mobility and memory from the Survey of Health, Ageing and Retirement in Europe (SHARE) we perform pairwise country comparisons of self-reported health, objective measures of health, and measures of health adjusted for the presence of reporting heterogeneity. Our study design focuses on comparisons of countries where there exist a discrepancy between the distribution of self-reported data and objective measures of health and assesses whether vignettes are able to reconcile this difference. Comparisons of distributions are based on first order stochastic dominance. In general, HOPIT and non-parametric estimation produce similar results in terms of first order stochastic dominance for the domains of both mobility and memory. Neither method consistently explains discrepancies across countries between self-reported and objective measures of health mobility and memory.

Keywords: reporting heterogeneity, anchoring vignettes, Hierarchical Ordered Probit, stochastic dominance, cross-country comparisons, health, SHARE

Corresponding Author: Professor Andrew Jones, Department of Economics and Related Studies, University of York, York, YO10 5DD (andrew.jones@york.ac.uk)

Acknowledgements: This research was funded by the Economic and Social Research Council under the Large Grant Scheme, (RES-060-25-0045). We would like to thank Ranjeeta Thomas, Marta Soares and Pedro Rosa Dias for valuable advice and participants of the 8th World Congress on Health Economics (Toronto, July 2011), the Annual Conference of the Italian Association of Health Economics (Naples, Sep 2011) and the Health, Econometrics and Data Group Seminar Series at the University of York for helpful comments.

1. Introduction

Self reported data are ubiquitous in social surveys. For example, respondents are often asked their opinion about their customer satisfaction, their job or life satisfaction, their satisfaction with public services or their health status. Although these kinds of question are widely used by social researchers, the comparability of responses across survey respondents is often questionable. Consider, as an example, self-assessed health status, which is often measured on an ordered 5-point categorical scale, ranging from very poor to excellent health. A common problem with such scales is that individuals faced with the self-reported health instrument are likely to interpret the meaning of the available response categories in a way that systematically differs across populations or population sub-groups (Salomon et al. 2004). This implies that individuals use different thresholds when mapping their `true' underlying level of health to the response scale available in the survey question. This phenomenon is variously referred to as differential item functioning, differential reporting behaviour, reporting heterogeneity or response cut-point shift. Where this occurs randomly across individuals it does not present a major concern when analysing health outcomes. Systematic variation in reporting behaviour across individuals is more problematic and can be particularly troublesome for cross-country comparative analysis where differences in social norms and expectations are likely to heavily influence the type of response scale (Salomon et al. 2004, Pfarr et al. 2011). Accordingly, analyses of health outcomes may produce invalid inference should country differences in reporting not be taken into account.

Anchoring vignettes have been proposed as a way to address the issue of reporting heterogeneity (King et al. 2004). Vignettes represent hypothetical descriptions of fixed levels of a latent construct such as health status. Since respondents rate the same fixed level described by a given vignette, differences in their ratings are assumed to be due to differences in reporting behaviour. Thus vignettes offer a useful means to assess systematic variation in ratings by relating respondent assessments to their socioeconomic and demographic characteristics. This information can then be used to adjust self-reported data to achieve greater cross-respondent comparability. A number of large-scale social surveys such as the Survey of Health, Ageing and Retirement in Europe (SHARE), the U.S. Health and Retirement Study (HRS), the English Longitudinal Study of Ageing (ELSA), and the World Health Survey (WHS) have introduced vignettes to be used alongside self-reported data.

The majority of studies that address the issue of reporting heretogeneity using vignettes have adopted the hierarchical ordered probit (HOPIT) model (Tandon et al. 2003). The HOPIT model is an extension of the standard ordered probit model that allows the cut-point thresholds (that separate the response categories) to vary across individuals as functions of respondent characteristics. In so doing, the model allows for systematic reporting behaviour to vary across respondents. The HOPIT model has been applied in several areas of economics and social science, for example, to investigate: self-reported data on health status (Iburg et al. 2002, King et al. 2004, Bago d'Uva et al. 2008, Peracchi and Rossetti 2009, Grol-Prokopczyk et al. 2011); healthy behaviours (van Soest et al. 2011); satisfaction with health systems performance (Valentine et al. 2003, Puentes Rosas et al. 2006, Sirven et al. 2008, Rice et al. 2012); work disability (Kapteyn et al. 2007, Kapteyn et al. 2009, Angelini et al. 2011a, Paccagnella 2011); political efficacy (King et al. 2004), job satisfaction (Kristensen and Johansson 2008); life satisfaction (Angelini et al. 2011b), satisfaction with income (Kapteyn et al. 2011a) and consumer satisfaction with products and services (Rossi et al. 2001).

While the parametric HOPIT model has dominated empirical applications of the vignette approach, non-parametric methods have also been developed to address the issue of reporting heterogeneity, but these have seldom been applied in the literature (Chevalier and Fielding 2011). In particular, King et al. (2004) and King and Wand (2007) have proposed an approach which exploits a respondent's ordering of the vignettes and the relative position of their self-assessment within this ordering. To our knowledge, the only other study that has applied this approach is Hudson (2011), which investigates reporting heterogeneity in parents` assessments of their children`s respiratory health. The non-parametric approach has the advantage of avoiding the parametric assumptions inherent in the HOPIT model. A potential disadvantage is that data is required on all respondents for both the self-assessments and ratings of the full set of vignette questions. In contrast, the HOPIT approach only requires respondents to answer the self reports together with a sub-set of vignettes, which is preferable for survey questionnaire design (King et al. 2004). In addition, the non-parametric approach requires the ability to order the vignettes from best to worst. This might not be straightforward for some concepts of interest. For example, if we consider the vignettes included in the World Health Survey questionnaire on respectful treatment as an indicator of the quality of health services, patient's valuations of different aspects of respectful care may very considerably leading to no natural ordering among the vignettes.

Despite the growing popularity of the vignette methodology to address the issue of reporting heterogeneity, the formal evaluation of the validity of the approach remains a topic of ongoing research. For the approach to be valid, two assumptions need to hold. The first, termed *vignette equivalence*, implies that "the level of the variable represented by any one vignette is perceived by all respondents in the same way and on the same unidimensional scale" (King et al. 2004, p. 194). This assumes that all respondents agree on the underlying latent level described by the vignette except for random error. The second assumption, termed *response consistency*, implies that individuals use the same mapping from the underlying latent scale to the available ordered response categories when responding to both the self-assessments and the vignette questions. This assumption allows the relationship between reporting behaviour and characteristics of respondents obtained using the responses to the vignettes to be used to adjust respondents' self-reports of the underlying construct of interest.

The empirical literature investigating the two assumptions is equivocal. While Murray et al. (2003), King et al. (2004), Kristensen and Johansson (2008), Rice et al. (2011) and Hudson (2011) provide evidence in support of the assumption of vignette equivalence, largely making use of non-parametric methods, Datta Gupta et al. (2010), Peracchi and Rossetti (2010) and Bago d'Uva et al. (2012) do not. Corrado and Weeks (2010) are sceptical about the comparability of survey responses across countries and develop a test that allows the identification of subsets of countries where the assumption of vignette equivalence holds. For response consistency Kapteyn et al. (2011b) and van Soest et al. (2011) provide supporting evidence, whereas Bago d'Uva et al. (2012) and Peracchi and Rossetti (2010) reject the assumption. It is notable that the studies which test response consistency (with the exception of Peracchi and Rossetti 2010) rely on the availability of objective measures of the concept of interest.[1] An important consideration, which is often overlooked is whether the use of vignettes may still aid in adjusting comparisons of self-reports closer to some `true' underlying difference even where the assumptions are rejected by statistical criteria. That is, are the methods helpful in improving comparability where the assumptions of response consistency and vignette equivalence fail on statistical criteria.

