DISCUSSION PAPER SERIES

IZA DP No. 3989

External Validation of the Use of Vignettes in Cross-Country Health Studies

Nabanita Datta Gupta
Nicolai Kristensen
Dario Pozzoli

February 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

External Validation of the Use of Vignettes in Cross-Country Health Studies

Nabanita Datta Gupta
Danish National Centre for Social Research and IZA

Nicolai Kristensen
Danish Institute of Governmental Research

Dario Pozzoli
Aarhus School of Business

Discussion Paper No. 3989
February 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3989
February 2009

ABSTRACT

External Validation of the Use of Vignettes in Cross-Country Health Studies[*]

Cross-country comparisons of subjective assessments are rendered difficult if not impossible because of sub-population specific response style. To correct for this, the use of vignettes has become increasingly popular, notably within cross-country health studies. However, the validity of vignettes as a means to re-scale across sample populations critically rests on the assumption of "response consistency" (RC): that vignettes and self-assessments are evaluated on the same scale. In this paper, we seek to test this assumption by applying objective measures of health along with subjective measures and vignettes. Our results indicate that the assumption of RC is not innocuous and that our extended model relaxing this assumption improves the fit and significantly changes the cross-country rankings of health vis-à-vis the standard Chopit model.

JEL Classification: C25, I10

Keywords: cross-country health comparison, vignettes, subjective and objective measures

Corresponding author:

Nabanita Datta Gupta
The Danish National Centre for Social Research
Herluf Trolles Gade 11
DK-1052 Copenhagen K
Denmark
E-mail: ndg@sfi.dk

[*] We thank Teresa Bago d'Uva, Hendrik Juerges, Gary King and seminar participants at the CIM 2008 Workshop in Aarhus, at the 2008 Health Econometrics Workshop in Milano (especially chair of session Bill Greene) and at AKF (Copenhagen) for helpful suggestions. The usual disclaimer applies.

1 Introduction

Cross-country comparisons of subjective responses frequently appear in empirical studies. Direct comparisons may be misleading, however, due to country-specific response style. In order to correct for systematic differences in response scales across subpopulations, King et al. (2004) suggested the use of anchoring vignettes which permit identification of country-specific threshold parameters in ordered probit models. The method of anchoring vignettes has subsequently been applied to achieve valid cross-country comparisons in various disciplines including political science (King et al., 2004), medicine (Salomon et al., 2004), work disability (Kapteyn et al., 2007), job satisfaction (Kristensen and Johansson, 2008), life satisfaction (Angelini et al., 2008) and notably health (see Bago d'Uva et al. (2008) and the references therein).

Since the seminal paper of King et al. (2004), some attention has been devoted to methodological improvements of the vignette approach. For instance, Javaras and Ripley (2007) introduce a multidimensional model, which allows Likert type data to reflect not just attitudes but also response style. Hopkins and King (2008) focus on the importance of question ordering and wording within the vignettes framework.

In this paper, we seek to test one of the fundamental assumptions underlying the vignette approach. The validity of the vignette approach hinges on important assumptions including the assumption about "response consistency". Response consistency implies that individuals use the same response categories for their subjective assessment (e.g. of own health) as the categories used for the hypothetical scenarios presented to them in vignettes.[1] A violation of this could arise, for example, in settings where individuals overplay their own health problems because they have a financial incentive to report themselves ill for the purpose of gaining windfall disability benefits (e.g. Kerkhofs and Lindeboom, 1995) but do not face similar incentives when it comes to rating the health problems of the vignette person.

In this paper, we evaluate the use of the vignettes as a means to appropriately re-scale self-assessments and obtain valid cross-country comparisons. We seek to test whether response consistency is a tenable assumption. In order to do this, we use cross-country health data, which include self-assessment of health (self-reported work disability) and vignettes, but also an objective measure of health (measured hand grip strength). Including this objective measure allows us to free up the assumption of response consistency.

A similar approach was first suggested by van Soest et al. (2007) in a study of self-assessments of drinking behavior among students in Ireland. The method applied in this paper follows their approach, but our application has several advantages compared to van Soest et al (2007).

Firstly, they use self-reported drinking as their objective measure and compare this to a self-assessment of how the respondent characterizes his or her drinking pattern over the last year. Self-reported drinking is at best semi-objective and bias can easily

[1] The other main assumption is that of Vignette Equivalence. This means that the domain levels represented in each vignette are understood in the same way by all respondents, irrespective of their country of residence or other sociodemographic variables. In this paper, we do not seek to test this assumption. While Murray et al. (2003) do report some systematic differences in how individuals rank vignettes by age, education and gender, the differences do not appear to be big enough to reject this assumption, therefore our focus on response consistency.

arise as a result of measurement error due to norms or social desirability (respondents report what is politically correct). Secondly, van Soest et al. (2007) only have two sub-populations: Irish and non-Irish students. However, it appears that the group of non-Irish students can be a blend of students from countries with higher levels than Irish students for what constitutes, say, ”severe drinking” and students from countries with lower levels of ”severe drinking” than Irish students.

Our application avoids these potential shortcomings and it therefore seems natural to assume that it is better equipped to assess whether response consistency is a tenable assumption. By using data across eight countries, we seek to validate the vignette method using the type of application where it has been most predominant, namely within Health Economics. Hence, this paper may also be seen as a sensitivity check of the burgeoning literature that uses vignettes to perform valid cross-country comparisons of health.

We find that the model log likelihood improves considerably when we do not impose the response consistency assumption. Model comparisons using both the Akaike and the Bayesian information criterion support a specification not imposing response consistency and vignette corrected response scales. A robustness check using alternative objective and self-reported measures of health confirms the main results.

The paper is organized as follows. In the next section ,we very briefly describe the data set. Next, in Section 3, we present an extended version of the vignette Chopit model, which we name Ochopit (objective-extended Chopit). In Section 4, we present the main results and the sensitivity analysis results. Section 5 discusses the findings and Section 6 concludes.

2 Data

Data for the empirical analysis comes from Release 2 of the Survey of Health, Ageing and Retirement in Europe (SHARE): it is a multi-disciplinary and cross-national data set which contains information on the individual life circumstances of, in principle, all eligible members of about 18,000 households. A household is eligible for participation in SHARE if at least one household member was born in or before 1954. An individual member of the household is eligible for interview if she or he, or his/her partner, was born in or before 1954.[2] Release 2 of the SHARE data was gathered in 2004 and consists of probability samples drawn from each participating country.[3] The survey contains information on over 26,000 individuals. SHARE covers 11 countries: Austria, Belgium, Denmark, France, Germany, Greece, Italy, the Netherlands, Spain, Sweden and Switzerland. The data set is designed after the Health and Retirement Study (HRS) and the English Longitudinal Study of Ageing (ELSA). The data include information about respondents' health overall as well as six specific domains of health (breath, pain, mobility, work disability, depression and memory). Vignettes have been collected for each of these domains as well.

An important feature of this study is to compare the self-reported health measure with an objective one. In order for this comparison to be valid, we mainly focus on self-

[2] SHARE contains information on a few respondents younger than 50 years of age. These spouses or partners of age-eligible respondents are omitted from the analysis.

[3] The data from Belgium and France were collected in 2004/2005.

assessed work disability which we tie to all the vignette questions on the same health domain. While many previous studies have found self-reported general health to be a good summary measure of individuals' underlying health and an accurate predictor of future mortality, we do not apply it here as the self-reported measure because the response categories on the vignettes relating to general health (1=none, 2=mild etc.) are not sufficiently close to the SRH categories (1=excellent, 2=very good) as required by response consistency. Instead, we focus on work disability where the vignette and self-reported categories match closely (1=none, 2=mild, 3=moderate, 4=severe and 5=extreme) and which is also the most crucial health variable for policy purposes given the increasing numbers of disability insurance claimants in welfare state countries. If the high costs to society of the lost work effort of work-disabled individuals has to be averted, it is necessary to know whether the observed differences in self-reported disability across a set of European countries are due to differences in reporting behavior (Kapteyn et al. 2007 find evidence of this type of heterogeneity comparing the Netherlands to the US). Furthermore, if individuals in generous welfare state countries have a financial incentive to misreport their disability status but not the disability status of a vignette individual, this could result in a violation of response consistency. We supplement the self-assessed work disability and the vignette-assessed work disability with the objective measure of the respondents' grip strength. We do this because grip strength in middle age has been found to predict rather closely late-life disability degree and mortality (Frederiksen et al., 2002, Rantanen et al., 1998). This reduces the sample size to about 4,000 individuals, since vignettes are available only for a small subsample.

