Netspar

Network for Studies on Pensions, Aging and Retirement

Netspar DISCUSSION PAPERS

Arthur van Soest and Hana Voňková

Testing Parametric Models Using Anchoring Vignettes against Nonparametric Alternatives

DP 02/2012-012

Testing Parametric Models Using Anchoring Vignettes against Nonparametric Alternatives

Arthur van Soest, Netspar, Tilburg University
Hana Voňková, Charles University in Prague

February, 2012

Summary

Comparing assessments of health, job satisfaction, etc. on a subjective scale across countries or socio-economic groups is often hampered by differences in response scales across groups. Anchoring vignettes help to correct for such differences, either in parametric models (CHOPIT and extensions) or non-parametrically, comparing rankings of vignette ratings and self-assessments across groups. We construct specification tests of parametric models, comparing non-parametric rankings with rankings using the parametric estimates. Applied to six domains of health, the test always rejects standard CHOPIT, but an extended CHOPIT performs better. This implies a need for more flexible (parametric or semi-parametric) models than standard CHOPIT.

Keywords: self-assessed health, reporting bias, CHOPIT model, specification testing
JEL-codes: C81, I10, C35

1 Introduction

In many studies in the social sciences using survey data, respondents' behaviours, attitudes, and well-being are measured in a qualitative and subjective manner. Respondents are typically asked to provide ratings on some subjective ordinal scale. An example that is commonly used in general socioeconomic surveys is a self-assessed health question on a five-point scale (from excellent to poor, for example). The use of subjective scales is widespread and not limited to health or well-being; other examples are evaluations of responsiveness of the health care system (Rice *et al.* (2010)), job satisfaction (Kristensen and Johansson (2008)), or political efficacy (King *et al.* (2004)).

Answers to questions with a subjective scale may depend on both the objective reality and the way in which respondents interpret the subjective answers, that is, the respondents' reporting behaviour. The latter is often referred to as differential item functioning (DIF; see Holland and Wainer (1993)). Usually, the aim is comparing the objective reality across socioeconomic groups or countries, and differences in reporting behaviour should be corrected for. To identify the differences in reporting behavior, King *et al.* (2004) have proposed to use anchoring vignettes. These are short descriptions of hypothetical persons or situations. Respondents are asked to evaluate one or more vignettes on the same subjective scale used to evaluate their own situation. Because the objective situation of the person described in the vignette(s) is the same for all respondents, systematic differences in vignette evaluations across respondents identify differences in reporting behaviour.

King *et al.* (2004) propose a parametric model as well as a nonparametric method for the use of anchoring vignettes to compare the distributions of the underlying objective reality of the phenomenon of interest (the survey reports corrected for DIF) in two or more countries or socio-economic groups. The parametric model is referred to as compound hierarchical ordered probit model (CHOPIT). Research using anchoring vignettes has grown rapidly in recent years, and virtually all applications use the CHOPIT model or parametric extensions of this model. This includes studies on several aspects of health (Bago d'Uva *et al.* (2008a), Bago d'Uva *et al.* (2008b), Vonkova

and Hullegie (2011)), health care responsiveness (Rice *et al.* (2010)), work disability (Kapteyn *et al.* (2007)), job satisfaction (Kristensen and Johansson (2008)), satisfaction with social contacts (Bonsang and Van Soest (2012)) and life satisfaction (Angelini *et al.* (2012)). The CHOPIT model consists of ordered probit equations (for vignette evaluations and the assessment of the own situation) with thresholds that are common to all equations and that can depend on the respondent's socio-economic characteristics to account for differences in reporting behaviour.

The nonparametric approach has been used much less often; exceptions are King *et al.* (2004) and King and Wand (2007). This method essentially compares the distributions in different socioeconomic groups of the rank of the respondent's self-evaluation amongst the same respondent's vignette evaluations. For example, suppose that only one vignette is evaluated by all respondents in two groups A and B (the same vignette in both groups); suppose almost everyone in group A evaluates their own health as better than that of the hypothetical vignette person (the benchmark), whereas in group B the majority evaluates the vignette person's health as better than their own. Then the non-parametric method immediately leads to the conclusion that group A is healthier than group B. This conclusion is still valid if the two groups use very different scales, since it is based upon the comparison with the vignette evaluation that (by assumption) uses the same scale as the self-assessment. This method does not require any model or covariates (other than those used to distinguish the two groups).

The non-parametric method relies on two assumptions: reporting behaviour of the respondents is the same in the self-assessment questions and the vignettes ("response consistency") and the objective reality represented in a vignette is perceived in the same way by all respondents ("vignette equivalence"). These can be called "identifying assumptions" in the sense that the interpretation of the non-parametric ranking comparison relies on them. These assumptions have been tested in recent studies, with mixed results (see Section 4 for some references). In this paper they are maintained (identifying) assumptions. The parametric model requires much more, in addition to these two assumptions. For example, it assumes that the objec-

tive reality can be modelled as a linear function of observed characteristics and an unobserved component; it assumes a specific functional form of the thresholds; and it assumes joint normality of the error terms.

In this paper, we compare the ranking distributions implied by the parametric model with the non-parametric rankings that come directly from the raw data, using the chi-squared diagnostic tests introduced in Andrews (1988). These can be seen as tests of the specification of the parametric model against non-parametric alternatives that lead to different rankings of the self-reports and vignette evaluations. While many alternative specification tests for the parametric model can be considered, our tests are motivated by the fact that they have power in a direction that matters: they reject the parametric model if misspecification is such that data generated by the parametric model would lead to biased conclusions concerning ranking comparisons across socioeconomic groups.

We run the tests for six health domains (breathing, cognition, depression, mobility, sleeping and bodily pains) on data on the population of ages 50 and older in eight European countries, from the 2004 wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). For each of the six domains and each respondent, we have self-assessments and evaluations of three vignettes describing different health levels of hypothetical persons.

We find that the standard CHOPIT model is always rejected, but a simple one parameter extension that allows for unobserved heterogeneity (used by Kapteyn *et al.* (2007), for example) is rejected for some health domains but not for others. This suggests that the standard CHOPIT model is misspecified and that conclusions about comparisons across countries or socioeconomic groups based upon this model may be biased. It also implies that the existing tests for vignette equivalence or response consistency that rely on the CHOPIT model may not be valid. On the other hand, the non-parametric method is not a viable alternative since it has limited applicability. It cannot be used with many covariates, or to produce counterfactual distributions of self-reported health with benchmark reporting scales. We therefore conclude that there is a need for future work on more flexible parametric or semi-parametric models that generalize the CHOPIT model.

The remainder of this paper is organized as follows. Section 2 explains the parametric CHOPIT model. In Section 3, we introduce our diagnostic tests. Section 4 relates the tests to the non-parametric approach for using anchoring vignettes. Section 5 presents the data for our application. Section 6 describes the results of some Monte Carlo simulations guiding how to implement the tests given the size and nature of our data. Our main results are discussed in Section 7. Section 8 concludes.