[1] Peracchi and Rossetti (2010) provide an important contribution by demonstrating how response consistency and vignette equivalence can be tested in the absence of objective measures. They exploit the fact that under these assumptions the model is over-identified where there are one or more vignettes available for the concept under analysis. The test applied to health domains in SHARE rejects the assumptions of response consistency and vignette equivalence.

This paper considers two research issues. First, we compare the relative performance of a non-parametric approach as an alternative to the HOPIT model to adjust for reporting heterogeneity. Secondly, we assess how well the methods adjust self-reports of health status towards `true' underlying health. We use objective measures of health as a benchmark in assessing these related issues. We do this by drawing on data from the Survey of Health, Ageing and Retirement in Europe (SHARE), a survey of household members born in or before 1954 across twelve European countries. The data are particularly useful as they contain information on self-reports of health status together with objective measures and vignettes for two domains of health. This enables us to undertake pairwise country comparisons of self-reported health (hereinafter, SRH), objective measures of health, and adjusted measures of SRH purged of reporting behaviour, with adjustment derived from both parametric (HOPIT) and non-parametric methods. We adopt a study design that focuses on pairwise comparisons of countries where there exists a discrepancy between comparisons based on self-reported health and comparisons based on objective measures of health. That is, where comparisons based on objective measures suggest a difference in health status, but self-assessments of health do not, or vice versa. Differences are based on first-order stochastic dominance of respective distributions of health. This set-up allows us to investigate whether vignettes are helpful in adjusting comparison of self-reports of health towards the `true' underlying differences observed in the objective measures (irrespective of assumptions about vignette equivalence and response consistency). It further allows us to assess the relative performance of the HOPIT model compared to a non-parametric approach.

2. Methods

The reporting of SRH is via an ordered categorical variable which is assumed to be a discrete representation of some underlying latent scale. Should individuals map the latent scale to the survey response categories in a consistent way, irrespective of their characteristics or circumstances, then we would observe homogeneous reporting behaviour. In these circumstances the standard ordered probit estimator that assumes a set of fixed thresholds applicable to all respondents would offer an appropriate method to model the data. However, where individuals systematically differ in the positioning of thresholds to map the latent construct to the available response categories, then reporting heterogeneity arises. This obfuscates meaningful analyses across groups of individuals and methods to correct for reporting heterogeneity are required to improve comparability.

2.1. Non-parametric methods

King et al. (2004) and King and Wand (2007) propose a non-parametric approach for examining the categorical self-reports and associated vignettes. The method exploits the ordering of the vignettes and the relative position of the self-assessment rating within this ordering. To implement the method, an individual's categorical self-assessed response is recoded to locate it relative to their rating of the set of vignettes. Accordingly, define y_i as the categorical self-assessment for respondent *i* and $r_{i1}, \dots, r_{iK}$ the corresponding set of *K* vignette responses. Respondents are presented the same set of response categories for both the self-assessment and the set of vignettes. Assuming all respondents order the vignettes in an identical way $\left(r_{i,k-1} < r_{i,k}, \ \forall i,k\right)$, then King et al (2004) define a recoded response C_i as

$$C_i = \begin{cases} 1 \ if \ y_i < r_{i1} \\ 2 \ if \ y_i = r_{i1} \\ 3 \ if \ r_{i1} < y_i < r_{i2} \\ 4 \ if \ y_i = r_{i2} \\ \vdots \\ 2K+1 \ if \ y_i > r_{iK} \end{cases} \quad (1)$$

Accordingly, C defines a scale that places the self-ratings relative to the respondent's assessment of the set of vignettes. The scale on which C lies includes a greater number of possible categories than the original scale on which the self assessments where made but importantly is purged of differential reporting behaviour. Once obtained, C_i can be used to perform direct comparisons across individuals or can be modelled using parametric methods, such as the ordered probit model. Where respondents fail to uniquely differentiate between vignettes leading to ties in their ratings, then it is suggested to define C_i by a set of values (or range) rather than a single value. For example, if $y_i = r_{i1} = r_{i2}$, then $C_i = \{2,3,4\}$. Moreover, in practice respondents might rank the vignettes inconsistently. To retain the information contained in these responses, King et al. (2004) suggest grouping across the vignettes causing the inconsistency and treating these as ties. However, dealing with a set of values instead of a scalar value presents challenges to implementing the approach. King et al. (2004) suggest allocating C_i by assuming a uniform distribution for the values across the specified range,

while King and Wand (2007) extend the approach by developing a generalisation of the ordered probit model. Their censored ordered probit model, models C_i by including the potential range of values that C_i might take within the thresholds that map the latent scale to the observed outcome assuming that the latent variable has a normal distribution. For the standard ordered probit model a latent outcome, y_i^*, is modelled as a conditional normal with mean $X_i\beta$ and variance 1. For identification the constant is set to zero. For scalar values of C_i, the observation mechanism relating the latent variable y_i^* to the observed category C_i can be expressed as:

$$C_i = c \quad \text{if} \quad \mu_{c-1} \leq y_i^* < \mu_c \tag{2}$$

for $= 1, \dots, 2K + 1$, with $\mu_0 = -\infty$ and $\mu_{2K+1} = \infty$

For vector values of C_i, the censored ordered probit model generalises the standard ordered probit model by extending the observation mechanism in equation (2) as follows:

$$C_i = c \quad \text{if} \quad \mu_{\min(c)-1} \leq y_i^* < \mu_{\max(c)} \tag{3}$$

Following King et al., (2004) and King and Wand (2007) we refer to the non-parametric approach as the C-estimator.

2.2. Hopit model

The HOPIT model (developed by Tandon et al. 2003) is an extension of the ordered probit model that allows the thresholds to vary across individuals as a function of respondent characteristics. The HOPIT model can be thought of as consisting of two parts. The first part uses vignettes to provide a source of exogenous information enabling the thresholds to be modelled as a function of relevant respondent covariates. The second part relates the set of regressors determining underlying latent health while controlling for differences in reporting behaviour by fixing the thresholds to the relationships obtained in the first part of the model.

For a formal description of the model see, as an example, King et al. (2004). This model has been described numerous times in the literature and we do not repeat this here.

3. Survey of Health, Ageing and Retirement in Europe (SHARE)

We make use of the first wave (2004-2005) of SHARE which is a survey of household members born in or before 1954 and covers twelve European countries (Austria, Belgium, Denmark, France, Germany, Greece, Israel, Italy, Netherlands, Sweden, Switzerland, Spain). The dataset has a similar design to the US Health and Retirement Study and contains information on the individual life circumstances of all eligible members of approximately 18,000 households. The data includes information relating to respondents' health overall and on six specific domains of health (breath, pain, mobility, work disability, depression and memory). For each domain three vignette questions are asked of respondents. In addition, the survey includes objective measures of health, notably in the domains of mobility and cognitive ability. We make use of data from Belgium, France, Germany, Greece, Italy, Netherlands, Spain, and Sweden, since only these countries fielded data on anchoring vignettes.