The explanatory variables are selected in order to keep the model relatively parsimonious. For this reason, we follow Bago d'Uva et al. (2008) and include age (represented by four categories), log of household income (normalized by household size), body mass index (here captured by separate parameters for height and weight for reasons discussed below), as well as indicators for education, gender, whether the respondent is employed or not at the date of the survey and for whether or not the respondent lives in an urban area and of course, country dummies for seven of the eight countries for which we have vignette information (Belgium, Germany, Sweden, the Netherlands, Spain, Italy, France and Greece).[4] In contrast to Juerges (2007) and Peracchi and Rossetti (2008), we do not include the reported health conditions among the covariates, as they could potentially be measured with error and may not be comparable across countries. If there are systematic differences in reporting health conditions across countries, this might bias our results (Kapteyn et al., 2007). Frequencies and sample means are reported in Table 1, which reveals that most observations adhere to France and Germany (29 and 26 percent) while very few respondents come from Sweden and Belgium (about 2 percent each).

[Table 1 about here].

Figure 1 shows the age-sex standardized distributions of self-reported work disability across the eight countries, i.e. the health distribution if each country had the same

[4] The vignette subsample does not include the following countries: Austria, Denmark and Switzerland.

age and sex distribution of individuals aged 50 and over (Juerges 2007). Countries are ordered by the fraction of respondents who say they have either none or only a mild work impairment which limits the amount of work they can do. According to the self-reports, the healthiest respondents live in Greece and the Netherlands. The least healthy respondents live in Spain and Germany.

[Figure 1 about here].

Figure 2 shows the age-sex standardized distribution of our objective measure of health, categorised into four age gender-specific quartiles. Again, countries are ordered by the fraction of respondents who have high or above average grip strength. According to grip strength, the healthiest respondents live in Germany, the Netherlands and Sweden while the least healthy live in Mediterranean countries (Greece, Italy and Spain). Hence, the ranking of the countries by self-reported health is somewhat at odds with the ranking by grip strength. This could be related to systematic differences in response scales across sub-populations but could also indicate that grip strength is not really a good objective measure for cross-country comparison. For instance, if Germans perform better when grip strength is applied, this could simply reflect that Germans on average are bigger (stronger) than the average person in Southern European countries. In order to rule out this potential alternative explanation, we include height and weight among our explanatory variables.

3 Methodology

The model presented here follows van Soest et al. (2007) and extends the Chopit model formulated by King et al. (2004) by allowing the threshold parameters in the self-assessment equation to differ from the thresholds in the vignette equation. In other words, we avoid the potentially disputable assumption of response consistency. Subsequently, we can test whether threshold parameters indeed are significantly different from each other by comparing with estimates from a model where the assumption of response consistency is maintained.

In order to identify two sets of threshold parameters, we need more information than self-assessments and answers to vignettes can provide. This is obtained by employing objective measures of hand grip strength.

Model for Subjective Self-Assessment The self-assessment measure of work disability (denoted Y_{si} for respondent i self-assessment s) is based on answers to the following question:

> Do you have any impairment or health problem that limits the kind or amount of work you can do? (1=none, 2=mild, 3=moderate, 4=severe, 5=extreme).

Relatively few reply ”severe” or ”extreme” so we combine these two categories with the ”moderate” one and continue with three categories.

The subjective answer is assumed to reflect an underlying continuous latent measure of health but will also mirror individual thresholds and an error term that captures the inherent noise related to subjective assessments. The model therefore becomes

$$Y^*_{si} = X_i\beta_s + \xi_{si} \tag{1}$$

$$Y_{si} = j \;\text{ if } \tau^{j-1}_{si} < Y^*_{si} \leq \tau^j_{si}, \quad j = 0, .., 3.$$

X_i includes a set of covariates describing the respondent and ξ_{si} denotes the error term (including unobserved heterogeneity) assumed to be i.i.d. normally distributed with variance, σ^2_ξ.

The thresholds τ^j_{si} are modeled as

$$\tau^0_{si} = -\infty,\; \tau^1_{si} = \gamma^1_s X_i,\; \tau^j = \tau^{j-1} + \exp(\gamma^j_s X_i),\; j = 2 \text{ and } \tau^3_{si} = \infty.$$

It is important to note that these response scales may differ across respondents, thus introducing Differential Item Functioning (DIF, King et al., 2004), i.e. the fact that there are differences in response scales.

Model for Vignettes The vignettes describe hypothetical persons in specific situations that reveal aspects of the hypothetical person's health. The respondents are asked to rank these vignette-persons' health on a similar five point scale (also collapsed into three points). As the same vignettes are used across all countries, the answers can be used to re-scale to adjust for DIF (see Appendix A for the exact phrasing of the vignettes).

Answers to the vignettes are also modeled as an ordered latent variable and can be written as

$$Y_{li}^{*} = \theta_l + \xi_{li} \tag{2}$$

$$Y_{li} = j \;\text{ if } \tau_{li}^{j-1} < Y_{li}^{*} \leq \tau_{li}^{j}, \quad j = 0, .., 3.$$

θ_l denotes vignette-specific parameters and ξ_{li} denotes the error term (including unobserved heterogeneity) assumed to be i.i.d. normally distributed with variance, σ_{ξ}^{2}, normalized to 1.

Similarly, the thresholds τ_{vi}^{j} for each of the v vignettes, v=(1,2,3) are modeled as

$$\tau_{vi}^{0} = -\infty,\; \tau_{vi}^{1} = \gamma_v^1 X_i,\; \tau_{vi}^{j} = \tau_{vi}^{j-1} + \exp(\gamma_v^j X_i),\; j = 2 \text{ and } \tau_{vi}^{3} = \infty.$$

response consistency would entail that

$$RC : \gamma_v^j = \gamma_s^j,\; j = 1, 2. \tag{3}$$

Equation 3 imposes the key assumption and it is the validity of this constraint we seek to evaluate in this paper.

Model for Objective Measure Response consistency is normally necessary for identification but with the availability of an objective measure of general health (hand grip strength), we can allow $\gamma_v^j \neq \gamma_s^j$.[5] We categorize grip strength as an ordered variable so that we can model it as an ordered probit

$$Y_{oi}^* = X_i \beta_o + \xi_{oi} \tag{4}$$

$$Y_{oi} = j \;\text{ if } \tau_o^{j-1} < Y_{oi}^* \leq \tau_o^j, \quad j = 0, .., 3.$$

where $\tau_o^0 = -\infty$ and $\tau_o^3 = \infty$. Note that the objective thresholds are constant across individuals and are chosen as the gender- and age-specific quartiles across the empirical distribution of grip strength. The error term ξ_{oi} is independent of X_i and ξ_{li}.[6] Again, following van Soest et al. (2007) we impose a one factor assumption which states that subjective and objective measures are driven by the same latent health (true health) process, i.e., that

$$OF : \beta_s = \beta_o. \tag{5}$$

It is assumed that (ξ_{oi}, ξ_{si}) is bivariate normally distributed and hence we allow ξ_{oi} to be correlated with ξ_{si}. The one factor assumption is necessary for the objective measure to yield identification when RC is not imposed. It may appear questionable to throw away information by creating groupings of an otherwise continuous objective measure, and then use a statistical model to infer back to the "unobserved" continuous measure. The reason for grouping is that this enables us to impose the one factor

[5] See Appendix B for details about how the test for hand grip strength was carried out.

[6] In this case, we normalize both the variance of ξ_{si} and the variance of ξ_{si} to one.

assumption meaningfully whereas we would not be able to impose this, and obtain identification, if we applied a linear measure. In addition, it would also be a very difficult model to solve.

The question may be raised why, when a continuous objective measure of health is available, we need the vignette method. The answer here is that while grip strength is an objective measure, it is not a golden standard for the truth. Hence, combining the different sources of information seems well worthwhile.

The Combined Ochopit Model The likelihood for the combined model where both self-assessments, vignettes and the objective measure enter can be written as the product of a bivariate ordered probit for self-assessment and the objective measure and an ordered probit model for the vignettes.

The likelihood for the self-assessment and objective components reads

$$\begin{aligned} L_{so} &= \Pi_{i=1}^{N}\Pi_{j=1}^{3}\Pi_{k=1}^{3}\Pr\left(subjective = j, objective = k\right) = \qquad (6)\\ &\Phi_2[c_{1j} - x_{1i}^{'}\beta_1,\ c_{2k} - x_{2i}^{'}\beta_2,\ \rho] - \\ &\Phi_2[c_{1j-1} - x_{1i}^{'}\beta_1,\ c_{2k} - x_{2i}^{'}\beta_2,\ \rho] - \\ &\Phi_2[c_{1j} - x_{1i}^{'}\beta_1,\ c_{2k-1} - x_{2i}^{'}\beta_2,\ \rho] + \\ &\Phi_2[c_{1j-1} - x_{1i}^{'}\beta_1,\ c_{2k-1} - x_{2i}^{'}\beta_2,\ \rho], \end{aligned}$$

where Φ_2 is the bivariate standard normal cumulative distribution function, ρ is the correlation between error terms from the self-assessed and the objective measures and the product is estimated over N individuals.