2 Parametric model

We first describe the CHOPIT model (King *et al.* (2004)) – the parametric model commonly used in studies with anchoring vignettes – and a one parameter extension of this model. For the sake of our exposition we assume that self-assessments and vignettes concern one given health domain (different domains are modelled completely separately), in line with our empirical application. The response scale is a five point scale going from no problem ('none') in the given health domain to an extreme problem ('extreme'; see Appendix B). The CHOPIT model consists of a self-assessment equation explaining the respondents' evaluation of their own health (in the given domain) and a vignette equation explaining the evaluation(s) of the health of (one or more) hypothetical vignette persons. The self-assessment of respondent i is modelled as follows:

$$
\begin{aligned}
Y_{si}^{\star} &= X_i^{'}\beta_s + \epsilon_{si} && (1)\\
Y_{si} &= j \Leftrightarrow \tau_i^{j-1} < Y_{si}^{\star} \leq \tau_i^{j} \quad j = 1, \ldots, J && (2)\\
\tau_i^{1} &= X_i^{'}\gamma^{1} + u_i && (3)\\
\tau_i^{j} &= \tau_i^{j-1} + \exp\left(X_i^{'}\gamma^{j}\right) \quad j = 2, \ldots, J-1 && (4)\\
\tau_i^{0} &= -\infty; \tau_i^{J} = \infty, && (5)
\end{aligned}
$$

Here $Y_{si}^{\star}$ is the latent health of respondent i in the given domain, modelled as the sum of a linear combination of explanatory variables X_i and an unobserved component ϵ_{si}, which may reflect unobserved heterogeneity, re-

porting error, or both. The observed value of self-assessed health Y_{si} is equal to $j(\in \{1,\ldots,J\})$; in our application, $J = 5$) if latent health is between thresholds τ_i^{j-1} and τ_i^j. The thresholds can vary with respondent characteristics X_i. Moreover, thresholds can vary with unobserved characteristics u_i. A large value of u_i means that all thresholds of respondent i are relatively high, implying that the respondent does not easily evaluate health problems in the given domain as severe or extreme. This unobserved heterogeneity term was not included in the standard CHOPIT model of King *et al.* (2004) but was introduced as an extension by Kapteyn *et al.* (2007).

The evaluations of K vignettes $v = 1,\ldots,K$ by respondent i are modelled as follows:

$$Y_{vi}^{\star} = \theta_v + \epsilon_{vi} \quad (v = 1,\ldots,K) \tag{6}$$

$$Y_{vi} = j \Leftrightarrow \tau_i^{j-1} \leq Y_{vi}^{\star} < \tau_i^j \quad (j = 1,\ldots,J) \tag{7}$$

where $Y_{vi}^{\star}$ is the latent health of the hypothetical person described in vignette v in the given domain, modelled as a sum of a vignette specific constant θ_v and an unobserved component ϵ_{vi}. θ_v does not vary across respondents since it is assumed that each vignette is interpreted in the same way by all respondents ("vignette equivalence"). In our application we use three vignettes ($K = 3$). Y_{vi} is the reported evaluation of vignette v by respondent i on the same J-point scale that is used for the self-assessments (in our application, $J = 5$). Y_{vi} is equal to $j = 1,\ldots,J$ if the latent health $Y_{vi}^{\star}$ is between thresholds τ_i^{j-1} and τ_i^j. The assumption of response consistency implies that the thresholds are the same as for the self-assessments.

The error terms $\epsilon_{si}, \epsilon_{vi}, v = 1,\ldots,K$, and the unobserved heterogeneity term u_i are assumed to be independent of each other and of the covariates X_i, with normal distributions that have mean zero and variances σ_s^2, σ_v^2 and σ_u^2, respectively. By means of normalization, we impose $\beta_{s,1} = 0$ and $\sigma_s = 1$. The standard way to estimate the (standard or extended) CHOPIT model is by maximum likelihood. Note that the extended CHOPIT model is a one parameter extension of standard CHOPIT, which sets σ_u^2 equal to zero.

3 Misspecification Tests

There are many ways to test the specification of a fully parametric model in general and of the (standard or extended) CHOPIT model in particular. For example, Lagrange multiplier tests can be performed against specific parametric extensions, such as models with heteroskedastic errors or errors with a non-normal distribution, in the spirit of, for example, Chesher and Irish (1987). Such tests will be powerful in the directions of the specific alternatives they are designed for but may be less powerful in other directions. Since one of the main goals of the parametric model is to compare health or well-being across countries or socio-economic groups after purging self-assessments for response scale differences, it seems more natural to look for tests with power in the directions of misspecification that lead to different conclusions concerning such comparisons.

A general category of misspecification tests are the goodness of fit tests of Andrews (1988). These tests first partition the product space of outcomes and regressors $Y \times X$ into cells; usually this is done by partitioning Y and X separately into M_Y and M_X cells respectively, and taking all $M_Y M_X$ products of the cells of the partitions of Y and X. Then the sample distribution over these cells is compared with the distribution generated by the estimated parametric model: For the given parameter estimates and for the given regressor values of each observation, the probability distribution of the dependent variable(s) is fully determined (by the distribution of error terms and unobserved heterogeneity) and the probabilities of each cell in the partition of Y can be computed. Averaging over all observations with regressor values in each given cell of the partition of X then gives the cell probabilities generated by the parametric model.

Under the null hypothesis that the parametric model is correctly specified, the sample distribution and the distribution generated by the model should be similar. If the parameters of the parametric model would all be known, this could be formalized with a Pearson chi-squared test. Andrews (1988) shows that the test statistic can be adjusted to correct for the fact that parameters are estimated using the same data. The appropriate test

statistic is a quadratic form which asymptotically has a chi-squared distribution under the null that the parametric model is correctly specified. If the parametric model is estimated by maximum likelihood (as in our case), Andrews shows that the test statistic can easily be computed by performing an auxiliary OLS regression. The left hand side variable in this regression is an n-dimensional vector $(1, 1, \ldots, 1)'$, where n is the number of observations. There are two groups of right hand side variables. First, for each cell, an $n-$dimensional vector with, for each of the n observations, the deviation between realizations (1 if the observation is in the given cell, 0 otherwise) and the cell probability according to the model (given the values of the regressors for that observation). Second, for each parameter, the vector of partial derivatives with respect to that parameter of the log likelihood contributions for all n observations (the "scores"); see Andrews (1988, p. 154) for details. The scores are added to correct for the fact that the parameters are estimated using the same sample. Under the null of no misspecification of the parametric model, n times the R^2 of this regression is asymptotically chi-squared distributed with degrees of freedom equal to the rank of the matrix of left hand side variables of the first type in the auxiliary regression. In the usual case where the cells are products of M_X cells partitioning X and M_Y cells partitioning Y, the number of degrees of freedom will be $(M_Y - 1)M_X$.

Different partitions of Y and X give different tests, with power in a different directions (directions that will lead to a different distribution over the cells in the chosen partition). We will use the partition of Y that is the basis for the nonparametric approach for comparing (DIF corrected) distributions of health in the given domain that will be described below. In a sense, this is a partition that "matters" since this kind of comparisons is one of the main goals of using anchoring vignettes. Since the test is asymptotic, the number of observations must be large to guarantee that the size of the test is approximately equal to the asymptotic size of 5%. In practice, this means that we will have to merge cells to guarantee that the number of observations in each cell is reasonably large. We have performed some simulations to compute the actual size of the test for various partitions and our choice of cells is based upon these simulation outcomes.

4 Nonparametric Approach

The nonparametric approach is explained by King *et al.* (2004) and King and Wand (2007). Its goal is to compare the distribution of the underlying objective reality purged for differences in reporting behaviour across countries or socio-economic groups. This is done by comparing where the self-assessments are placed on the scale fixed by the vignette evaluations in each country or group. A stylized numerical example with only one vignette is as follows.