We use the two health-related domains mobility and memory for which 4408 and 4413 individuals, respectively, responded to both the questions on vignettes and their own health conditions. For self-reported mobility individuals are asked “Overall in the last 30 days, how much of a problem did you have with moving around?” while for memory the question reads “Overall in the last 30 days how much difficulty did you have with concentrating or remembering things”. For both questions, the available response categories are “None”, “Mild”, “Moderate”, “Severe”, “Extreme”. Table 1 reports the frequencies of responses to these categories across the eight counties for mobility and memory, respectively. These descriptive statistics show substantial variability in the reporting of health. For example, while 74% of Greek respondents report no problems with mobility, only 38% of Swedish respondents do likewise. However, 56% of Swedish respondents report no problems with memory contrasted against 34% of respondents in Belgium. The variation in reporting will reflect differences both in underlying health status and reporting behaviour. The purpose of using the methods outlined above is to purge the latter to reveal only differences in underlying health.

Respondents are asked to evaluate three vignettes related to the mobility domain and three related to memory. An illustrative vignette for mobility is: "Tom has a lot of swelling in his legs due to his health condition. He has to make an effort to walk around his home as his legs feel heavy. Overall in the last 30 days, how much of a problem did Tom have with moving around?". For memory, an illustrative vignette is: "Lisa can concentrate while watching TV, reading a magazine or playing a game of cards or chess. Once a week she forgets where her keys or glasses are, but finds them within five minutes. Overall in the last 30 days, how much difficulty did Lisa have with concentrating or remembering things?" When rating vignettes, individuals choose among the same categories as those available for the self-reports. Table 2 presents the percentage of respondents, by country, reporting each of the five categories for the two illustrative vignettes. This table clearly shows heterogeneity across respondents in reporting of the vignettes. For example, approximately 16% of respondents rate mobility problems described in the vignette as extreme, while around 49% and 30% rate this as severe and moderate, respectively. Given the fixed and exogenous nature of the vignettes this variation in respondents' ratings provides prime-facie evidence of differential reporting behaviour. Reporting behaviour also appears to vary across countries and while approximately 30% of respondents in Greece judge mobility described in the illustrative vignette as extreme, in France only around 4% of respondents chose this response category.

Our objective measure of mobility is based on a walking speed test. This is aimed at eliciting information about mobility and functioning of the lower limbs, and is implemented by a timed walk over a short distance (2.5 m). The measure is available only for those aged 75 years and over or younger respondents with self-reported mobility limitations. Since this measure is available for a subsample of individuals only we supplement it with a measure of hand grip strength. Hand grip strength is measured using a hand-held dynamometer, which shows strength in kilograms and this test is applied to all individuals without an age restriction.[2] Hand grip strength has been found to predict the severity of late-life disability and mortality (Frederiksen et al., 2002; Rantanen et al., 1998) and recently has been used elsewhere as an objective indicator of mobility (Bago d`Uva et al. 2012). For the memory domain, SHARE contains objective measures of cognitive ability, such as verbal fluency and numeracy. Among these, we exploit information provided by a memory test termed "ten

[2] Respondents are asked to squeeze a lever as hard as they can for a couple of seconds and then let go. Grip strength is calculated from this.

words list learning".[3] This test reports on the number of words an individual can recall from 10 words previously provided to respondents.

SHARE also contains rich information on individual socio-economic characteristics. In common with a great deal of the literature on reporting behaviour, we restrict our attention to a set of key variables and include age (as a continuous variable), gender (a dummy variable for men), household income (in Euros), education (a dummy variable for having a number of years of school below the average), and marital status (a dummy variable for being married or living with a partner). Table 3 reports summary statistics for these variables for each of the eight countries considered.

4. Study design

To evaluate how well the parametric and non-parametric approaches perform in adjusting for differential reporting behaviour we undertake pairwise country comparisons of SRH, of objective measures of health, and of SRH adjusted for reporting behaviour. The adjusted measures are derived from both the parametric and non-parametric models and comparisons are made for both mobility and memory health domains. Our study design focuses on pairwise comparison of countries where there exists a discrepancy between assessments based on self-reported health and those based on objective measures. Since the objective measure of mobility is recorded only for people aged over 75 years or younger respondents with self-reported mobility limitations, we supplement this measure with a measure of hand grip strength and consider comparisons where a discrepancy exists between self-reported mobility and the two objective measures.

We test for differences in health distributions using first order stochastic dominance. For self-reports and objective measures, which are measured on a categorical scale, we apply the Kolmogorov-Smirnov test (Kolmogorov 1933, Smirnov 1948). Strictly the Kolmogorov-Smirnov test is a test of equality of distributions and not specifically a test of stochastic dominance. Therefore, when the test suggests rejecting the null hypothesis of equality we confirm stochastic dominance by a visual inspection of graphs of the two distributions to

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

disregard comparisons where curves may cross. Comparisons of the adjusted health distributions are undertaken using the method developed by Anderson (1996). Since the non-parametric C estimator results in both scalar and vector values (for tied observations and inconsistent orderings), and the HOPIT model produces counterfactual predictions of the probability of belonging to each health status category, the Kolmogorov-Smirnov test is not applicable. The approach of Anderson (1996) is based on the population frequencies for each of the categories of the outcome of interest. We apply this test to the sample frequencies derived from the HOPIT and non-parametric model. The intuition behind Anderson`s test is the following. Let H be the range space of health levels from the health distributions of country A and country B, with cumulative distributions $F_A(h)$ and $F_B(h)$, respectively. First order stochastic dominance of B with respect to A is equivalent to and requires a test of the following condition:

$$F_A(h) \leq F_B(h),\ F_A(h_j) \neq F_B(h_j) \text{ for some j} = 1, \ldots, \text{J},\ \forall \text{ h} \in \text{H} \qquad (4)$$

The test is straightforward to implement. The combined sample for A and B is partitioned into *j* equal intervals and, in each interval, the empirical frequencies of the A and B samples is computed: for example, $p_{ij}^A = \frac{x_j^A}{n^A}$ j =1, …J , where x_j^A is the number of observations in country A in the j^{th} interval *j*, and $n^A$is the total number of observations in country A. Let I_f be a k x k lower-triangular matrix of ones. Since the cumulative distribution function at a point *l* can be computed as $F(c_l) = \sum_{l=1}^{l} p_l$, a test of condition (7) requires testing the following hypothesis:

$$H_0 : I_f\,(p^A - p^B) = 0 \quad \text{against} \quad H_1 : I_f\,(p^A - p^B) \leq 0 \qquad (5)$$

In particular, first-order dominance of distribution B over A requires that no element of the vector $I_f\,(p^A - p^B)$ be significantly greater than zero, while at least one element is significantly negative.[4] The significance of the test statistic $I_f\,(p^A - p^B)$ can be assessed through the use of the studentized maximum modulus distribution (Stoline and Ury 1979).