The likelihood component for the vignettes reads

$$L_v = \Pi_{i=1}^{N} \Pi_{l=1}^{9} \Pi_{k=1}^{3} [\Phi(\tau_l^k) - \Phi(\tau_l^{k-1})]^{\mathbf{I}(v_{l,j}=k)} \tag{7}$$

The joint likelihood therefore becomes

$$L = L_{so} \times L_v. \tag{8}$$

We name this model, first formulated by van Soest et al. (2007), Ochopit, in short for Objective-extended Chopit. The maximization routine is written in stata.

4 Results

4.1 Main Results

We estimate three models: a standard ordered probit model, a Chopit model using vignettes and an Ochopit model using both vignettes and objective measures and relaxing the response consistency assumption.[7]

[Tables 2, 3 and 4 about here].

In the ordered probit model, i.e., without the DIF correction, the probability of

[7] We always take account of the complex survey design. The potentially biasing effects on descriptive statistics and estimates are accounted for by using sampling weights in the data set: these weights being approximately equal to the inverse of the probability of selection of each individual into the sample. We use calibrated weights for the main and vignette samples together to compensate for unit nonresponse to some extent.

reporting work disability is found to be significantly higher for the reference group, respondents aged 65 and higher, and decreases significantly with household income. The probability of work disability is also higher at low levels of education vis-a-vis higher educational levels and lower for respondents who are employed at the date of the survey and who are living in urban areas. These parameters generally enter with expected and similar signs across all three sets of parameter estimates, although the Ochopit model yields insignificant parameters for "above average" education and the urban area indicator. We also note that height enters with a negative parameter estimate in all three models although with varying significance while weight generally enters with a positive parameter, with exception of the Ochopit model.

Women are found to be significantly more likely to report work disability. However, the DIF-corrected results imply that health is not significantly different between men and women (Chopit). When we correct for DIF and relax the response consistency assumption (Ochopit), we find that the female dummy is underestimated in the oprobit model since a higher initial threshold is used by women.

In the threshold equation we control for all the covariates included in the main equation. When response consistency is imposed, the "low education" dummy enters with a positive sign and significantly in the equation for the first threshold parameter, τ_1. This means that according to the Chopit model, the lowest educated have a higher standard for what constitutes the second-lowest level of work disability compared to the lowest (as the issue is whether one is above or below the first threshold). For the other Chopit threshold-equation, education is not significant. This result appears counter-intuitive. Low educated generally have a higher tendency of manual work, and a small injury could therefore be expected to have a bigger impact on their work ability than a similar injury would have for high educated who generally have less physically demanding work. Indeed, when we relax the RC assumption we find the opposite result in the equations for the lowest threshold (τ_1), though the coefficients are not statistically significant (cf. the right-most column of Table 2, lower panel). This is also true for the age variable. When we impose the RC assumption, we find that younger respondents have a higher initial threshold. This is at odds with the common finding that older respondents tend to have a milder view of their health, i.e. they tend to rate their health as better than otherwise comparable younger respondents (Groot, 2000; Van Doorslaer and Gerdtham, 2003). This is not the case when we relax the RC assumption and impose the OF one, as the age dummies enter with a negative sign and significantly in the equation for the first threshold parameter.

The female indicator does not enter significantly in the Chopit model's threshold equations. This is consistent with the almost identical Ordered probit and Chopit parameter estimates for female in the main equation. In the Ochopit model, on the other hand, females were found to be *more* likely to be work-disabled and we should therefore expect the gender dummy to enter significantly in the threshold equations and to differ between vignettes and self-reported. Indeed, in the self-reported equation (τ_1 and τ_2), females are found to have a higher standard for when to cross these threshold limits for work disability.

The above results could indicate that the response consistency assumption is not very plausible in this application; an interpretation supported by the fact that most of the parameter estimates in the Ochopit vignette threshold-part differ greatly from their corresponding parameter estimates in the Ochopit self-reported threshold part. A Wald test for equality of coefficients reveals that they are significantly different, cf. Table 3.

As far as the country dummies are concerned, although they generally are significant in the threshold equation for the Chopit model (Table 2, the mid columns), the results reveal that the country ranking only differs very little between the Chopit model and the ordered probit model, cf. Table 4. Testing for rank correlation (Kendall's tau), we cannot reject that they have the same order. According to these country rankings, one of the healthiest countries is Greece while Germany is the least healthy. This is at odds with what has been found in Juerges (2007). Applying a generalized ordered probit model to the first release of SHARE, Juerges (2007) computes a cross-country comparable health index and according to this he finds that Germany (Greece) is the healthiest (least healthy) country.

When we relax the assumption of response consistency, the country rankings shuffle around much more, cf. the right-most column of Table 4. We also find that a rank order test rejects equality of country rankings between the Ochopit country ranking and the two other models' country rankings. In addition, the country rankings obtained from the Ochopit model are more consistent with what has been found in Juerges (2007).

The correlation coefficient between the error terms in the self-reported and the objective models is estimated to be about 0.3 and very significant. This is clearly smaller than the estimate of 0.6 found in van Soest et al. (2007) but the high correlation in their study could partly be a reflection of a non-ideal objective measure.

Interestingly, the log likelihood improves a lot when we do not impose the response consistency assumption compared to the chopit model. As the models are non-nested, we cannot use a likelihood ratio test, but AIC and BIC tests indicate that our Ochopit is the preferred model, cf. Table 3.

4.2 Sensitivity Analysis

This section briefly discusses i) the results obtained using the self-assessment and vignette question on *mobility* to examine the sensitivity of the main results to an alternative definition of health; and ii) the results when we relax the one factor assumption and impose response consistency.

As a sensitivity check, the specification now includes an objective measure of mobility (walking speed) instead of grip strength.[8] Walking speed, which declines rapidly with age, is an excellent measure of general mobility. In this case, the sample is relatively small, only about 500 observations in total, and the mean age is very high (almost 79 years), given the walking speed is available only for respondents aged 75 and over or respondents with self-reported mobility limitations.[9] We perform the analysis using this objective measure, despite this small sample size because the one factor assumption, which is a key assumption for the Ochopit model to be valid, seems most

[8] See Appendix B for details about how the test for walking speed was carried out.

[9] Given that the sample size is relatively small, we only include the most relevant explanatory variables both in the main and the threshold equation and exclude household income and whether the respondent is employed at the date of the survey.

likely to hold when walking speed is used as the objective measure.[10]

[Tables 5, 6 and 7 about here].

As in the main analysis, we find that the country ranking differs very little between the Chopit model and the ordered probit model, but when we relax the response consistency assumption, the country ranking shuffles around much more. The results also confirm that the Ochopit model is significantly better than the Chopit and the Ochopit relaxing the one factor assumption according to AIC and BIC, so the response consistency would be rejected under the maintained assumption of one factor, which seems plausible in this case given that the correlation coefficient is estimated to be 0.40.

Next, we estimate the Ochopit model relaxing the one factor assumption and imposing the response consistency assumption. The results of this sensitivity exercise reveal that the log-likelihood is much higher for the initial version of the Ochopit model compared to the Ochopit log-likelihood where OF is substituted by RC. Again, this supports our approach. We discuss the choice of OF versus RC further in the following section.

[10] We also tried to estimate alternative models where the use the self-assessments and vignette questions on the full set of health domains (pain, mobility, sleeping problems, shortness of breath, concentration problems, depression and work limitations) and grip strength. However, the estimated correlation between unobservables was very low, around 10%. We interpreted this result as an indication that the one factor assumption is not very plausible in these cases.

5 Discussion

5.1 Economic Incentives for Misreporting?

The examples of disability and mobility above show that country rankings change when the response consistency assumption is relaxed. Our motivating example of when a violation of response consistency could potentially arise was if individuals from countries with social transfers were more likely to self-report disability (i.e. opportunistic behaviour). To test whether there is evidence of this type of behavior in our setting, we compare vignette to self-reported thresholds in terms of the estimated country threshold dummies in the Ochopit specification. Our objective is to see whether any consistent pattern emerges when comparing the welfare state countries to other countries. From the results in Table 3, it can been seen for the health measure of self-reported work-disability that relative to Italy, individuals in the northern European countries of Germany, the Netherlands, France and Belgium tend to use lower thresholds when assessing their own disability status than when judging the disability status of the vignette person.[11] Whereas, in Greece and Spain, the vignette person's health is not rated significantly higher than the respondent's own health relative to Italy. This result holds for all cases for the two thresholds estimated in Table 2. Thus, it would seem that individuals in countries in which social expenditures are high relative to GDP tend to rate their own health below that of the vignette person. Although Sweden appears to be an exception to this rule, a possible explanation is the relative strictness with which vocational assessments are made in disability cases in Sweden compared to other

[11] Social expenditures constitute at least one fifth of GDP in these countries (OECD, 2003).

SHARE countries (Börsch-Supan, 2007). On the other hand, in Tables 5 and 6 where we consider the more narrowly-defined health measure of self-reported mobility, and where the sample is confined to the oldest old, a group which is presumably not motivated by strategic considerations when reporting health, we hardly see any differences between welfare state countries and other countries in the relative health ratings. While individuals in Sweden and the Netherlands claim significantly less mobility than the vignette person, in Germany, France, Spain, Greece and Belgium, the vignette person is not ranked higher than individuals rank themselves. Strategic behavior, therefore, seems in part to underlie the country differences in the thresholds when the health measure is work disability, and where the sample consists of the working-age elderly.