Distribution (in %) of Self-assessments and Vignette Evaluations in Countries A and B

Country A	vignette					
	1	2	3	4	5	all
self assessment	(none)	(mild)	(moderate)	(severe)	(extr.)	
1 (no problem)	4	4	4	4	4	20
2 (mild problem)	4	4	4	4	4	20
3 (moderate problem)	4	4	4	4	4	20
4 (serious problem)	4	4	4	4	4	20
5 (extreme problem)	4	4	4	4	4	20
all	20	20	20	20	20	100

Country B	vignette					
	1	2	3	4	5	all
self assessment	(none)	(mild)	(moderate)	(severe)	(extr.)	
1 (no problem)	16	4	4	4	0	28
2 (mild problem)	8	4	4	4	0	20
3 (moderate problem)	8	4	4	4	0	20
4 (serious problem)	8	4	4	4	0	20
5 (extreme problem)	0	4	4	4	0	12
all	40	20	20	20	0	100

These cross-tabulations give the joint distributions of self-assessments and vignette evaluations of health problems in a given domain in (hypo-

thetical) countries A and B. Looking at the (marginal) distribution of the self-assessments only (the final column) would lead to the conclusion that respondents in country B face fewer problems in this health domain than respondents in country A – under the assumption that they use the same response scales. The difference in the marginal distribution of the vignette evaluations (the final row), however, shows that this assumption is incorrect: respondents in country A evaluate a given health problem as more problematic, on average, and this may be an alternative explanation for the cross-country difference in the self-assessments.

The nonparametric approach simply entails comparing the relative distributions (how do the self-assessments rank compared to the vignette evaluations?) in the two countries. The relative rankings (RR) are as follows:

Relative Ranking (RR) of Self-assessments
and Vignette Evaluations (in %) by Country

	Country A	Country B
RR=1: Self-ass. < Vignette eval.	40	24
RR=2: Self-ass. = Vignette eval.	20	28
RR=3: Self-ass. > Vignette eval.	40	48

The distribution of RR in country A is stochastically dominated by that in country B, showing that, once differences in response behaviour are accounted for, the health problems in country B appear to be more serious than in country A. This is the reverse of the conclusion based upon the self-assessments only.

The example above has only one vignette. King et al. (2004) consider the case with $K > 1$ vignettes, assuming that the evaluations of the vignettes are ranked in the same way by each respondent and that each respondent evaluates different vignettes differently. In that case a self-assessment can fit in any of $2K + 1$ positions in the given ranking of the vignette evaluations (better than any vignette or worse than any vignette (2 possibilities), between two vignettes ($K - 1$ possibilities), or equal to one of the K vignettes (K possibilities). The nonparametric approach then boils down to comparing

the distributions over the $2K+1$ positions across countries or socio-economic groups. See King et al. (2004, pp.195-196) for an empirical illustration.

King and Wand (2007) discuss the more realistic case with *ties*, that is, situations where a respondent assigns the same rating to several vignettes, or where a respondent rates the vignettes in a way that does not respect the ranking of the vignette evaluations used by the majority (which is often the natural ranking, given the wordings of the vignettes). For $K = 3$ vignettes, Table 1 presents a complete listing of all possible rankings of vignettes and self-assessments, generalizing Table 1 in King and Wand (2007). The natural ordering of the vignette ratings is $Y_1 < Y_2 < Y_3$. The seven situations in the left upper panel respect this ordering; the remainder looks at ties. Some of these are non-problematic since all that matters is the position of the self-assessment Y_s. For example, the situation $Y_s < Y_2 < Y_1 < Y_3$ puts Y_s in the same position as $Y_s < Y_1 < Y_2 < Y_3$. For the nonparametric comparison, the two will be merged, which is indicated by assigning 1 to both of them (column C). Similarly, $Y_2 < Y_1 < Y_s < Y_3$ puts Y_s in the same place as $Y_1 < Y_2 < Y_s < Y_3$, and both get 5. But in other situations, the position of Y_s is more ambiguous. For example, take the case $Y_3 < Y_s < Y_1 < Y_2$. If $Y_3 < Y_s$ (and Y_1 and Y_2 are misreported), we are in situation 7 (Y_s is worse than all vignettes), but if $Y_s < Y_1$ (and Y_3 is misreported), we are in situation 1 (Y_s better than all vignettes). We therefore cannot say anything about the true position of Y_s and classify this case as 1-7. Another example is $Y_s = Y_1 = Y_2 < Y_3$. In this case, due to rounding, we cannot distinguish whether Y_s is the same as Y_1, the same as Y_2, or in between the two, so we are in situation 2, 3 or 4. However, we can plausibly conclude that we are not in situations 1, 5, 6 or 7. This ranking is therefore coded as 2-4.

The nonparametric method categorizes observations into specific cells for each group or country. The 19 labels in Table 1 define a partition of the set Y of possible realizations of the observed dependent variables (self-assessments and vignette evaluations) into 19 cells. If the population consists of countries A and B and X is a country dummy (the only "regressor"), then the two countries form a partition of the set of all values of the regressor X. The nonparametric comparison then partitions $Y \times X$ into 19*2=38 cells. Ide-

ally, the number of observations in the twelve cells other than those labelled $1, 2, ..., 7$ should be so small that these cells can be discarded. This is helpful for the comparison because the position of Y_s in the remaining cells respects a clear ordering, and we can say that the distribution of the health domain considered is better in country A than in country B if the distribution over the cells $\{1, ..., 7\}$ in country A stochastically dominates that in country B.

To interpret the nonparametric results, we need response consistency and vignette equivalence, the two assumptions underlying the use of anchoring vignettes to correct for response scale differences. These are the identifying assumptions in this framework and we will consider them as maintained hypotheses; tests of response consistency and vignette equivalence are discussed elsewhere and are not the topic of this paper. See, for example, Van Soest *et al.* (2011) or Datta Gupta *et al.* (2010) for tests on response consistency using additional information in the form of a measure on an objective scale; see Peracchi and Rossetti (2010) for an analysis of vignette equivalence, exploiting overidentifying restrictions if respondents get more than one vignette; and see Bago d'Uva *et al.* (2011) or Corrado and Weeks (2010) who test response consistency conditional on vignette equivalence by testing the joint significance of covariates added to the equation for vignette responses.

Finally, we want to emphasize that the non-parametric approach has limited applicability compared to the parametric models and cannot replace them. It cannot deal with many covariates, nor produce counter-factual distributions of self-reported health with benchmark reporting scales (as in, e.g., Kapteyn *et al.* (2007)). It is useful only for making comparisons across a few socioeconomic groups or (groups of) countries, and only if such comparison is not hampered by ties that make it very difficult or impossible to interpret the non-parametric results.

5 Data

We use data from Survey of Health, Ageing and Retirement in Europe (SHARE) collected in 2004. SHARE is a broad socioeconomic survey among the population of ages 50 and older and their spouses in 11 European coun-

tries; see Börsch-Supan and Jürges (2005) for details on the design and set up. All respondents first got a personal interview and were then asked to complete a short paper and pencil questionnaire. In eight countries, Belgium, France, Germany, Greece, Italy, Netherlands, Spain, and Sweden, random subsamples were given an additional questionnaire with self-assessments and vignettes on several aspects of health (not in the context of work) and on work disability. Here we focus on the health questions, which were also used by, for example, Lardjane and Dourgnon (2007) and Bago d'Uva *et al.* (2008a). Self-assessments and three vignettes were collected for six health domains - breathing, concentration, depression, mobility, bodily pains and sleeping. The wordings of these questions are given in Appendix B. All questions use the same five-point scale: none, mild, moderate, severe, or extreme. The three vignettes in each domain are ordered, with one vignette (labellled $k = 1$) describing a mild health problem, the second describing a worse problem ($k = 2$), and the third describing the most severe health problem ($k = 3$). This order is used as the natural order in the nonparametric approach. About 50 percent of the respondents got all their vignette questions in the order from mild to severe; the other 50 percent got them in the reverse order. Vignette questions always came after the corresponding self-assessment. The total size of the vignette subsample is 4544 respondents. Due to missing observations, we have about 4370 respondents for each domain. Precise sample sizes and descriptive statistics for self-assessments and vignettes are presented in Table 2. For each domain, most respondents rate their own health problems as "none" or "mild." Severe or extreme health problems were reported by about 6.5 percent of the respondents, on average across the domains (from 3.57 percent for breathing to 9.23 percent for sleep). The majority reported no problem with mobility or breathing, but for the other domains, the "none" answers are a minority. In particular, pain problems are quite prevalent, with less than one third reporting "none."