Sample frequencies for the categories of self-reported health adjusted for reporting hererogeneity can be computed from the simulations (predictions) from the HOPIT model. These simulated frequencies are obtained assuming all individuals adopt the reporting

[4] Since this is symmetric, in order for distribution A to dominate distribution B, no element of the vector $I_f\,(p^A - p^B)$ should be significantly negative, while at least one element should be significantly positive.

behavior of a selected baseline country. A common reporting style for both countries is adopted in order to try to purge these simulated frequencies from reporting heterogeneity (see, for example, Rice et al. 2012).

Issues due to ties and inconsistencies in the ordering of the vignettes are dealt with by following the two approaches suggested by King et al (2004) and King and Wand (2007). The first approach assigns a discrete value to an individual where inconsistencies or ties have resulted in C taking a range or vector of values by assuming the values over the range have a uniform distribution (see King et al., 2004). We refer to this as the "uniform distribution approach". The greater number of cases with interval values and the larger the intervals, the more uniform the resulting distribution will look. The second approach, proposed by King and Wand (2007), involves estimating the censored ordered probit model described above to regress C_i on a set of individual characteristics (those used in the HOPIT model). The resulting predictions from the model are then used to compute the sample frequencies.[5]

Estimation of the HOPIT model was undertaken using STATA 11, while the non-parametric estimator C was computed using the package "anchors" available in R and developed by Wand, King and Lau (2011).

5. Results

Cross-country comparisons of unadjusted SRH and their respective objectives counterparts lead to five pairwise comparisons for mobility and twelve for memory. As an illustration, Figure 1 compares mobility in the Netherlands to Italy and shows the cumulative distribution of SRH, of the objective measures of health (walking speed and hand grip strength) and of the predicted frequencies of health derived from the HOPIT model and the non-parametric C estimator (for brevity, we show only the results from assuming ties and inconsistencies are dealt with by imposing a uniform distribution). Visual inspection of the figure suggests that neither country stochastically dominates for SRH, but that The Netherlands dominates Italy on the objective measures of mobility (walking speed and hand grip strength) indicating greater mobility in The Netherlands compared to Italy. Measures of SRH adjusted for reporting behaviour also suggests stochastic dominance for The Netherlands over Italy. This set of results imply that adjusting SRH for differences in

[5] While neither model is devoid of parametric assumptions, these are only used to deal with the problems of ties and inconsistencies in the ordering of the vignettes and we continue to refer to the underlying approach as non-parametric.

reporting leads to comparisons of self-assessments more closely aligned to the comparison of the objective measures than the comparison based on unadjusted SRH. This supports the use of vignettes as a credible way to adjust self-reported data for cultural differences in norms and expectations.

The preliminary evidence reported in Figure 1 is supported by the results we get when running formal tests of stochastic dominance. Table 4 reports D (the largest difference between the distribution functions of the two countries under comparison) and the p-values for the Kolmogorov-Smirnov first order stochastic dominance tests for SRH, walking speed and hand grip strength. The null hypothesis is a test of equality of the cumulative distribution functions of the two countries (we test the corresponding test statistics at the 5% critical value). Table 5 shows the corresponding statistics for the Anderson`s test, reported for predictions from the HOPIT model and the C non-parametric estimator. For the latter model, we consider predictions from both the 'uniform distribution approach' and the censored ordered probit model. Since both kinds of predictions produce very similar results, we focus only on results derived from the approach that assumes a uniform distribution for ties and inconsistencies. In the column for the HOPIT model each statistic corresponds to one of the five categories of self-reported health, while in the columns for the C estimator each statistic corresponds to one of the seven possible values taken by C. The significance of these test statistics is assessed through the use of the studentized maximum modulus distribution (Stoline and Ury 1979). With 5 and 7 statistics each (and infinite degrees of freedom), the 5% critical value of the studentized maximum modulus distribution are 2.80 and 3.03 for the HOPIT and the C estimator, respectively.

For ease of reference Table 6 presents a summary of the stochastic dominance results presented in Tables 4 and 5. For SRH mobility we observe stochastic dominance for The Netherlands compared to Sweden (The Netherlands dominates) and Italy compared to Greece (Greece dominates). These are not, however, supported by the objective measures where stochastic dominance is not established. The converse holds for the other paired country comparisons where dominance is found when comparing across objective measures (The Netherlands dominates Italy and Belgium, while Belgium dominates Italy) but no stochastic dominance is observed for comparisons of SRH. We are interested in assessing whether the application of the HOPIT model or the non-parametric C estimator results in adjusted distributions of health that more closely resemble the comparison of the distributions of the objective measures than comparisons of the raw unadjusted self-reports.

In terms of tests of stochastic dominance, in general, both the HOPIT and the C estimator result in similar conclusions and there does not appear to be any clear advantage in using a less restrictive non-parametric approach over the more standard HOPIT model. If we compare The Netherlands with Italy, or Belgium or Sweden, and compare Italy with Greece, both estimators indicate the presence of stochastic dominance (Belgium, Sweden and Italy are dominated by The Netherlands and Italy is dominated by Greece, respectively). Only two of these comparisons, however, produce results that correspond to those found when comparing the objective measures of mobility. Comparisons for Italy with Greece, and The Netherlands with Sweden produce adjusted distributions that more closely resemble the comparisons of the unadjusted distributions than those from the objective measures. Here the use of vignettes does not appear to aid in enhancing the comparability of the data. A similar result is found when comparing Italy with Belgium where again, comparisons using stochastic dominance criteria result in adjusted distributions of health that resemble the unadjusted SRH more closely than the distributions of the objective measures.

For memory, we observe stochastic dominance for three out of the twelve country comparisons when considering SRH and nine comparisons for the objective recall memory test. As with mobility, in general, application of the non-parametric C estimator and the HOPIT model results in comparisons of adjusted health distributions that agree in terms of stochastic dominance criteria (there are two exceptions: Germany and Italy, and Spain and Italy).[6] However, once again no clear pattern is discernable when comparing the dominance of the distributions for the vignette adjusted measures of memory compared to the comparisons of the objective measures.

6. Discussion and Conclusions

Many social surveys require respondents to rate their satisfaction with various aspects of life which tend to be measured on an ordered categorical scale. A common problem with such scales is that individuals may interpret the meaning of the available response categories in a way that systematically differs across populations or population sub-groups (Salomon et

[6] When comparing Spain with Belgium with regard to the C estimator it is not clear if one country dominates the other or not. The Anderson (1996) test suggests the presence of stochastic dominance, however a visual inspection of the graph of the two distributions suggests that no country clearly dominates over the other. The Anderson test may reject a null of equality of distributions due to either dominance indeterminacy or through curve-crossing (Yalonetzky 2011).

al. 2004). This reporting heterogeneity hinders comparability, particularly where comparison is sought across countries where social norms and expectations may differ markedly. The use of anchoring vignettes has been proposed as a means to address the issue of reporting heterogeneity (King et al. 2004). The most common approach to incorporating information contained in responses to vignettes is via the use of the HOPIT model (Tandon et al. 2003). More recently an approach termed the non-parametric C estimator has been proposed as an alternative that relaxes the parametric assumptions underlying the use of the HOPIT model (King and Wand 2007). For both approaches, identification relies on the assumptions of response consistency and vignette equivalence. However, an important practical consideration is whether, even where these assumptions do not hold, the application of the vignette approach moves self-reported data closer to the 'true' underlying latent construct it purports to measure.