5.2 Is OF any better than RC?

Is the cure any better than the disease here? This question arises since the validity of our approach here over the original (King et al., 2004) vignette approach entirely hinges on the substitution of one identifying assumption (RC) with another identifying assumption (OF). We have shown that the likelihood increases when we impose OF instead of RC but no formal test is available as the models are non-nested. As such, even though we have good arguments for why OF is more reasonable than RC, this is not definitely conclusive. An interesting future research agenda could be to compare how much bias plausible deviations from OF and RC, respectively, generate.

[Figure 3 about here].

5.3 Policy Implications

So far, we have shown that the response consistency assumption can be violated in very general cases when comparing self-reported to objectively measured health. But what are the implied consequences of imposing an assumption of response consistency for policy purposes? We explore this question in Figure 3 by plotting simple cross-country correlations between the predicted health index of disability (1=no disability, 0=severe disability) from each of the models - Oprobit, Chopit and Ochopit - and health expenditures in percent of GDP available from World Bank Development Indicators (WBI), 2004. In the DIF uncorrected case, a one percentage point increase in health expenditures is associated with a decrease in the health index (proportion non-disabled) by about 0.15ths of a percentage point, but this is statistically insignificant. In the DIF corrected case, assuming response consistency, the relationship is still negative and statistically insignificant corresponding to a decrease in the proportion non-disabled by 0.82 pct. points. In the Ochopit model, relaxing response consistency produces a considerably stronger positive and statistically significant correlation and shows an increase in the proportion non-disabled by 2.4 pct. points for every one percent increase in health expenditures. Although the figures depicted here are mere correlations and should not be used to infer causal relationships, it is worth pointing out that the strength of the correlation between health expenditures and disability rates is considerably underestimated assuming response consistency and leads to the erroneous conclusion that increasing investments in health do not go hand in hand with better health. When this assumption is relaxed, it can be seen that significant improvements in health in terms of lowered work disability in the population are associated with rising health care expenditures.

6 Conclusion

In this paper, we have investigated the validity of anchoring vignettes, which have been used to correct for systematic differences in response scales across individuals when answering questions on a subjective scale. Following the approach suggested by van Soest et al. (2007), we seek to test the validity of anchoring vignettes assessing whether the key identifying assumption of response consistency is a tenable assumption or not. In order to do this, we use cross-country health data which include self-assessment of work disability and vignettes, but also an objective measure of health (measured hand-grip strength). Including this objective measure allows us to free up the assumption of response consistency.

We find that the model log likelihood improves considerably when we do not impose the response consistency assumption. Model comparisons using both the Akaike and the Bayesian information criterion support a specification not imposing response consistency and vignette corrected response scales. A robustness check using an alternative objective and self-reported measures of health, i.e. walking speed and mobility, confirms the main result. We also find that a rank order test rejects equality of country rankings between the Ochopit country ranking and the two other models' country rankings.

Our results indicate that the assumption of RC is not innocuous and has important implications for health policy. The results indicate that our extended model relaxing this assumption improves the fit and significantly changes the cross-country rankings of health vis-á-vis the standard Chopit model. We find this is in part due to strategic misreporting of work disability by the working-age elderly in welfare state nations. Disentangling whether deviations from RC cause greater bias than deviations from OF is a question we leave for future research.

References

[1] Angelini, V., Cavapozzi, D., Corazzini, L. and Paccagnella, O. (2008). Do Danes and Italians Rate Life Satisfaction in the Same Way? Using Vignettes to Correct for Individual-Specific Scale Biases. Manuscript. University of Padua.

[2] Bago d'Uva, T. , Doorslear, E.v., Lindeboom, M. and O'Donnell, O. (2008). Does Reporting Heterogeneity bias the Measurement of Health Disparities?. Health Economics, 17, 351-375.

[3] Börsch-Supan, A. (2007). Work Disability, Health and Incentive Effects. Sonder Forschungs Bereich 504, 07-23.

[4] Frederiksen, H., Gaist D. and Petersen, H. C. (2002). Hand grip strength: a phenotype suitable for identifying genetic variants affecting mid- and late-life physical functioning. Genetic Epdemiology 23, 110-122.

[5] Groot, W. (2000). Adaption and Scale of Reference Bias in Self-Assessments of Quality of Life. Journal of Health Economics 19, 403-420.

[6] Hopkins, D.J. and King, G. (2008). Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability. Manuscript, Harvard University.

[7] Javaras, K. N. and Ripley, B.D. (2007) An "unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response style. Journal of the American Statstical Association 102 (478), 454-463.

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

[9] Kerkhofs, M.K.M. and Lindeboom, M. (1995). Subjective health measurements and state dependent reporting errors. Health Economics 4, 221-235.

[10] King, G. A., Murray, C. J. L., Salomon, J.A. and Tandon, A. (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. American Political Science Review, 98 (1), 191-207.

[11] Kristensen, N. and Johansson, E. (2008). New Evidence on cross-country differences in job satisfaction using anchoring vignettes. Labour Economics, 15, 96-117.

[12] Juerges H. (2007). True Health vs Response Styles: Exploring cross-country differences in self-reported health. Health Economics 16 (2) 2007, 163-178.

[13] Murray, C.J.L., Ozaltin, E., Tandon, A. and Salomon, J. (2003). Empirical evaluation of the anchoring vignettes approach in health surveys. In Health Systems Performance Assessment: Debates, Methods and Empiricism, Murray CJL, Evans, DB (eds). World Health Organization, Geneva.

[14] Organisation for Economic Co-operation and Development. Health at a GlanceOECD Indicators 2003. OECD. Paris, France 2003b.

[15] Peracchi, F. and Rosetti, C. (2008). Gender and regional differences in self-rated health in Europe. Manuscript, Tor Vergata University.

[16] 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.

[17] Salomon, J. A., Tandon, A., and Murray, C.J. (2004). Comparability of Self Related Health: Cross Sectional Multi-Country Survey Using Anchoring Vignettes. British Medical Journal, 328, 258-264.

[18] van Doorslaer, E. and Gerdtham, U. G. (2003). Does Inequality in Self-Assessed Health Predict Inequality in Survival by Income? Evidence from Swedish Data. Social Science and Medicine, 57, 1621-1629.

[19] van Soest, A., Delaney, A., Harmon, C., Kapteyn, A. and Smith, J.P. (2007). Validating the use of vignettes for subjective threshold scales. Tilburg University, Discussion Paper 43.

World Development Indicators (WDI), World Bank.

Table 1: Descriptive statistics of covariates.

Variables	**Observations**	**Mean**	**Std. Dev.**	**Min**	**Max**
Personal characteristics					
Age	3918	64.684	9.929	50	99
Low education	3918	0.352	0.478	0	1
Average education	3918	0.171	0.377	0	1
Above average education	3918	0.303	0.460	0	1
High education	3918	0.173	0.378	0	1
Gender (1, female)	3918	0.548	0.498	0	1
Log of household income normalized by household size	3856	11.035	1.847	3.842	15.173
Living in a urban area	3918	0.307	0.461	0	1
Weight	3879	73.280	13.478	40	160
Height	3918	166.561	8.605	100.68	210
Employed	3918	0.279	0.448	0	1
Countries					
Germany	3918	0.256	0.436	0	1
Sweden	3918	0.021	0.144	0	1
the Netherlands	3918	0.045	0.207	0	1
Spain	3918	0.138	0.345	0	1
Italy	3918	0.189	0.392	0	1
France	3918	0.288	0.453	0	1
Greece	3918	0.036	0.187	0	1
Belgium	3918	0.027	0.161	0	1

Notes: Weighted results. Source: SHARE release 2.

Table 2: Work disability equation: Ordered Probit, Chopit and O-Chopit with OF.