The vignette evaluations reflect the level of the health problems of the hypothetical persons in the vignettes. As expected, the person in the third vignette in each domain was typically evaluated as least healthy, followed by the person in the second vignette.

SHARE is quite a rich survey, with many background variables collected for all respondents. For the parametric model in Section 2, we use the background variables in Table 3, which also presents some descriptive statistics. The average age of the respondents is 63 years. We distinguish three education levels; 35 percent of the sample obtained low education (ISCED levels 0 and 1), 45 percent intermediate education (ISCED 2 and 3), and 20 percent high education (ISCED 4, 5 or 6). Most respondents are women (55.6 percent) and do not live alone (74.4 percent).

6 Simulations: how to choose the cells?

Our tests rely on asymptotic theory keeping the number of cells fixed, with the number of observations going to infinity (see Andrews (1988)). As a consequence, the finite sample properties of the test may be poor if some cells have few observations. The same problem arises with the classical Pearson chi-squared test, where a common rule of thumb is that no more than 20% of the expected cell counts should be less than 5 and all expected counts should be at least 1 (Yates *et al.* (1999), p. 734). Such a rule of thumb is not available for the Andrews test. We performed some Monte Carlo simulations to determine whether, for several given choices of the cells, the actual size of the tests in our finite sample of about 4400 observations approximates the asymptotic size of 5%.

First, we estimated the CHOPIT model for one health domain – pain – using the actual data on the 4368 complete observations. The estimation results are presented in Table 1 of the Online Appendix. Using these estimates and the actual values of the covariates X, we generated 300 new data sets, all with the same covariates as the real data, but with different dependent variables (three vignette evaluations and one self-assessment for each observation), constructed from the values of the covariates and independent draws of the error terms in the CHOPIT model. These data sets are all generated using the CHOPIT model, so that they satisfy the null hypothesis of no misspecification. For each given choice of cells, we then perform the Andrews test on all 300 data sets, using a nominal size of 5%. The number of

times the null hypothesis is rejected divided by 300 is approximately equal to the actual size of the test for the given sample size and given choice of cells. (Approximately since 300 is not infinity; experimenting with larger numbers of simulated data sets did not change our conclusions.) The difference between actual and nominal (5%) size is due to the deviation between the finite sample and the asymptotic distribution of the test statistic.

The results are presented in Table 4. The final column (D10: 19 cells) uses the 19 cells of Y that form the basis of the nonparametric approach in Section 4, combined with several partitions of X. The actual size of the test is very different from the nominal size (5%) – it is always larger than 25% so that the tests heavily over reject. The reason is that cell sizes are often too small: even without partitioning X, 15 out of the 19 cells have expected cell size less than five, so that the rule of thumb for the simple Pearson chi-squared goodness of fit test is not satisfied at all. See Table 2 of the Online Appendix. The sizes of the cells corresponding to ties where the natural order of the evaluations of the three vignettes is not respected are particularly small (supporting the quality of data). The problem of small cell sizes gets even worse if we further split up the cells by partitioning X.

To avoid the problem of small cell sizes, we can merge cells and partition Y into fewer cells. This can be done in many different ways. Nine of them are presented in Table 5. The resulting actual sizes of the tests using these partitions of Y combined with various partitions of X are given in the additional columns of Table 4. For example, the column "D1 (3 cells)" partitions Y into three categories: rankings where the self-assessment unambiguously indicates less pain problems than all three vignettes, rankings where the self-assessment is rated the same as the vignette with the least pain problems, and all other rankings. The former two are by far the most frequent rankings in the data (with 34% and 24% of the observations). The column "D9 (10 cells)" takes the seven rankings that respect the natural vignette ordering as separate cells, and merges the 13 cells of Y into three. This is already enough to guarantee that the rule of thumb that at most 20% of all cells have expected cell size less than five is satisfied. The other columns are intermediate cases where Y is partitioned in four to nine cells. When Y is partitioned

into at most six cells, the expected cell size is always larger than ten; in the other cases, the expected cell size is often between five and ten, particularly if X is partitioned into three cells using education.

The results show that merging cells help to bring the actual size of the tests closer to the nominal size of 5%. The actual size varies between 4.7% (Y partitioned into 9 cells; X partitioned using gender) and 11.3% (Y partitioned in 10 cells; X partitioned using education). The test still tends to reject too often: the actual size is almost always larger than 5%, and increases somewhat when the number of cells in Y is larger than six. It seems safe to conclude that we can perform the tests using the given partitions of Y into three to six cells, though we should take into account that the actual size of the test with nominal size 5% varies from around 5% to about 10%.

7 Empirical Results

This section presents the test results for each domain and for various partitions of Y and X. At the end of the section we also perform some sensitivity checks. For each health domain, we first estimated the parametric model of Section 2. As an example, the parameter estimates for concentration and memory skills are presented in Table 6; estimates for the other domains are available upon request. The first column shows how, according to the parametric model, problems with concentration and memory skills are associated with individual characteristics and country dummies, keeping response scales constant. Most results here are plausible and confirm findings in the literature (e.g., Bago d'Uva *et al.* (2008a)). For example, problems fall with education and rise with age. There are substantial differences across countries. In particular, respondents suffer much less from problems with concentration and memory skills in Sweden than in other countries.

The other columns present the estimates of the parameters determining the thresholds. Many variables are significant, implying that not accounting for DIF would lead to biased estimates of the parameters of main interest in the first column. Particularly the estimates of γ_1 are important since $X_i\gamma_1$ affects all thresholds in the same way (see equation 3). They imply that, for

example, Swedish respondents use lower thresholds than others, so that they tend to evaluate a given concentration and memory problem as more serious than respondents in other countries. Correcting their self-assessments for this makes them even better off than the self-assessment data would suggest. These findings are not new to the current paper – similar models have been analyzed in, for example, Bago d'Uva *et al.* (2008a). The vignette dummies in the bottom panel have the expected ranking, corresponding to the fact that the first vignette describes the mildest problem, etc.

Finally, note that the standard deviation of the unobserved heterogeneity term is quite precisely estimated, with a 95 percent confidence interval [0.380, 0.426]. This suggests that extending the standard CHOPIT model with this unobserved heterogeneity term is useful, even though the role of this unobserved heterogeneity term is smaller than the roles of the noise terms ϵ_{si} and $\epsilon_{vi}, v = 1, \ldots, 3$.

This paper does not focus on parameter estimates but on the specification tests. For each misspecification test characterized by a different partition of $Y \times X$, the cell probabilities according to the parametric model were computed (numerically), using the given values of the regressors for all observations in the sample and the estimated coefficients of the parametric model. Each test compares such a distribution generated by the parametric model with the corresponding distribution in the raw data.

We performed many tests for many different partitions and health domains and two different model specifications, and the test statistics are often not independent of each other since they rely on the same data (with the same or correlated dependent variables). The results should be interpreted with some care due to the issue of multiple hypothesis tests – the tests should be interpreted in isolation and different tests cannot be combined into a joint conclusion about some common hypothesis.

The test results for the standard CHOPIT model are easy to summarize (and do not require a table): the null hypothesis of a correctly specified model is always rejected at the 5% or even the 1% level, for all health domains and for all partitions of $Y \times X$. The p-values for the CHOPIT model extended with unobserved heterogeneity in the thresholds show more variation and are

presented in Table 7.

First consider the tests partitioning Y only, not using X. For concentration and memory skills, the null is not rejected at the 1% level for any of the partitions, and not at the 5% level for the three partitions with the smallest numbers of cells. And if we take into account that the simulations have shown that for the given sample sizes the tests tend to over reject, even the p-values of 0.029 and 0.030 seem to be supportive of this parametric model. For breathing, mobility, and sleep, the null hypothesis of a correct parametric specification is always rejected, no matter which partition of Y is used. For the other two domains, the results are mixed: the null is not rejected at the 5% level for a partition into only three cells (depression) or into three or four cells (pain), but it is rejected at even the 1% level for the partitions with more cells.