This paper attempts to evaluate the performance of both parametric and non-parametric estimators for adjusting for differential reporting behaviour to assess whether either approach consistently dominates the other. Using data from SHARE across eight European countries we perform pairwise country comparisons where there exist discrepancies in terms of stochastic dominance of the distributions of SRH and the distributions of objective measures of health. These distributions are compared to those derived from SRH adjusted for reporting behaviour using the HOPIT model and the non-parametric C estimator. Based on the assumption that the objective measures are true reflections of underling health status in these domains, we fail to find a consistent pattern in the results to suggest that either approach satisfactorily addresses the issue of differential reporting. For some pairwise country comparisons the predictions obtained through the HOPIT estimator and the C estimator replicate better the results obtained through the objective measures of health than those obtained through the self-reported measures of health. However, for other pairwise comparisons this does not hold. In general, we observe similar results when using the HOPIT model and the non-parametric C estimator to adjust self-reports for differential reporting behaviour.

Our research design selects pairwise country comparisons based on differences in distributions of SRH but not in distributions of objective measures, or vice versa, where differences are defined by tests of first order stochastic dominance supported by graphical inspection of the distributions. Tests of stochastic dominance of vignette adjusted differences in self-reports, using either the HOPIT model or the non-parametric approaches, based on

both statistical criteria and visual graphical inspection (more conservative) or the former alone, do not suggest a consistent pattern supportive of the use of vignettes in reconciling SRH towards objective measures. This holds for results from either the HOPIT model or the non-parametric approach. In general, either approach produces similar conclusions.

The lack of clear-cut results with regard to the performance of the two approaches to adjusting for reporting might imply that the two fundamental assumptions underlying the use of the methods (response consistency and vignette equivalence) are not tenable, or at least do not hold across all country comparisons. Results presented in Peracchi and Rossetti (2010), Datta Gupta et al. (2010), and Bago d'Uva et al. (2012), do cast some doubt on the validity of the assumptions. In part, this might be due to the vignettes not describing sufficiently well health problems to which respondents are equally able to relate. For the HOPIT approach, the set of characteristics used to model systematic variation in reporting behaviour may be too blunt to sufficiently capture the nuances across individuals in their reporting styles and further investigation of the determinants of reporting behaviour is a potentially fruitful area for future research. Similarly, ranking SRH between vignette ratings in the non-parametric approach may ignore potentially valuable information, particularly where respondents' value their SRH closer to one of the two vignettes between which it is placed. Additional information may be obtained by asking individuals to identify one of the two vignettes that most closely resemble their underlying health status. A further potential explanation for the inconsistent performance of the vignette approach is that the measures used as objective indicators of health do not fully reflect the underlying concept of the domains of health addressed in the self-reports. This is unlikely, however, for the two domains used here where the objective measures would appear a reasonable description of the underlying health problem.

References

Anderson G. (1996) “Nonparametric tests of stochastic dominance in income distributions”, *Econometrica*, 64(5), 1183-1193.

Angelini V., Cavapozzi D., Paccagnella O. (2011a) “Dynamics of reporting work disability in Europe”, *Journal of the Royal Statistical Society, A*, 174, Part 3, 621–638.

Angelini V., Cavapozzi D., Corazzini L., Paccagnella O. (2011b), “Do Danes and Italians Rate Life Satisfaction in the Same Way? Using Vignettes to Correct for Individual-Specific Scale Bias”, *HEDG Working Paper*, 11/20.

Bago d'Uva, T., van Doorlsaer, E., Lindeboom, M. and O'Donnell, O. (2008) “Does reporting heterogeneity bias the measurement of health disparities?”, *Health Economics*, 17 (3), 351-375.

Bago D`Uva T., Lindeboom M., O`Donnell O., van Doorslaer E., (2012) "Slipping anchors? Testing the vignettes approach to identification and correction of reporting heterogeneity", *Journal of Human Resources*, 46, *forthcoming*.

Chevalier A, Fielding A. (2011) “An introduction to anchoring vignettes”, *Journal of the Royal Statistical Society*, A, 174, Part 3, 569-574.

Corrado L., Weeks M. (2010) “Identification Strategies in Survey Response Using Vignettes”, *CWPE Working Paper*, 1031.

Datta Gupta N., Kristensen N., Pozzoli D. (2010) “External Validation of the use of vignettes in cross-country health studies”, *Economic Modelling*, 27, 854-865

Frederiksen, H., Gaist, D., Petersen, H.C. (2002) “Hand grip strength: a phenotype suitable for identifying genetic variants affecting mid- and late-life physical functioning”, *Genetic Epidemiology* 23, 110–122.

Grol-Prokopczyk H., Freese J., Hauser R.M. (2011) “Using Anchoring Vignettes to Assess Group Differences in General Self-Rated Health”. *Journal of Health and Social Behavior*, 52, 246–261.

Hudson E. (2011), “Examining the Effect of Socioeconomic Status on Child Health Using Anchoring Vignettes”, unpublished paper, (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2024685)

Iburg K. M., Salomon, J., Tandon, A. and Murray, C. J. L. (2002) “Cross-country comparability of physician-assessed and self-reported measures of health”. In: *Summary measures of population health: concepts, ethics, measurement and applications* (eds C. J. Murray, J.A. Salomon, C.D. Mathers and A.D. Lopez), pp 433-448. Geneva: The World Health Organization.

Kapteyn A., Salomon, J., van Soest, A. (2007) “Vignettes and self-reports of work disability in the US and the Netherlands”, *American Economic Review*, 97(1), 461–473.

Kapteyn A., Salomon, J., van Soest A. (2009) “Work Disability, Work, and Justification Bias in Europe and the U.S.”, *NBER Working Paper*, No. 15245.

Kapteyn A., Smith J.P., Van Soest A. (2011a), “Are Americans Really Less Happy With Their Incomes?”, *Rand Working paper WR-858*.

Kapteyn A., Smith J.P., Van Soest A., Vonkova H. (2011b), “Anchoring Vignettes and Response consistency”, *Rand Working paper WR-840*.

King G., Wand J. (2007) “Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes”, *Political Analysis*, 15, 46-66.

King G., Murray C. J. L., Salomon J., Tandon A. (2004) “Enhancing the validity and cross-cultural comparability of measurement in survey research”, *American Political Science Review*, 98(1), 184-91.

Kolmogorov A. (1933) "Sulla determinazione empirica di una legge di distribuzione", Gionale.dell`Istituto Italiano degli.Attuari, 4, 83.

Kristensen, N., Johansson, E. (2008) “New evidence on cross-country differences in job satisfaction using anchoring vignettes”, *Labour Economics*, 15, 96–117

Murray C. J. L., Ozaltin E., Tandon A., Salomon J., Sadana R., Chatterji S. (2003) “Empirical evaluation of the anchoring vignettes approach in health surveys” In *Health systems performance assessment: debates, methods and empiricism* (eds C.J.L. Murray and D.B. Evans), pp 369-399. Geneva: World Health Organisation

Paccagnella O. (2011) “Anchoring vignettes with sample selection due to non-response”, *Journal of the Royal Statistical Society*, A, 174, Part 3, 665–687.