Health domain: Work disability	**Ordered Probit**		**Chopit**		**O-Chopit (with OF)**			
	Coeff	*Std. Err.*	*Coeff*	*Std. Err.*	*Coeff*	*Std. Err.*		
Personal characteristics								
Age1 (50-54)	-0.299**	0.095	-0.384**	0.158	-0.945**	0.095		
Age2 (55-59)	-0.327**	0.083	-0.406**	0.144	-0.771**	0.092		
Age3 (60-64)	-0.407**	0.072	-0.434**	0.127	-0.564**	0.075		
Employed	-0.329**	0.078	-0.568**	0.135	-0.240**	0.077		
Weight	0.011**	0.002	0.019**	0.004	-0.010**	0.003		
Height	-0.005	0.005	-0.008	0.008	-0.035**	0.006		
Low education	0.461**	0.087	0.913**	0.155	0.259**	0.091		
Average education	0.305**	0.091	0.606**	0.165	0.216**	0.093		
Above average education	0.232**	0.080	0.377**	0.141	0.124	0.082		
Gender (1, female)	0.143**	0.070	0.181	0.118	1.703**	0.079		
Log of household income	-0.034**	0.015	-0.084**	0.024	-0.039**	0.017		
Living in a urban area	-0.111**	0.047	-0.148	0.102	0.045	0.063		
Country indicator (ref: Italy):								
Germany	0.337**	0.094	0.358**	0.167	-0.433**	0.102		
Sweden	0.057	0.104	-0.332	0.175	-0.055	0.105		
the Netherlands	-0.063	0.095	-0.206	0.175	-0.252**	0.101		
Spain	0.097	0.098	-0.477**	0.169	0.428**	0.106		
France	0.092	0.083	0.056	0.137	0.000	0.088		
Greece	-0.454**	0.097	-0.953**	0.165	0.193	0.101		
Belgium	0.266**	0.087	0.364**	0.157	-0.209**	0.095		
Thresholds								
Threshold 1			**Vignette=Self reported**		**Vignette**		**Selfreported**	
Age1 (50-54)			0.115	0.098	0.013	0.106	-0.448**	0.120
Age2 (55-59)			0.133	0.098	0.078	0.107	-0.323**	0.112
Age3 (60-64)			0.229**	0.108	0.192	0.114	-0.071	0.103
Employed			-0.085	0.092	-0.035	0.098	0.008	0.101
Weight			0.005	0.002	0.005	0.003	-0.020**	0.003
Height			-0.002	0.005	-0.003	0.006	-0.031**	0.007
Low education			0.309**	0.103	0.333**	0.110	-0.180	0.120
Average education			0.174	0.127	0.192	0.136	-0.104	0.132
Above average education			0.014	0.090	0.008	0.095	-0.117	0.111
Gender (1, female)			-0.035	0.089	-0.016	0.095	1.537**	0.102
Log of household income			-0.030	0.019	-0.025	0.020	-0.008	0.022
Living in a urban area			0.039	0.077	0.047	0.082	0.161	0.081
Germany			-0.111	0.130	-0.085	0.138	-0.775**	0.134
Sweden			-0.341**	0.113	-0.447**	0.126	0.014	0.140
the Netherlands			-0.190	0.138	-0.191	0.148	-0.248	0.137
Spain			-0.684**	0.111	-0.818**	0.133	0.416**	0.140
France			-0.009	0.103	-0.026	0.108	-0.006	0.113
Greece			-0.231**	0.109	-0.294**	0.116	0.752**	0.133
Belgium			-0.087	0.113	-0.081	0.120	-0.539**	0.124
Constant	-0.200	0.749	-2.128**	0.843	-2.042**	0.897	-0.388	0.810
Threshold 2								
Age1 (50-54)			0.035	0.065	0.052	0.065	-0.898**	0.127
Age2 (55-59)			0.037	0.058	0.049	0.058	-0.586**	0.111
Age3 (60-64)			0.112	0.057	0.119	0.057	-0.261**	0.104
Employed			-0.034	0.055	-0.040	0.055	0.194	0.110
Weight			0.000	0.002	0.000	0.002	-0.023**	0.003
Height			0.001	0.003	0.001	0.003	-0.032**	0.007
Low education			0.061	0.064	0.058	0.064	-0.248	0.126
Average education			0.080	0.065	0.074	0.064	-0.083	0.140
Above average education			0.030	0.055	0.031	0.055	-0.104	0.119
Gender (1, female)			-0.038	0.053	-0.039	0.053	1.578**	0.106
Log of household income			-0.033**	0.011	-0.034	0.011	0.002	0.022
Living in a urban area			-0.004	0.044	-0.005	0.044	0.151	0.083
Germany			-0.248**	0.073	-0.247**	0.072	-0.792**	0.141
Sweden			-0.526**	0.074	-0.520**	0.074	-0.272	0.147
the Netherlands			0.066	0.070	0.068	0.069	-0.045	0.142
Spain			-0.558**	0.073	-0.542**	0.073	0.216	0.141
France			-0.199**	0.063	-0.195**	0.063	-0.204	0.119
Greece			-0.361**	0.068	-0.348**	0.068	0.405**	0.138
Belgium			-0.014	0.065	-0.015	0.064	-0.413**	0.128
Constant	0.544	0.749	-0.819	0.530	-0.846	0.531	0.867	0.910
Theta values			Yes		Yes			
Thresholds objective								
Threshold 1					-7.375**	0.924		
Threshold 2					-5.707**	0.907		
Sigma self reported			1.523**	0.071	1.000	0.000		
Sigma vignette			1.000	0.000	1.000	0.000		
Sigma objective					1.000	0.000		
Rho self reported and objective					0.250**	0.034		
Log pseudo likelihood	-4401.167		-1.16E+08		-1.05E+08			

Notes: **: significant at two-sided 5-percent level. Weighted results. Source: SHARE release 2.

Table 3: Log likelihood values and tests of equality of thresholds coefficients.

Health domain	**Models**	**Log Pseudo Likelihood**	**AIC**	**BIC**
Work disability	Chopit (parameters)	-1.16E+08	232000110	232000409.2
	O-Chopit with RC (parameters)	-1.17E+08	233600198	233600409.2
	O-Chopit with OF (parameters)	-1.05E+08	209000198	209000409.2

Wald test	P-values	**One-sided test**	P-values
All coefficients		**All coefficients**	
Threshold1(vignette) = Threshold1(selfreported)	0.000	Threshold1(vignette) ≥ Threshold1(selfreported)	0.000
Threshold2(vignette) = Threshold2(selfreported)	0.000	Threshold2(vignette) ≥ Threshold2(selfreported)	0.000
Germany		**Germany**	
Threshold1(vignette) = Threshold1(selfreported)	0.000	Threshold1(vignette) ≥ Threshold1(selfreported)	0.000
Threshold2(vignette) = Threshold2(selfreported)	0.001	Threshold2(vignette) ≥ Threshold2(selfreported)	0.000
Sweden		**Sweden**	
Threshold1(vignette) = Threshold1(selfreported)	0.014	Threshold1(vignette) ≥ Threshold1(selfreported)	1.000
Threshold2(vignette) = Threshold2(selfreported)	0.129	Threshold2(vignette) ≥ Threshold2(selfreported)	1.000
the Netherlands		**the Netherlands**	
Threshold1(vignette) = Threshold1(selfreported)	0.774	Threshold1(vignette) ≥ Threshold1(selfreported)	0.000
Threshold2(vignette) = Threshold2(selfreported)	0.464	Threshold2(vignette) ≥ Threshold2(selfreported)	0.000
Spain		**Spain**	
Threshold1(vignette) = Threshold1(selfreported)	0.000	Threshold1(vignette) ≥ Threshold1(selfreported)	1.000
Threshold2(vignette) = Threshold2(selfreported)	0.000	Threshold2(vignette) ≥ Threshold2(selfreported)	1.000
France		**France**	
Threshold1(vignette) = Threshold1(selfreported)	0.895	Threshold1(vignette) ≥ Threshold1(selfreported)	0.000
Threshold2(vignette) = Threshold2(selfreported)	0.944	Threshold2(vignette) ≥ Threshold2(selfreported)	0.000
Greece		**Greece**	
Threshold1(vignette) = Threshold1(selfreported)	0.000	Threshold1(vignette) ≥ Threshold1(selfreported)	1.000
Threshold2(vignette) = Threshold2(selfreported)	0.000	Threshold2(vignette) ≥ Threshold2(selfreported)	1.000
Belgium		**Belgium**	
Threshold1(vignette) = Threshold1(selfreported)	0.008	Threshold1(vignette) ≥ Threshold1(selfreported)	0.000
Threshold2(vignette) = Threshold2(selfreported)	0.000	Threshold2(vignette) ≥ Threshold2(selfreported)	0.000

Notes: Source: SHARE release 2.

Table 4: Country rankings.

	Work disability		
Rank	**Ordered Probit**	**Chopit**	**O-Chopit**
1	Greece	Greece	Germany
2	the Netherlands	Spain	the Netherlands
3	Italy	Sweden	Belgium
4	Sweden	the Netherlands	Sweden
5	France	Italy	France
6	Spain	France	Italy
7	Belgium	Germany	Greece
8	Germany	Belgium	Spain

Notes: Source: SHARE release 2. A low rank (e.g. rank=1) indicates *less* work disability.

Table 5: Mobility Equation: Ordered Probit, Chopit and O-Chopit relaxing the response consistency assumption.