For the partition of Y into four cells, the observed and predicted probabilities on which the tests are both presented in Table 8. For pain, concentration and depression the maximum difference is about 1 percentage point; for mobility and sleep it increases to 1.5 percentage points and for breathing to 2.4 percentage points. Comparing the two distributions suggests that the differences are not that big even for domains such as breathing where the null is firmly rejected. Because of the large sample sizes, however, even these modest differences can apparently be sufficient to reject the null hypothesis.

The tests using partitions of Y only essentially explore whether the parametric model is able to reproduce the ranking distribution of vignette evaluations and self-assessments, a feature of the marginal distribution of the dependent variables, not involving any regressors. This does not yet correspond to the nonparametric approach – which compares two such rankings, distinguished on the basis of X (for example, two countries or groups of countries; men and women; high and low educated respondents; etc.). This is why we also want to consider partitions of $Y \times X$. Each different partition of X gives a test with power in a specific direction, corresponding to the cells used by the nonparametric approach for comparing specific groups.

The remaining rows in Table 7 present the p-values for such tests. To guarantee that sample sizes are large enough, we only consider partitions of

X into two or three cells, leading to a partition of $Y \times X$ into between $2 \times 3 = 6$ and $3 \times 6 = 18$ cells. First, since cross-country comparison is what anchoring vignettes have traditionally been used for (cf. King et al. (2004)), we consider a partition $Y \times$ country, where countries are divided into two groups - southern Europe (Greece, Spain and Italy) and the remaining countries (Belgium, France, Germany, Netherlands and Sweden). This north-south division corresponds to the systematic differences found in many SHARE studies; see, for example, many chapters in Börsch-Supan *et al.* (2005). In addition, we perform the test for partitions $Y \times$ sex; $Y \times$ age with age categorized into younger than 56, 56-65, and older than 65; $Y \times$ education, with education categorized into low, middle and high; and finally $Y \times$ alone, distinguishing respondents living alone or not.

As expected (based on the results for the partition of Y only) the null hypothesis is rejected at the 5% or even the 1% significance level for all considered partitions of $Y \times X$ for breathing, sleep and mobility: if the parametric model is already not able to reproduce the marginal distribution over the Y cells, it cannot give a good fit to the bivariate $Y \times X$ cells either. For the other domains, pain, concentration and memory skills, and depression, the results are mixed. P-values exceeding 5 percent are found for partitions $Y \times$ sex and $Y \times$ age for pain and for concentration and memory skills. On the other hand, for all domains, the null hypothesis is rejected for partitions using a partition of X by country group or by education level. The tests therefore do not support the use of the parametric model for comparison across (southern versus northern) countries or different education groups.

Table 9 presents the predicted and actual distribution over the four cells in partition D2 for southern and northern countries, illustrating the magnitude of the differences. In most cases, the differences do not affect the qualitative conclusions on cross-country differences. For breathing for example, 83.8% of southern respondents report less problems than any of the vignettes, compared to 63.8% in the north, suggesting that, after correcting for response scale differences, people aged 50 or older have larger problems with breathing in the North than in the South. Using the parametric model's predictions, the percentages are 83.0% and 65.9% and the conclusion remains

the same. Similar comparisons for other domains lead to the same qualitative conclusion: whether using the raw data or the parametric predictions sometimes changes where a cell frequency is larger, but does not affect our conclusion on which countries have better health in the given domain. Take sleep, for example: according to the raw data, 7.3% have similar problems as the best vignette in the South, compared to 12.2% in the North. The parametric model predicts the reverse order: 10.6% in the South, 10.3% in the North. Still, if we look at the rankings as a whole, we would conclude from the raw data as well as the parametric predictions that sleep problems in the North and South are very similar. Similarly, whether using the raw data or the parametric predictions, we find that pain and memory and concentration problems are clearly larger in the South, depression related health problems are slightly larger in the South, mobility is somewhat worse in the North.

Tables 3-6 of the Online Appendix make the same comparisons using other partitions of X. Again, the raw data and the predictions using the parametric model sometimes lead to different conclusions about some of the relative cell sizes by population subgroup, but generally lead to the same conclusion on which group is healthier in the given domain.

To get more insight why the tests often reject the parametric model, we performed several sensitivity checks. First, Table 9 indicates that predicted and actual cell frequencies are not hugely different, but the differences are apparently large enough to reject the null hypothesis. A possible reason for this is that the Andrews test accounts for the fact that parameters are estimated so as to fit the (same) data, by including the likelihood scores in the auxiliary regression used to compute the test statistic (see Section 3), increasing the value of the test statistic. To see to which extent this matters, we also computed the test statistic without the likelihood scores. Table 7 of the Online Appendix shows the resulting p-values, which are often much higher than the correct p-values in Table 7. For sleep, for example, the fact that parameters are estimated apparently makes it likely that under the null, predicted and observed frequencies are very similar. Adjusting for this implies that the null hypothesis is already rejected for quite modest differences between predicted and actual cell sizes. For breathing, on the

other hand, the p-values remain virtually zero for all partitions. Here the discrepancies are such that the null would also be firmly rejected if parameter values were given instead of estimated.

Finally, to study whether the results of the tests are driven by wrongly ordered vignettes (leading to ties in the rankings that are typically not used when interpreting the nonparametric results; see Section 4), we re-estimated the parametric model using the observations with $Y_1 \leq Y_2 \leq Y_3$ only, and re-computed p-values for all partitions in Table 7. The p-values for this subsample are similar to those for the whole sample (see Table 8 of the Online Appendix). It therefore seems unlikely that wrongly ordered vignettes are the source of the misspecification of the parametric model.

8 Conclusion

Comparing self-reported survey measures of well-being, health, or other aspects of perceived quality of life or society often suffers from the fact that different groups use different reporting scales (DIF). Anchoring vignettes are an increasingly popular tool to correct for these differences. In the literature, there are two ways to use anchoring vignettes: parametric models (the CHOPIT model and its extensions) and a nonparametric approach based upon ranking vignette evaluations and self-assessments in subsamples characterized by values of control variables (such as country, age, gender, or education level). In this paper, we consider tests for misspecification of the parametric model based upon comparing the distribution generated by the parametric model with the actual distribution of the dependent variables. An attractive feature of our tests is that they compare exactly those features of this distribution that drive the nonparametric method. In that sense, the tests have power in directions that matter: rejecting the null implies that the rankings of self-assessments and vignette evaluations implied by the parametric model are inconsistent with the rankings in the raw data used for the nonparametric approach.

We apply the tests using data on the 50+ population in eight European countries, with self-assessments and three vignettes on six domains of health.

A simulation study demonstrates that cells need to be combined to guarantee a reasonable finite sample performance of the tests, leading to tests with actual size between 5% and 10% where the nominal asymptotic size is 5%. Our results are mixed. The specification of the standard CHOPIT model is always rejected, but the CHOPIT model extended with unobserved heterogeneity in the reporting scales performs better. For some socioeconomic characteristics and some health domains, we cannot reject the null hypothesis that the parametric model generates the same distributions of the rankings as the raw data. But in other cases, even the marginal distribution of the rankings of vignettes and self-assessments is captured poorly enough by the parametric model to reject the null of a correct specification, even though the cell frequencies of the rankings predicted by the parametric model are not hugely different from the observed frequencies in the data and typically do not change the conclusions on which countries or demographic groups are healthier in the given domain.