Peracchi F., Rossetti C. (2009) “Gender and regional differences in self-rated health in Europe,” CEIS Working Paper No. 142.

Peracchi F., Rossetti C. (2010) “The heterogeneous thresholds ordered response model: Identification and inference”, *EIEF Working Paper 10/12*.

Pfarr C., Schmid A., Schneider U. (2011) “Reporting Heterogeneity in Self-Assessed Health among Elderly Europeans: The Impact of Mental and Physical Health Status”, *Discussion Paper 02-11, University of Bayreuth*.

Puentes Rosas E., Gómez Dantés, O. Garrido Latorre F. (2006) “Trato a los usuarios en los servicios públicos de salud en México”, *Rev Panam Salud Publica*., 9(6), 394–402.

Rantanen T., Masaki K., Foley D., Izmirlian G., White L., Guralnik J.M. (1998) “Grip strength changes over 27 years in Japanese-American men”, *Journal of Applied Physiology*, 85, 2047–2053.

Rice N., Robone S., Smith P.C. (2011) “Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness”, *European Journal of Health Economics*, 12 (2), 141-162.

Rice N., Robone S., Smith P.C. (2012) “Vignettes and health system responsiveness in cross-country comparative analysis”, *Journal of the Royal Statitistical Society (A)*, forthcoming.

Rossi P. E., Gilula Z.,Allenby G.M. (2001) “Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach”, *Journal of the American Statistical Association*, 96(453), 20-31.

Salomon J., Tandon A., Murray C. J. L., World Health Survey Pilot Study Collaborating Group (2004) “Comparability of self-rated health: Cross sectional mutli-country survey using anchoring vignettes”, *British Medical Journal,* 328(258).

Sirven N., Santos-Eggimann B., Spagnoli J. (2008) “Comparability of Health Care Responsiveness in Europe Using anchoring vignettes from SHARE”, *IRDES working paper*

van Soest A., Delaney L., Harmon C., Kapteyn A., Smith J.P. (2011), “Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions”, *Journal of the Royal Statistical Society, A,* 174, Part 3, 575–595.

Stoline M.R., Ury H.A. (1979) “Tables of the Studentised Maximum Modulus Distribution and an Application to Multiple Comparisons Among Means”, *Technometrics,* 21, 87-93.

Smirnov N.V. (1948) “Tables for estimating the goodness of fit of empirical distributions", *Annals of Mathematical Statistic*, 19, 279.

Tandon A., Murray C. J. L., Salomon J. A., King G. (2003) “Statistical models for enhancing cross-population comparability”, In *Health systems performance assessment: debates, methods and empiricism* (eds C.J.L. Murray and D.B. Evans), pp 727-746. Geneva: World Health Organisation.

Valentine N.B., Ortiz J.P., Tandon A., Kawabata K., Evans D.B., Murray C.J.L. (2003) Patient Experiences with Health Services: Population Surveys from 16 OECD Counties. In *Health systems performance assessment: debates, methods and empiricism* (eds C.J.L. Murray and D.B. Evans), pp 643 – 652. Geneva: World Health Organisation.

Wand J., King G., Lau O. (2011) “Anchors: Software for Anchoring Vignette Data” *Journal of Statistical Software, 42(3).*

Yalonetzky G. (2013) “Stochastic dominance with ordinal variables: conditions and a test”, *Econometric Reviews, forthcoming*

Table 1: Distribution of self reported health across countries

Mobility problems	Belgium	France	Germany	Greece	Italy	The Netherlands	Spain	Sweden
none	55.4%	67.2%	46.3%	74.2%	58.6%	57.7%	52.6%	38.1%
mild	26.2%	15.1%	27.0%	15.5%	20.3%	25.3%	19.9%	38.6%
moderate	12.5%	13.6%	19.1%	5.3%	11.0%	11.3%	17.7%	18.0%
severe	4.5%	3.6%	7.2%	3.6%	7.2%	4.3%	8.7%	4.6%
extreme	1.4%	0.5%	0.4%	1.4%	2.8%	1.3%	1.1%	0.8%

Memory problems	Belgium	France	Germany	Greece	Italy	The Netherlands	Spain	Sweden
none	33.8%	38.8%	44.6%	52.6%	41.8%	42.1%	44.2%	56.1%
mild	45.4%	35.8%	36.1%	31.4%	35.0%	48.4%	24.0%	21.9%
moderate	19.2%	21.6%	15.7%	13.4%	16.1%	6.8%	22.1%	12.5%
severe	1.4%	3.5%	3.6%	2.6%	5.4%	1.9%	9.5%	8.7%
extreme	0.2%	0.2%	0.0%	0.0%	1.6%	0.8%	0.2%	0.7%

Note: For self-reported mobility individuals are asked “Overall in the last 30 days, how much of a problem did you have with moving around?” while for memory they are asked “Overall in the last 30 days how much difficulty did you have with concentrating or remembering things

Table 2: Vignette ratings for *mobility* and *memory* (Vignette 1), by country

Mobility problems	Belgium	France	Germany	Greece	Italy	The Netherlands	Spain	Sweden
	%	%	%	%	%	%	%	%
none	1.78	2.59	1.19	0.56	4.38	1.69	0.65	0.00
mild	4.26	8.37	7.16	5.01	12.21	2.44	5.19	2.95
moderate	35.52	37.15	26.24	22.95	20.51	29.27	25.11	14.50
severe	43.34	47.29	56.66	41.03	51.38	39.21	59.74	58.23
extreme	15.10	4.60	8.75	30.46	11.52	27.39	9.31	24.32

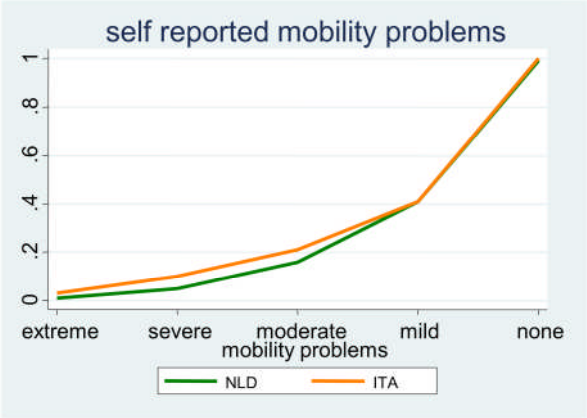
Memory problems	Belgium	France	Germany	Greece	Italy	The Netherlands	Spain	Sweden
	%	%	%	%	%	%	%	%
none	17.97	15.86	23.06	41.31	27.98	22.03	16.67	5.38
mild	64.23	53.82	49.30	39.36	43.35	68.55	38.31	24.45
moderate	15.30	25.73	24.45	15.72	20.18	8.29	33.33	45.72
severe	2.31	3.88	2.58	3.62	7.57	1.13	11.47	23.96
extreme	0.18	0.71	0.60	0.00	0.92	0.00	0.22	0.49

Note: The description of the vignette for mobility is: “Tom has a lot of swelling in his legs due to his health condition. He has to make an effort to walk around his home as his legs feels heavy. Overall in the last 30 days, how much of a problem did Tom have with moving around?”. For memory, the description for the vignette is as follows: “Lisa can concentrate while watching TV, reading a magazine or playing a game of cards or chess. Once a week she forgets where her keys or glasses are, but finds them within five minutes. Overall in the last 30 days, how much difficulty did Lisa have with concentrating or remembering things?”.