Health domain: Mobility	**Ordered Probit**		**Chopit**		**Ochopit (with OF)**			
	Coeff	*Std. Err.*	*Coeff*	*Std. Err.*	*Coeff*		*Std. Err.*	
Personal characteristics								
Age	0.002	0.010	0.005	0.014	0.028**		0.010	
Low education	0.392	0.336	0.625	0.416	1.598**		0.332	
Average education	0.787	0.381	1.068**	0.481	1.320**		0.366	
Above average education	0.844	0.346	1.036	0.513	0.920		0.349	
Gender (1, female)	0.199	0.203	0.117	0.277	0.047		0.201	
Weight	0.040	0.008	0.045**	0.010	0.026**		0.006	
Height	-0.024	0.013	-0.026	0.015	-0.020		0.010	
Living in a urban area	0.075	0.172	0.100	0.192	0.122		0.138	
Country indicator (ref: Italy):								
Germany	-0.099	0.300	0.044	0.475	0.057		0.340	
Sweden	0.374	0.312	0.330	0.427	-1.047**		0.350	
the Netherlands	-0.227	0.309	-0.003	0.401	-1.041**		0.308	
Spain	-0.093	0.299	-0.083	0.404	-0.371		0.317	
France	0.202	0.246	0.204	0.371	-0.112		0.264	
Greece	-0.133	0.302	0.028	0.410	-0.340		0.308	
Belgium	-0.008	0.280	0.512	0.384	-0.337		0.319	
Thresholds								
Threshold 1			**Vignette=Self reported**		**Vignette**		**Selfreported**	
Age			-0.009	0.013	-0.009	0.015	0.022	0.014
Low education			0.212	0.344	0.418	0.489	1.102**	0.379
Average education			0.106	0.378	0.392	0.494	0.318	0.464
Above average education			-0.208	0.353	-0.050	0.482	-0.255	0.426
Gender (1, female)			-0.170	0.192	-0.199	0.222	-0.094	0.192
Germany			0.270	0.392	0.284	0.451	0.189	0.453
Sweden			-0.042	0.434	-0.273	0.512	-1.280**	0.477
the Netherlands			0.273	0.345	0.046	0.416	-0.744**	0.383
Spain			0.059	0.337	-0.012	0.374	-0.250	0.375
France			-0.025	0.310	-0.312	0.382	-0.191	0.319
Greece			0.002	0.299	-0.341	0.373	-0.248	0.386
Belgium			0.588	0.339	0.426	0.397	-0.183	0.390
Constant	-0.685	2.264	-0.798	1.145	-1.463	1.138	-0.617	1.957
Threshold 2								
Age			0.011	0.009	0.010	0.009	0.038	0.015
Low education			0.177	0.253	0.147	0.244	1.151**	0.351
Average education			0.349	0.332	0.293	0.325	0.672	0.483
Above average education			0.354	0.290	0.315	0.291	0.312	0.444
Gender (1, female)			-0.011	0.162	-0.009	0.162	0.030	0.186
Germany			-0.026	0.317	-0.023	0.311	-0.141	0.542
Sweden			-0.115	0.323	-0.069	0.319	-1.654**	0.489
the Netherlands			0.285	0.258	0.336	0.257	-1.047**	0.428
Spain			-0.075	0.254	-0.072	0.249	-0.445	0.406
France			-0.049	0.234	0.000	0.231	-0.535	0.359
Greece			0.331	0.227	0.396	0.223	-0.425	0.397
Belgium			0.479	0.250	0.520**	0.242	-0.541	0.443
Constant	0.110	2.273	-1.711**	0.826	-2.128**	0.739	-1.070	2.022
Theta values			Yes		Yes			
Thresholds objective								
Threshold 1					1.720	1.901		
Threshold 2					2.775	1.904		
Sigma self reported			1.117**	0.166	1.000	0.000		
Sigma vignette			1.000	0.000	1.000	0.000		
Sigma objective					1.000	0.000		
Rho self reported and objective					0.397**	0.089		
Log pseudo likelihood	-3469.424		-7.537E+06		-5475061.8			

Notes: **: significant at two-sided 5-percent level. Weighted results. Source: SHARE release 2.

Table 6: Log likelihood values and tests of equality of thresholds coefficients.

Health domain	**Models**	**Log Pseudo Likelihood**	**AIC**	**BIC**
Mobility	Chopit (parameters)	-7.537E+06	15074918.8	15075218.03
	Ochopit with RC (parameters)	-7.538E+06	15076131.8	15076343.03
	Ochopit with OF (parameters)	-5475061.8	10950321.6	10950532.83

Wald test	P-values	**One-sided test**	P-values
All coefficients		**All coefficients**	
Threshold1(vignette) = Threshold1(selfreported)	0.000	Threshold1(vignette) ≥ Threshold1(selfreported)	0.000
Threshold2(vignette) = Threshold2(selfreported)	0.000	Threshold2(vignette) ≥ Threshold2(selfreported)	0.000
Germany		**Germany**	
Threshold1(vignette) = Threshold1(selfreported)	0.875	Threshold1(vignette) ≥ Threshold1(selfreported)	1.000
Threshold2(vignette) = Threshold2(selfreported)	0.837	Threshold2(vignette) ≥ Threshold2(selfreported)	1.000
Sweden		**Sweden**	
Threshold1(vignette) = Threshold1(selfreported)	0.115	Threshold1(vignette) ≥ Threshold1(selfreported)	0.000
Threshold2(vignette) = Threshold2(selfreported)	0.000	Threshold2(vignette) ≥ Threshold2(selfreported)	0.000
the Netherlands		**the Netherlands**	
Threshold1(vignette) = Threshold1(selfreported)	0.171	Threshold1(vignette) ≥ Threshold1(selfreported)	0.000
Threshold2(vignette) = Threshold2(selfreported)	0.005	Threshold2(vignette) ≥ Threshold2(selfreported)	0.000
Spain		**Spain**	
Threshold1(vignette) = Threshold1(selfreported)	0.663	Threshold1(vignette) ≥ Threshold1(selfreported)	1.000
Threshold2(vignette) = Threshold2(selfreported)	0.414	Threshold2(vignette) ≥ Threshold2(selfreported)	1.000
France		**France**	
Threshold1(vignette) = Threshold1(selfreported)	0.816	Threshold1(vignette) ≥ Threshold1(selfreported)	1.000
Threshold2(vignette) = Threshold2(selfreported)	0.217	Threshold2(vignette) ≥ Threshold2(selfreported)	1.000
Greece		**Greece**	
Threshold1(vignette) = Threshold1(selfreported)	0.867	Threshold1(vignette) ≥ Threshold1(selfreported)	1.000
Threshold2(vignette) = Threshold2(selfreported)	0.066	Threshold2(vignette) ≥ Threshold2(selfreported)	1.000
Belgium		**Belgium**	
Threshold1(vignette) = Threshold1(selfreported)	0.277	Threshold1(vignette) ≥ Threshold1(selfreported)	1.000
Threshold2(vignette) = Threshold2(selfreported)	0.029	Threshold2(vignette) ≥ Threshold2(selfreported)	0.000

Notes: Source: SHARE release 2.

Table 7: Country rankings II.

	Mobility		
Rank	**Ordered Probit**	**Chopit**	**Ochopit**
1	the Netherlands	Spain	Sweden
2	Greece	the Netherlands	the Netherlands
3	Germany	Italy	Spain
4	Spain	Greece	Greece
5	Belgium	Germany	Belgium
6	Italy	France	France
7	France	Sweden	Italy
8	Sweden	Belgium	Germany

Notes: Source: SHARE release 2.

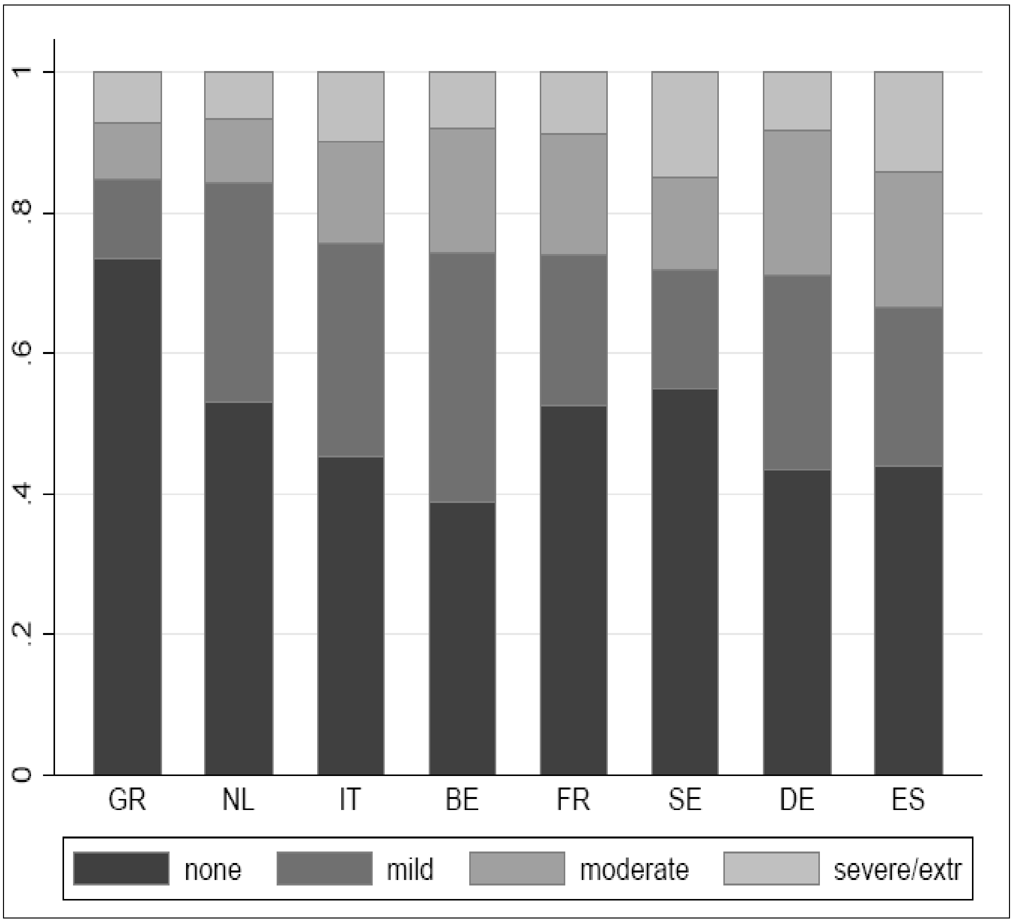

Figure 1: Self-reported work disability, by country.