What does this imply for studies using anchoring vignettes? Not that the parametric models should be replaced by the non-parametric method: As explained in Section 4 the non-parametric approach has limited applicability and gives results that are hard to interpret when there are many ties. The latter seems a major reason for considering parametric models: if vignettes are not ordered consistently by all respondents, the ties make it impossible to draw conclusions from the nonparametric comparisons. This problem does not arise in the parametric models, where idiosyncratic errors can explain any violation of the natural ordering of the vignette evaluations.

The fact that the standard CHOPIT model is always rejected does not necessarily mean it does not help to correct for differences in reporting behaviour (DIF). In fact, the modest size of the differences between predicted and actual rankings suggests it does. But there is room for improvement: more flexible parametric or semi-parametric models may reduce the misspecification bias and lead to better corrections for DIF. Moreover, the misspecification of the CHOPIT model also implies that the tests for vignette equivalence or response consistency developed in the recent literature, which rely on the CHOPIT model as a maintained auxiliary hypothesis, may not be

valid: the null hypothesis of (e.g.) response consistency could be rejected not because of lack of response consistency but because the parametric model in which the test is applied is misspecified. Using more flexible models than the standard CHOPIT model as a basis for such tests can give more robust tests for response consistency and vignette equivalence.

We interpret our findings as a motivation for future work on developing and applying flexible models that are more general than standard CHOPIT but share its advantages – models that can be used to construct counterfactual distributions in combination with many covariates, and that can deal with idiosyncratic errors and the ties that result from them in the same way as the CHOPIT model. The CHOPIT model with unobserved heterogeneity in the thresholds is a simple example of such an extension: with only one additional parameter, it already helps to reduce misspecification problems substantially. But many other extensions of the CHOPIT model can be considered. Error terms could be heteroskedastic and systematic parts could be made more flexible using interactions and polynomial expansions. Unobserved heterogeneity can be incorporated with a more flexible discrete distribution in the spirit of Heckman and Singer (1984), or error terms can be given more general, non-normal, distributions, using, for example, semi-nonparametric specifications as in Gallant and Nychka (1987). Exploring which extensions help to pass the misspecification tests is beyond the scope of the current paper and remains a topic of future research.

References

Andrews, D. W. K. (1988) Chi-square diagnostic tests for econometric models. *Journal of Econometrics*, **37**, 135–156.

Angelini, V., Cavapozzi, D., Corazzini, L. and Paccagnella, O. (2012) Age, health and life satisfaction among older Europeans. *Social Indicators Research*, **105**, 293–308.

Bago d'Uva, T., Lindeboom, M., O'Donnell, O. and Van Doorslaer, E. (2011) Slipping anchor? Testing the vignettes approach to identification and cor-

rection of reporting heterogeneity. *Journal of Human Resources*, **46**, 872–903.

Bago d'Uva, T., O'Donnell, O. and Van Doorslaer, E. (2008a) Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. *International Journal of Epidemiology*, **37**, 1375–1383.

Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M. and O'Donnell, O. (2008b) Does reporting heterogeneity bias the measurement of health disparities? *Health Economics*, **17**, 351–375.

Bonsang, E. and Van Soest, A. (2012) Satisfaction with social contacts of older Europeans. *Social Indicators Research*, **105**, 273–292.

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

Börsch-Supan, A. and Jürges, H. (2005) *The Survey of Health, Ageing, and Retirement in Europe - Methodology*. Mannheim Research Institute for the Economics of Aging (MEA).

Chesher, A. and Irish, M. (1987) Residual analysis in the grouped and censored normal linear model. *Journal of Econometrics*, **34**, 33–61.

Corrado, L. and Weeks, M. (2010) Identification strategies in survey response using vignettes. Cambridge Working Papers in Economics 1031, University of Cambridge.

Datta Gupta, N., Kristensen, N. and Pozzoli, D. (2010) External validation of the use of vignettes in cross-country health studies. *Economic Modelling*, **27**, 854–867.

Gallant, R. and Nychka, D. (1987) Semi-nonparametric maximum likelihood estimation. *Econometrica*, **55**, 363–390.

Heckman, J. and Singer, B. (1984) A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, **52**, 271–320.

Holland, P. and Wainer, H. (1993) *Differential item functioning*. Hillsdale, N.J.: Lawrence Erlbaum.

Kapteyn, A., Smith, J. and Van Soest, A. (2007) Vignettes and self-reports of work disability in the U.S. and the Netherlands. *American Economic Review*, **97**, 461–473.

King, G., Murray, C. J. L., Salomon, J. and Tandon, A. (2004) Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, **98**, 191–207.

King, G. and Wand, J. (2007) Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, **15**, 46–66.

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

Lardjane, S. and Dourgnon, P. (2007) Les comparaisons internationales d'état de santé subjectif sont-elles pertinentes? Une évaluation par la méthode des vignettes-étalons. *Economie et Statistique*, **403-404**, 165–177.

Peracchi, F. and Rossetti, C. (2010) The heterogeneous thresholds ordered response model: Identification and inference. Mimeo, Tor Vergata University.

Rice, N., Robone, S. and Smith, P. (2010) International comparison of public sector performance: The use of anchoring vignettes to adjust self-reported data. *Evaluation*, **16**, 81–101.

Van Soest, A., Delaney, L., Harmon, C. P., Kapteyn, A. and Smith, J. P. (2011) Validating the use of anchoring vignettes for the correction of re-

sponse scale differences in subjective questions. *Journal of the Royal Statistical Society Series A*, **174**, 575–595.

Vonkova, H. and Hullegie, P. (2011) Is the anchoring vignette method sensitive to the domain and the choice of the vignette. *Journal of the Royal Statistical Society Series A*, **174**, 597–620.

Yates, D., Moore, D. and McCabe, G. (1999) *The practice of statistics (1st Ed.)*. New York: W.H. Freeman.