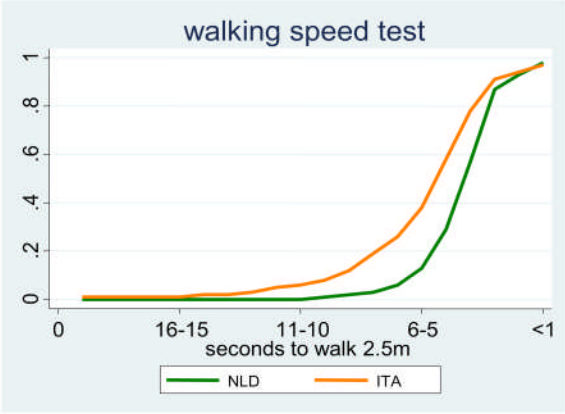
Table 3: Average of socio-demographic variables within countries

	Belgium	France	Germany	Greece	Italy	The Netherlands	Spain	Sweden
n. obs	554	826	497	718	428	529	462	394
age (years)	63.5	64.5	63.4	61.8	63.5	62.3	64.7	63.8
gender (men)	44%	43%	43%	46%	44%	48%	42%	48%
household income (1000 euros)	40.7	45.5	51.8	29.2	29.7	49.8	39.9	49.0
education (years)	10.7	9.4	13.4	9.9	7.2	12.3	7.2	11.2
married or cohabiting	74%	69%	78%	74%	73%	84%	72%	78%

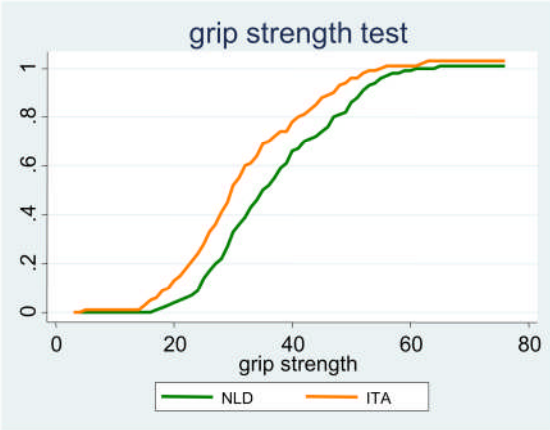
Figure 1: Pairwise comparisons, The Netherland vs Italy



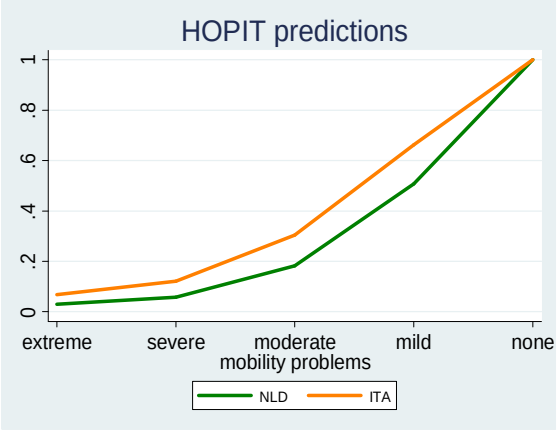
No stochastic dominance



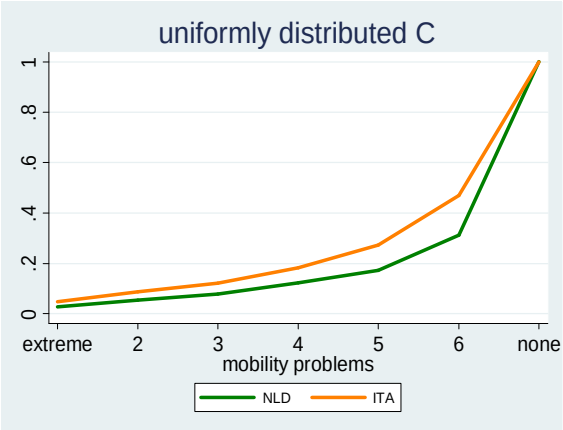
Stochastic dominance



Stochastic dominance



Stochastic dominance



Stochastic dominance

Note:In the graph related to the HOPIT predictions, respondents in Italy are assumed to have the same reporting behaviour of respondents in The Netherlands

Table 4: Kolmogorov Smirnov stochastic dominance test (D statistics and p-values).

MOBILITY	Self reported mobility			Walking speed test			Grip strength test		
	D	p-value	Dominating country	D	p-value	Dominating country	D	p-value	Dominating country
The Netherlands vs Italy	0.050	0.544	none	0.340	0.000	The Netherlands	0.216	0.000	The Netherlands
The Netherlands vs Belgium	0.022	1.000	none	0.187	0.000	The Netherlands	0.107	0.004	The Netherlands
The Netherlands vs Sweden	0.204	0.000	The Netherlands	0.073	0.307	none	0.081	0.103	none
Italy vs Greece	0.160	0.000	Greece	0.085	0.307	none	0.075	0.105	none
Italy vs Belgium	0.046	0.639	none	0.167	0.000	Belgium	0.130	0.001	Belgium

MEMORY	Self reported memory			Memory test		
	D	p-value	Dominating country	D	p-value	Dominating country
Germany vs France	0.06	0.197	none	0.224	0.000	Germany
Greece vs Belgium	0.188	0.000	Greece	0.032	0.891	none
Sweden vs Greece	0.068	0.162	none	0.089	0.029	Sweden
Sweden vs France	0.173	0.000	none	0.172	0.000	Sweden
Germany vs Italy	0.038	0.883	none	0.354	0.000	Germany
Spain vs Italy	0.087	0.059	none	0.137	0.000	Italy
Spain vs Belgium	0.004	0.110	none	0.353	0.000	Belgium
Germany vs The Netherlands	0.099	0.011	The Netherlands	0.060	0.305	none
Spain vs France	0.065	0.147	none	0.274	0.000	France
Italy vs France	0.033	0.912	none	0.137	0.000	France
Italy vs Belgium	0.080	0.078	none	0.216	0.000	Belgium
France vs Belgium	0.050	0.346	none	0.086	0.015	Belgium

Note: the table reports two-sample Kolmogorov-Smirnov tests for equality of distribution functions.

Table 5: Anderson`s (1996) test statistics.