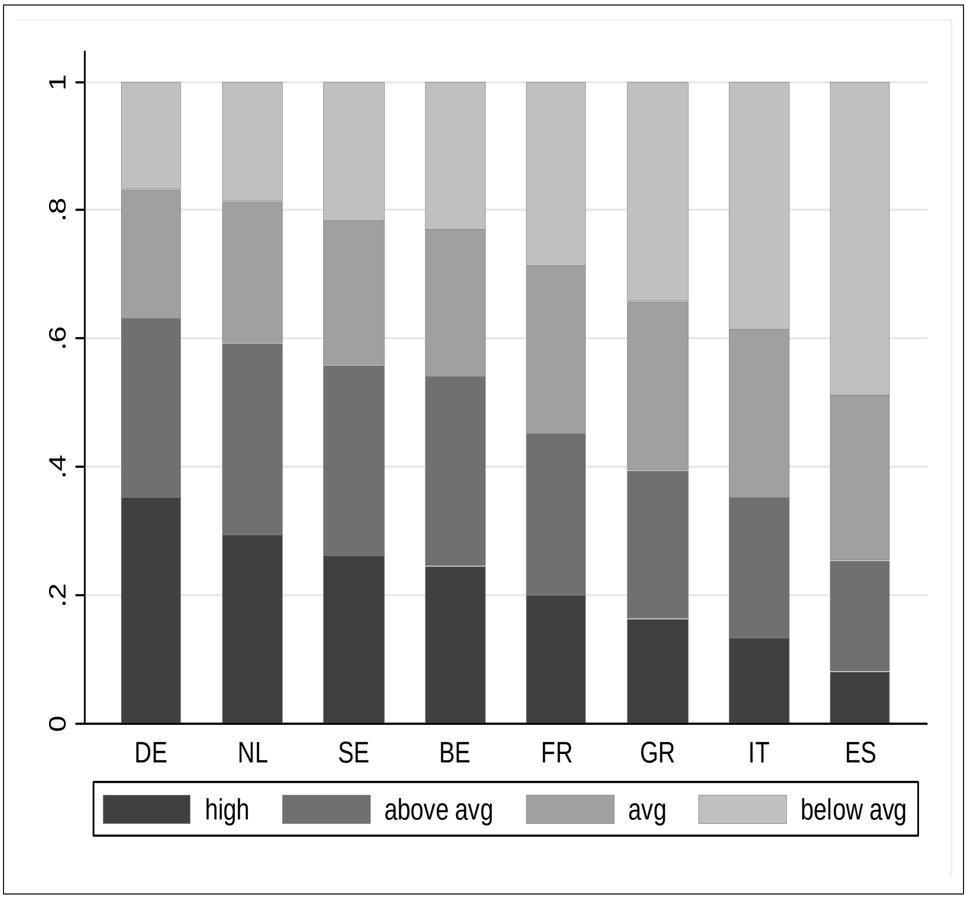

Figure 2: Categorised grip strength, by country.

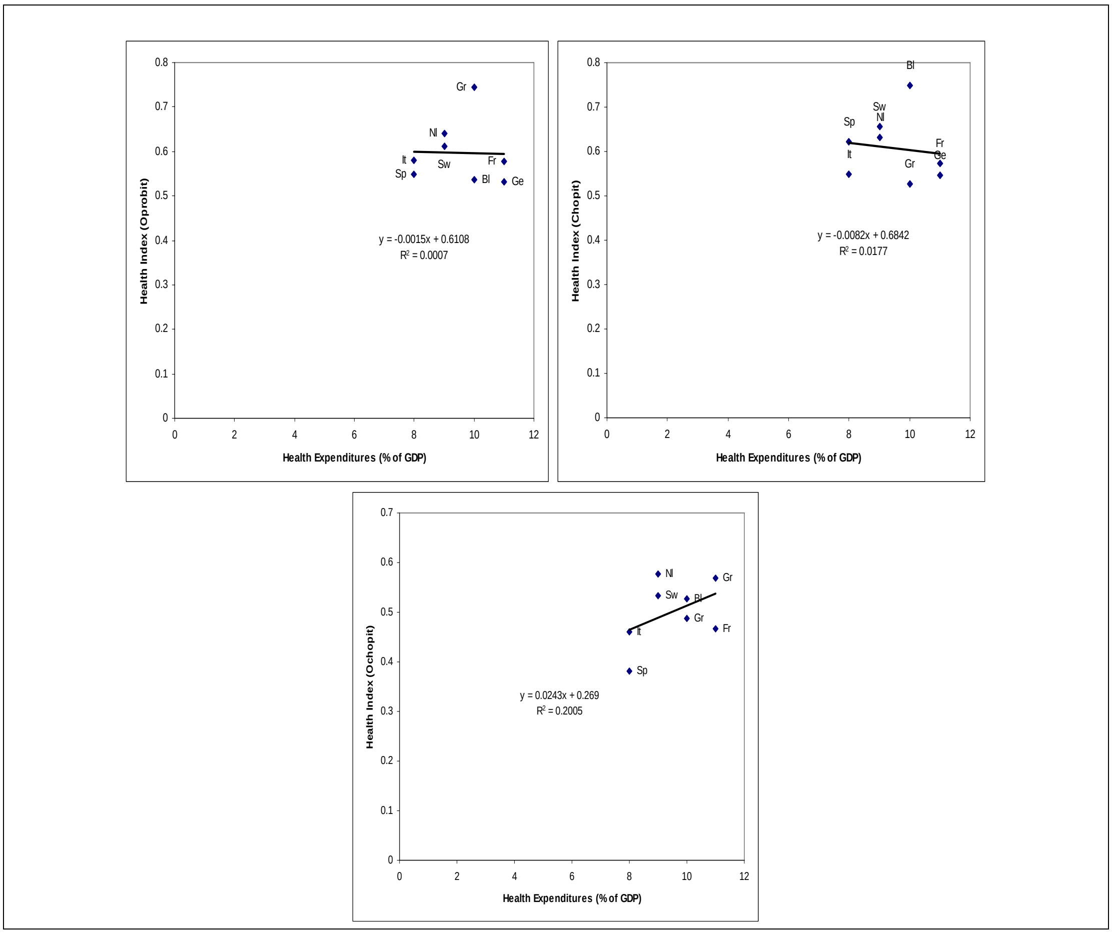

Figure 3: Health expenditures and population work disability levels.

Appendix A: The Work Disability and Mobility Vignettes

In the main analysis, we use the following set of vignettes relating to work disability:

1. ”Alice has almost constant pain in her back and this sometimes prevents her from doing her work. How much is Alice limited in the kind or amount of work he could do?” None, Mild, Moderate, Severe and Extreme.

2. ”Kevin suffers from back pain that causes stiffness in his back especially at work but is relieved with low doses of medication. He does not have any pains other than this generalized discomfort. How much is Kevin limited in the kind or amount of work he could do?” None, Mild, Moderate, Severe and Extreme.

3. ”Lisa has pain in her back and legs and the pain is present almost all the time. It gets worse while she is working. Although medication helps, she feels uncomfortable when moving around, holding and lifting things at work. How much is Lisa limited in the kind or amount of work she could do?.” None, Mild, Moderate, Severe and Extreme.

4. ”Tom feels worried all the time. He gets depressed once a week at work for a couple of days in a row, thinking about what could go wrong and that his boss will disapprove of his condition. But he is able to come out of this mood if he concentrates on something else. How much is Tom limited in the kind or amount of work he could do?” None, Mild, Moderate, Severe and Extreme.

5. ”Tamara has mood swings on the job. When she gets depressed, everything she does at work is an effort for her and she no longer enjoys her usual activities at work.