A Tables

Table 1: Rankings of self-assessment Y_s and vignette evaluations Y_1, Y_2, Y_3

ranking	C	C label	ranking	C	C label
$Y_s < Y_1 < Y_2 < Y_3$	1	1	$Y_s < Y_1 < Y_3 < Y_2$	1	1
$Y_s = Y_1 < Y_2 < Y_3$	2	2	$Y_s = Y_1 < Y_3 < Y_2$	2	2
$Y_1 < Y_s < Y_2 < Y_3$	3	3	$Y_1 < Y_s < Y_3 < Y_2$	3	3
$Y_1 < Y_s = Y_2 < Y_3$	4	4	$Y_1 < Y_s = Y_3 < Y_2$	3-6	16
$Y_1 < Y_2 < Y_s < Y_3$	5	5	$Y_1 < Y_3 < Y_s < Y_2$	3-7	17
$Y_1 < Y_2 < Y_s = Y_3$	6	6	$Y_1 < Y_3 < Y_s = Y_2$	4-7	19
$Y_1 < Y_2 < Y_3 < Y_s$	7	7	$Y_1 < Y_3 < Y_2 < Y_s$	7	7
$Y_s < Y_2 < Y_1 < Y_3$	1	1	$Y_s < Y_2 < Y_3 < Y_1$	1	1
$Y_s = Y_2 < Y_1 < Y_3$	1-4	8	$Y_s = Y_2 < Y_3 < Y_1$	1-4	8
$Y_2 < Y_s < Y_1 < Y_3$	1-5	9	$Y_2 < Y_s < Y_3 < Y_1$	1-5	9
$Y_2 < Y_s = Y_1 < Y_3$	2-5	13	$Y_2 < Y_s = Y_3 < Y_1$	1-6	10
$Y_2 < Y_1 < Y_s < Y_3$	5	5	$Y_2 < Y_3 < Y_s < Y_1$	1-7	11
$Y_2 < Y_1 < Y_s = Y_3$	6	6	$Y_2 < Y_3 < Y_s = Y_1$	2-7	15
$Y_2 < Y_1 < Y_3 < Y_s$	7	7	$Y_2 < Y_3 < Y_1 < Y_s$	7	7
$Y_s < Y_3 < Y_1 < Y_2$	1	1	$Y_s < Y_3 < Y_2 < Y_1$	1	1
$Y_s = Y_3 < Y_1 < Y_2$	1-6	10	$Y_s = Y_3 < Y_2 < Y_1$	1-6	10
$Y_3 < Y_s < Y_1 < Y_2$	1-7	11	$Y_3 < Y_s < Y_2 < Y_1$	1-7	11
$Y_3 < Y_s = Y_1 < Y_2$	2-7	15	$Y_3 < Y_s = Y_2 < Y_1$	1-7	11
$Y_3 < Y_1 < Y_s < Y_2$	3-7	17	$Y_3 < Y_2 < Y_s < Y_1$	1-7	11
$Y_3 < Y_1 < Y_s = Y_2$	4-7	19	$Y_3 < Y_2 < Y_s = Y_1$	2-7	15
$Y_3 < Y_1 < Y_2 < Y_s$	7	7	$Y_3 < Y_2 < Y_1 < Y_s$	7	7
$Y_s < Y_1 = Y_2 < Y_3$	1	1	$Y_s < Y_3 < Y_1 = Y_2$	1	1
$Y_s = Y_1 = Y_2 < Y_3$	2-4	12	$Y_s = Y_3 < Y_1 = Y_2$	1-6	10
$Y_1 = Y_2 < Y_s < Y_3$	5	5	$Y_3 < Y_s < Y_1 = Y_2$	1-7	11
$Y_1 = Y_2 < Y_s = Y_3$	6	6	$Y_3 < Y_s = Y_1 = Y_2$	2-7	15
$Y_1 = Y_2 < Y_3 < Y_s$	7	7	$Y_3 < Y_1 = Y_2 < Y_s$	7	7
$Y_s < Y_1 = Y_3 < Y_2$	1	1	$Y_s < Y_2 < Y_1 = Y_3$	1	1
$Y_s = Y_1 = Y_3 < Y_2$	2-6	14	$Y_s = Y_2 < Y_1 = Y_3$	1-4	8
$Y_1 = Y_3 < Y_s < Y_2$	3-7	17	$Y_2 < Y_s < Y_1 = Y_3$	1-5	9
$Y_1 = Y_3 < Y_s = Y_2$	4-7	19	$Y_2 < Y_s = Y_1 = Y_3$	2-6	14
$Y_1 = Y_3 < Y_2 < Y_s$	7	7	$Y_2 < Y_1 = Y_3 < Y_s$	7	7
$Y_s < Y_1 < Y_2 = Y_3$	1	1	$Y_s < Y_2 = Y_3 < Y_1$	1	1
$Y_s = Y_1 < Y_2 = Y_3$	2	2	$Y_s = Y_2 = Y_3 < Y_1$	1-6	10
$Y_1 < Y_s < Y_2 = Y_3$	3	3	$Y_2 = Y_3 < Y_s < Y_1$	1-7	11
$Y_1 < Y_s = Y_2 = Y_3$	4-6	18	$Y_2 = Y_3 < Y_s = Y_1$	2-7	15
$Y_1 < Y_2 = Y_3 < Y_s$	7	7	$Y_2 = Y_3 < Y_1 < Y_s$	7	7
$Y_s < Y_1 = Y_2 = Y_3$	1	1	$Y_s = Y_1 = Y_2 = Y_3$	2-6	14
$Y_1 = Y_2 = Y_3 < Y_s$	7	7			

Note: There are 19 different cells according to nonparametric approach. We define their labels (column "C label") as follows: if C=1,...,7 labels remain the same. In case of ties C=1-4, 1-5, 1-6, 1-7, 2-4, 2-5, 2-6, 2-7, 3-6, 3-7, 4-6, 4-7 labels are defined as 8, 9, ..., 19, resp.

Table 2: Distributions of self-assessments and vignette evaluations

	breathing				concentration			
	s	v_1	v_2	v_3	s	v_1	v_2	v_3
none	64.58	10.81	2.29	2.45	43.98	22.15	5.25	2.01
mild	22.23	24.11	5.15	2.22	35.17	48.63	27.08	8.90
moderate	9.62	38.00	19.76	8.59	16.22	22.79	44.37	29.79
severe	3.04	24.08	52.20	44.25	4.24	6.09	20.73	47.33
extreme	0.53	3.00	20.60	42.49	0.39	0.34	2.58	11.98
	depression				mobility			
	s	v_1	v_2	v_3	s	v_1	v_2	v_3
none	49.51	6.45	2.31	2.20	58.39	9.64	2.31	1.55
mild	28.70	44.17	13.44	2.59	22.17	34.73	11.83	5.91
moderate	14.97	36.19	45.75	10.80	13.04	42.84	38.80	27.49
severe	5.29	11.56	33.53	42.57	5.23	11.97	40.37	48.80
extreme	1.53	1.63	4.97	41.84	1.16	0.82	6.69	16.24
	pain				sleep			
	s	v_1	v_2	v_3	s	v_1	v_2	v_3
none	32.28	15.59	2.31	1.12	42.68	2.65	1.92	1.87
mild	35.81	56.96	18.06	5.31	28.06	21.50	9.73	7.31
moderate	23.10	22.09	50.76	26.10	20.04	47.98	29.02	27.00
severe	7.14	4.78	25.71	48.63	7.36	24.03	42.49	41.70
extreme	1.67	0.57	3.16	18.84	1.87	3.84	16.84	22.12

Note: s is self-assessment; v_1, v_2 and v_3 are vignettes 1, 2 and 3, resp. The size of the vignette subsample of the SHARE sample is 4544. We work with around 4370 respondents for each domain (4368 for breathing, 4384 for concentration, 4369 for depression, 4379 for mobility, 4368 for pain and 4377 for sleep).

Table 3: Background variables (Mean (age) or percentage equal to 1 (other variables)

Belgium	12.48	male	44.43
France	19.48	education low	35.54
Germany	11.18	education mid	44.60
Greece	15.85	education high	19.86
Italy	9.79	age - mean	63.06
Netherlands	11.84	age - std dev	10.01
Spain	10.21	alone	25.58
Sweden	9.18	not alone	74.42

Note: The descriptive statistics are given for the 4544 respondents in the vignette subsample of SHARE 2004. *Education mid* corresponds to the ISCED levels 2 and 3, *education high* to the levels 4, 5 and 6. *Not alone* is based on observed marital status of respondent, it corresponds to categories married and living together with spouse, or living together with registered partnership. *Alone* corresponds to all other categories.

Table 4: Simulation results: percentage of rejections of the null hypothesis at significance level 5%

	partitions									
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
	3 cells	4 cells	5 cells	5 cells	6 cells	7 cells	8 cells	9 cells	10 cells	19 cells
	partitions based on Y only									
	8.33	7.33	7.67	7.33	7.33	10.00	10.67	8.00	8.67	25.33
	partitions based on $Y \times X$									
X										
sex	7.67	6.00	6.67	6.33	5.33	6.33	5.67	4.67	6.67	70.33
country	7.67	6.67	6.67	6.67	6.67	8.33	9.67	8.67	9.33	64.67
alone	8.67	6.33	6.67	7.00	6.33	9.00	9.00	9.33	10.33	77.67
education	8.33	7.67	9.00	8.67	8.00	10.00	8.67	10.00	11.33	96.67
age	6.33	7.00	6.67	7.00	6.67	9.33	9.33	10.33	11.00	93.00

Note: How the simulations are carried out and the p-values are obtained is explained in Section 6. The partitions $D1$ – $D10$ are defined in Table 5. For partitions $Y \times X$, countries are divided into two groups - southern Europe (Greece, Spain and Italy) and the remaining countries (Belgium, France, Germany, Netherlands and Sweden), age is categorized into younger than 56, 56-65, and older than 65 and education is categorized into low, middle and high.