MOBILITY	Hopit	C (Uniformly allocated intervals)	C (Censored oprobit predictions)
	test statistics	test statistics	test statistics
The Netherland vs Italy (The Netherlands dominates)	-2.247 -3.256 -4.117 -4.471 0.000	5.095 3.839 2.636 2.295 2.083 1.679 0.000	5.094 3.394 2.005 1.565 1.737 1.590 0.000
The Netherland vs Belgium (The Netherlands dominates)	-1.188 -1.919 -2.619 -2.978 0.000	3.786 1.758 1.621 1.335 1.166 0.761 0.000	3.987 1.866 1.821 1.618 1.257 0.570 0.000
The Netherland vs Sweden (The Netherlands dominates)	-1.713 -3.152 -4.500 -5.117 0.000	6.833 3.236 1.909 0.222 -0.067 -2.117 0.000	6.955 3.211 2.122 0.840 0.218 -2.117 0.000
Italy vs Greece (Greece dominates)	2.649 4.870 6.335 7.918 0.000	-7.203 -6.085 -4.201 -3.327 -3.195 -2.551 0.000	-7.105 -5.417 -3.440 -2.448 -2.745 -2.794 0.000
Italy vs Belgium (no country dominates)	0.928 1.574 1.948 2.072 0.000	-1.564 -2.237 -1.154 -1.083 -1.031 -1.004 0.000	-1.411 -1.597 -0.368 -0.118 -0.567 -0.943 0.000

Note: The column for the Hopit model corresponds to the five categories of self-reported health, while the columns for the C estimator correspond to the seven possible values assumed by C. Under the null hypothesis of lack of first order stochastic dominance, the categories in the Hopit column and the values in the C column are distributed as a studentized maximum modulus distribution. With 5 and 7 multiple comparisons (and infinite degrees of freedom) the 5% critical value of the distributions in the Hopit column and the C column are 2.800 and 3.031 respectively.

Table 5: (cont.)

MEMORY	Hopit	C (Uniformly allocated intervals)	C (Censored oprobit predictions)		Hopit	C (Uniformly allocated intervals)	C (Censored oprobit predictions)		Hopit	C (Uniformly allocated intervals)	C (Censored oprobit predictions)
	test statistics	test statistics	test statistics		test statistics	test statistics	test statistics		test statistics	test statistics	test statistics
Germany vs France (no country dominates)	-0.343 -0.547 -0.788 0.051 0.000	0.053 0.670 1.244 1.848 2.249 1.007 0.000	0.300 1.006 1.443 1.596 2.282 1.190 0.000	Germany vs Italy (Germany dominates for C unif. distr. only)	-0.227 -1.219 -2.242 -2.032 0.000	1.181 1.503 1.345 2.597 3.087 1.653 0.000	1.319 1.872 1.385 2.194 2.979 1.703 -0.005	Spain vs France (no country dominates)	-0.225 -0.698 -0.686 -1.192 0.000	2.497 0.341 0.208 -0.102 -0.119 2.204 0.000	2.551 0.219 0.261 -0.516 -0.282 2.288 0.000
Greece vs Belgium (no country dominates)	0.199 0.500 0.805 1.046 0.000	0.910 -1.308 -0.803 -1.895 -1.587 -2.036 0.000	0.908 -1.300 -0.916 -2.181 -1.807 -2.097 0.000	Spain vs Italy (Spain dominates for C only)	-0.599 -1.558 -1.982 -2.026 0.000	3.296 1.207 0.457 0.942 1.125 2.556 0.000	3.254 1.198 0.380 0.500 0.966 2.520 0.000	Italy vs France (no country dominates)	0.423 0.951 1.501 1.156 0.000	-1.253 -1.030 -0.313 -1.175 -1.411 -0.942 0.000	-1.165 -1.115 -0.165 -0.958 -1.263 -0.834 0.000
Sweden vs Greece (Sweden dominates)	-3.410 -9.035 -10.469 -10.164 0.000	11.460 7.321 5.611 3.752 3.381 3.291 0.000	11.643 7.134 5.810 4.163 3.731 3.334 0.000	Spain vs Belgium (clear absence of dominance only for Hopit)	-0.099 0.088 0.251 -0.012 0.000	1.026 -0.568 -1.879 -1.077 -0.665 4.286 0.000	1.326 -1.237 -1.546 -0.742 0.000 4.265 0.000	Italy vs Belgium (no country dominates)	0.491 1.417 2.282 2.457 0.000	0.693 -1.929 -1.589 -2.713 -2.274 -1.510 0.000	0.827 -1.323 -1.060 -1.988 -1.912 -1.478 0.000
Sweden vs France (Sweden dominates)	-3.825 -8.928 -11.692 -12.279 0.000	2.587 3.585 4.499 6.947 7.949 10.801 0.000	2.587 3.936 3.901 7.289 7.347 10.841 0.000	Germany vsThe Netherlands (The Netherlands dominates)	0.413 1.873 3.680 4.272 0.000	0.725 -3.903 -4.654 -3.005 -1.493 -1.886 0.000	0.991 -2.315 -3.837 -2.699 -1.177 -1.748 0.000	France vs Belgium (no country dominates)	0.177 0.543 0.908 0.911 0.000	0.000 -0.657 -2.199 -1.650 -1.145 2.322 0.000	-1.191 -0.744 -1.893 -1.276 -0.404 2.305 0.000

Table 6: Summary table for stochastic dominance

Mobility					C	
Country comparison	Self reported mobility	Walking speed test	Grip strength test	Hopit predictions	Uniformly allocated intervals	Censored oprobit predictions
The Netherland vs Italy	NO SD	**SD**	**SD**	**SD**	**SD**	**SD**
The Netherland vs Belgium	NO SD	**SD**	**SD**	**SD**	**SD**	**SD**
The Netherland vs Sweden	**SD**	NO SD	NO SD	**SD**	**SD**	**SD**
Italy vs Greece	**SD**	NO SD	NO SD	**SD**	**SD**	**SD**
Italy vs Belgium	NO SD	**SD**	**SD**	NO SD	NO SD	NO SD

Note: SD = stochastic dominance, NO SD = absence of stochastic dominance

Memory				C	
Country comparison	Self reported mobility	memory test	Hopit predictions	Uniformly allocated intervals	Censored oprobit predictions
Germany vs France	**SD**	NO SD	NO SD	NO SD	NO SD
Greece vs Belgium	**SD**	NO SD	NO SD	NO SD	NO SD
Sweden vs Greece	NO SD	**SD**	**SD**	**SD**	**SD**
Sweden vs France	NO SD	**SD**	**SD**	**SD**	**SD**
Germany vs Italy	NO SD	**SD**	NO SD	**SD**	NO SD
Spain vs Italy	NO SD	**SD**	NO SD	**SD**	**SD**
Spain vs Belgium	NO SD	**SD**	NO SD	NO SD/**SD**	NO SD/**SD**
Germany vs The Netherlands	**SD**	NO SD	**SD**	**SD**	**SD**
Spain vs France	NO SD	**SD**	NO SD	NO SD	NO SD
Italy vs France	NO SD	**SD**	NO SD	NO SD	NO SD
Italy vs Belgium	NO SD	**SD**	NO SD	NO SD	NO SD
France vs Belgium	NO SD	**SD**	NO SD	NO SD	NO SD

Note: SD = stochastic dominance, NO SD = absence of stochastic dominance. These are supported both by respective test statistics and graphical inspection. NO SD/SD indicates situations where the Anderson (1996) test suggests stochastic dominance, however a visual inspection of the graph suggests otherwise.