These mood swings are not predictable and occur two or three times during a month. How much is Tamara limited in the kind or amount of work she could do?" None, Mild, Moderate, Severe and Extreme.

6. "Anthony generally enjoys his work. He gets depressed every 3 weeks for a day or two and loses interest in what he usually enjoys but is able to carry on with his day-to-day activities on the job. How much is Anthony limited in the kind or amount of work he could do?" None, Mild, Moderate, Severe and Extreme.

7. "Eve has had heart problems in the past and she has been told to watch her cholesterol level. Sometimes if she feels stressed at work she feels pain in her chest and occasionally in her arms. How much is Eve limited in the kind or amount of work she could do?" None, Mild, Moderate, Severe and Extreme.

8. "Mark has been diagnosed with high blood pressure. His blood pressure goes up quickly if he feels under stress. Mark does not exercise much and is overweight. How much is Mark limited in the kind or amount of work he could do?" None, Mild, Moderate, Severe and Extreme.

9. "Anna has undergone triple bypass heart surgery. She is a heavy smoker and still experiences severe chest pain sometimes. How much is Anna limited in the kind or amount of work she could do?" None, Mild, Moderate, Severe and Extreme.

As far as the sensitivity analysis is concerned, we use the following set of vignettes related to mobility :

1."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?" None, Mild, Moderate, Severe and Extreme.

2."Kevin does not exercise. He cannot climb stairs or do other physical activities because he is obese. He is able to carry the groceries and do some light household work.

Overall in the last 30 days, how much of a problem did Kevin have with moving around?" None, Mild, Moderate, Severe and Extreme.

3. "Rob is able to walk distances of up to 200 meters without any problems but feels tired after walking one kilometer or climbing more than one flight of stairs. He has no problems with day-to-day activities, such as carrying food from the market.

Overall in the last 30 days, how much of a problem did Rob have with moving around?" None, Mild, Moderate, Severe and Extreme.

Appendix B: Grip Strength and Walking Speed

The objective measure we use in the main analysis is hand grip strength. It is measured using a hand-held dynamometer, where respondents are asked to press a lever as hard as they can. The dynamometer shows grip strength in kilograms. We take the maximum of up to four measurements: two on the left hand and and two on the right hand. This variable is missing if the respondent does not have two measurements on at least one hand or if these differ by more than 20 kg or had implausible values.

In the analyses, we drop missing values and we categorize grip strength taking the age and gender specific quartiles across its empirical distribution.

For the robustness check, we use walking speed. This is a measure of mobility and functioning of the lower limbs that strongly declines with age (available only for those 75 and over or respondents with self-reported mobility limitations). It is measured by a timed walk over a short distance (2.5m). Two measurements were made, of which we take the fastest.

Appendix C: Ochopit – imposing RC and relaxing OF

Table 1C: Work disability equation: Ochopit assuming the response consistency assumption and relaxing the one factor assumption.

Health domain: Work disability	**Ochopit (with RC)**			
	Self-reported		**Objective**	
	Coeff	*Std. Err.*	*Coeff*	*Std. Err.*
Personal characteristics				
Age1 (50-54)	-0.241	0.122	-0.948**	0.095
Age2 (55-59)	-0.259**	0.111	-0.773**	0.092
Age3 (60-64)	-0.263**	0.100	-0.565**	0.076
Employed	-0.428**	0.105	-0.241**	0.077
Weight	0.015**	0.003	-0.010**	0.003
Height	-0.015**	0.002	-0.037**	0.006
Low education	0.650**	0.114	0.253**	0.091
Average education	0.424**	0.127	0.212**	0.093
Above average education	0.246**	0.108	0.120	0.082
Gender (1, female)	0.028	0.076	1.697**	0.079
Log of household income	-0.074	0.018	-0.040	0.017
Living in a urban area	-0.107	0.080	0.044	0.063
Country indicator (ref: Italy):				
Germany	0.219	0.132	-0.434**	0.103
Sweden	-0.310**	0.134	-0.053	0.106
the Netherlands	-0.152	0.140	-0.250**	0.102
Spain	-0.492**	0.129	0.428**	0.107
France	0.004	0.108	-0.002	0.089
Greece	-0.747**	0.127	0.193	0.101
Belgium	0.249	0.124	-0.210**	0.096
Thresholds				
Threshold 1	**Vignette=Self reported**			
Age1 (50-54)	0.114	0.087		
Age2 (55-59)	0.122	0.088		
Age3 (60-64)	0.203**	0.097		
Employed	-0.093	0.083		
Weight	0.004**	0.002		
Height	-0.004	0.004		
Low education	0.276**	0.091		
Average education	0.149	0.111		
Above average education	0.012	0.079		
Gender (1, female)	-0.053	0.076		
Log of household income	-0.029	0.017		
Living in a urban area	0.025	0.068		
Germany	-0.100	0.115		
Sweden	-0.307**	0.099		
the Netherlands	-0.167	0.122		
Spain	-0.593**	0.095		
France	-0.015	0.093		
Greece	-0.223**	0.098		
Belgium	-0.076	0.101		
Constant	-1.665**	0.650		
Threshold 2				
Age1 (50-54)	0.034	0.065		
Age2 (55-59)	0.040	0.058		
Age3 (60-64)	0.116	0.057		
Employed	-0.027	0.056		
Weight	0.000	0.002		
Height	0.000	0.003		
Low education	0.057	0.064		
Average education	0.073	0.065		
Above average education	0.026	0.055		
Gender (1, female)	-0.047	0.052		
Log of household income	-0.033	0.011		
Living in a urban area	-0.002	0.044		
Germany	-0.243**	0.073		
Sweden	-0.528**	0.075		
the Netherlands	0.071	0.071		
Spain	-0.562**	0.072		
France	-0.198**	0.064		
Greece	-0.357**	0.069		
Belgium	-0.015	0.065		
Constant	-0.651	0.485		
Theta values	Yes			
Thresholds objective				
Threshold 1	-7.622**	0.909		
Threshold 2	-5.948**	0.893		
Sigma self reported	1.000	0.000		
Sigma vignette	1.000	0.000		
Sigma objective	1.000	0.000		
Rho self reported and objective	0.230**	0.031		
Log pseudo likelihood	-1.17E+08			

Notes: **: significant at two-sided 5-percent level. Weighted results. Source: SHARE release 2.

Table 2C: Mobility equation: Ochopit assuming the response consistency assumption and relaxing the one factor assumption.

Health domain: Mobility	**O-Chopit (with RC)**			
	Self-reported		**Objective**	
	Coeff	*Std. Err.*	*Coeff*	*Std. Err.*
Personal characteristics				
Age	0.003	0.012	0.022**	0.010
Low education	0.555	0.375	1.639**	0.326
Average education	0.947**	0.439	1.353**	0.364
Above average education	0.915**	0.439	0.926**	0.344
Gender (1, female)	0.038	0.217	-0.073	0.212
Weight	0.041**	0.008	0.014**	0.007
Height	-0.029**	0.007	-0.020	0.012
Living in a urban area	0.083	0.174	0.162	0.184
Country indicator (ref: Italy):				
Germany	-0.538	0.384	0.411	0.294
Sweden	-0.281	0.347	-0.692	0.339
the Netherlands	-0.539	0.343	-0.694	0.293
Spain	-0.708**	0.332	-0.089	0.322
France	-0.605	0.358	0.286	0.317
Greece	-0.398	0.277	0.177	0.253
Belgium	-0.517	0.355	-0.007	0.320
Thresholds				
Threshold 1	**Vignette=Self reported**			
Age	-0.010	0.012		
Low education	0.195	0.327		
Average education	0.112	0.363		
Above average education	-0.189	0.336		
Gender (1, female)	-0.157	0.189		
Germany	-0.347	0.298		
Sweden	-0.602	0.367		
the Netherlands	-0.340	0.271		
Spain	-0.554	0.310		
France	-0.601	0.318		
Greece	-0.627**	0.227		
Belgium	-0.583	0.259		
Constant	-0.687	1.069		
Threshold 2				
Age	0.010	0.009		
Low education	0.168	0.252		
Average education	0.324	0.332		
Above average education	0.334	0.288		
Gender (1, female)	-0.014	0.161		
Germany	-0.520**	0.264		
Sweden	-0.620**	0.290		
the Netherlands	-0.211	0.204		
Spain	-0.576**	0.240		
France	-0.506**	0.247		
Greece	-0.546**	0.198		
Belgium	-0.174	0.204		
Constant	-1.588**	0.786		
Theta values	Yes			
Thresholds objective				
Threshold 1	0.407	2.033		
Threshold 2	1.454	2.037		
Sigma self reported	1.000	0.000		
Sigma vignette	1.000	0.000		
Sigma objective	1.000	0.000		
Rho self reported and objective	0.427**	0.085		
Log pseudo likelihood	-7.538E+06			

Notes: **: significant at two-sided 5-percent level. Weighted results. Source: SHARE release 2.