Table 5: Merging 19 original cells into larger cells

division	nr larger cells	description of merging 19 original cells
D1	3	$\{1\},\{2\},\{3,\ldots,19\}$
D2	4	$\{1\},\{2\},\{3,4,8,\ldots,14\},\{5,6,7,15,\ldots,19\}$
D3	5	$\{1\},\{2\},\{3,4\},\{5,6,7\},\{8,\ldots,19\}$
D4	5	$\{1,2\},\{3,4\},\{5,6,7\},\{8,\ldots,15\},\{16,\ldots,19\}$
D5	6	$\{1\},\{2\},\{3,4\},\{5,6,7\},\{8,\ldots,15\},\{16,\ldots,19\}$
D6	7	$\{1\},\{2\},\{3,4\},\{5,6,7\},\{8,\ldots,11\},\{12,\ldots,15\},\{16,\ldots,19\}$
D7	8	$\{1\},\{2\},\{3\},\{4\},\{5,6,7\},\{8,\ldots,11\},\{12,\ldots,15\},\{16,\ldots,19\}$
D8	9	$\{1\},\{2\},\{3\},\{4\},\{5,6\},\{7\},\{8,\ldots,11\},\{12,\ldots,15\},\{16,\ldots,19\}$
D9	10	$\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8,\ldots,11\},\{12,\ldots,15\},\{16,\ldots,19\}$
D10	19	$\{1\},...,\{19\}$

Note: For the description labels of 19 original cells are used (see Table 1).

Table 6: Estimates of the parametric model for concentration

	β_s		γ_1		γ_2		γ_3		γ_4	
	coef	t-val	coef	t-val	coef	t-val	coef	t-val	coef	t-val
const	0	-	-0.1848	-3.9786	0.2037	5.5851	0.1117	2.8092	0.1820	3.1522
Belgium	0.0362	0.5715	-0.0961	-2.0688	0.1161	3.0912	-0.1385	-3.4545	-0.0849	-1.3969
Germany	0.1044	1.5750	0.2324	5.0074	-0.1477	-3.6080	-0.2057	-4.8098	-0.0913	-1.4099
Greece	0.0489	0.8311	0.4653	11.6894	-0.2131	-5.6407	-0.3526	-9.0218	-0.1239	-2.1732
Italy	0.1160	1.7402	0.2563	5.6073	-0.1153	-2.7832	-0.2790	-6.6135	-0.0896	-1.3732
Netherlands	-0.2155	-3.2028	-0.0843	-1.7336	0.0843	2.0973	-0.2964	-6.5579	-0.3917	-6.4416
Spain	-0.1173	-2.0938	0.0171	0.3475	-0.2393	-5.2428	-0.1450	-3.8082	0.2399	4.2768
Sweden	-0.8136	-11.3444	-0.4506	-6.6301	-0.4259	-6.5886	-0.3165	-6.0508	0.1623	3.2131
male	-0.0819	-2.1414	0.0560	2.2164	-0.0346	-1.5160	0.0076	0.3226	-0.0459	-1.5449
age 55-	-0.0145	-0.3038	0.0457	1.4469	-0.0341	-1.1783	-0.0448	-1.4627	0.0024	0.0615
age 66-75	0.1942	3.9206	-0.0862	-2.6178	0.0162	0.5655	0.0764	2.6049	0.0434	1.1049
age 76+	0.4663	7.4514	-0.0352	-0.8465	-0.0236	-0.5972	0.0371	0.9748	0.0849	1.7471
education low	0.2557	6.2358	0.0807	2.6240	-0.1078	-3.8602	0.0121	0.4370	-0.0562	-1.4797
education high	-0.1710	-3.1905	0.0046	0.1325	-0.0381	-1.2653	-0.0771	-2.2560	-0.0066	-0.1679
alone	0.1622	3.7273	0.0302	1.0282	-0.0396	-1.4765	-0.0065	-0.2473	0.0648	1.8565

	coef	se
θ_{v_1}	0.5436	0.0250
θ_{v_2}	1.3237	0.0237
θ_{v_3}	2.0452	0.0294
σ_s	1	-
σ_u	0.4030	0.0115
σ_v	0.6968	0.0151

Table 7: Goodness of fit for all health domains

	breathing					concentration				
	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5
Y only	*r*	*r*	*r*	*r*	*r*	0.106	0.212	0.187	0.029	0.030
$Y\times$ sex	*r*	*r*	*r*	*r*	*r*	0.172	0.336	0.210	0.085	0.068
$Y\times$ country	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*	0.011	*r*
$Y\times$ alone	*r*	*r*	*r*	*r*	*r*	0.027	0.073	0.122	0.119	0.025
$Y\times$ education	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*
$Y\times$ age	*r*	*r*	*r*	*r*	*r*	0.099	0.268	0.074	0.007	0.014

	depression					mobility				
	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5
Y only	0.050	0.008	0.003	*r*	*r*	*r*	*r*	*r*	*r*	*r*
$Y\times$ sex	*r*	*r*	*r*	*r*	*r*	*r*	0.002	0.001	*r*	*r*
$Y\times$ country	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*	0.011	*r*
$Y\times$ alone	0.139	0.024	0.013	*r*	*r*	*r*	*r*	*r*	*r*	*r*
$Y\times$ education	*r*	*r*	*r*	*r*	*r*	0.001	0.004	*r*	*r*	*r*
$Y\times$ age	0.012	0.006	0.004	0.002	*r*	0.003	0.007	0.001	*r*	*r*

	pain					sleep				
	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5
Y only	0.168	0.304	0.009	0.003	0.003	*r*	*r*	*r*	*r*	*r*
$Y\times$ sex	0.064	0.176	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*
$Y\times$ country	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*
$Y\times$ alone	0.030	0.051	0.001	*r*	*r*	*r*	*r*	*r*	*r*	*r*
$Y\times$ education	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*	*r*
$Y\times$ age	0.364	0.524	0.002	*r*	*r*	*r*	*r*	*r*	*r*	*r*

Note: "r" labels that the null hypothesis that the parametric model is correctly specified is rejected on significance level < 0.001. All results are presented for the extended CHOPIT model with unobserved heterogeneity. For the definition of D1,. . ., D5 see Table 5.

Table 8: Observed and predicted distributions (cells constructed using partition D2 of Y only)

	breathing		concentration		depression	
D2 cells	observed	predicted	observed	predicted	observed	predicted
$\{1\}$	71.09	72.12	38.12	38.84	57.73	58.60
$\{2\}$	13.74	14.60	23.56	22.53	17.01	16.64
$\{3, 4, 8, \ldots, 14\}$	12.13	9.73	26.32	26.58	18.59	17.39
$\{5, 6, 7, 15, \ldots, 19\}$	3.04	3.55	12.00	12.05	6.68	7.38

	mobility		pain		sleep	
	observed	predicted	observed	predicted	observed	predicted
$\{1\}$	63.26	63.03	34.43	34.93	61.37	62.93
$\{2\}$	13.25	14.79	23.81	22.79	10.40	10.40
$\{3, 4, 8, \ldots, 14\}$	15.00	13.59	25.09	25.50	17.61	16.00
$\{5, 6, 7, 15, \ldots, 19\}$	8.50	8.59	16.67	16.80	10.62	10.67

Note: Predicted distributions are presented for the extended CHOPIT model with unobserved heterogeneity.

Table 9: Example of observed and predicted distributions (cells constructed using partition D2 of $Y\times$ country)

	South		North	
	observed	predicted	observed	predicted
D2 cells		breathing		
$\{1\}$	83.79	82.96	63.80	65.92
$\{2\}$	7.30	10.40	17.44	16.99
$\{3,4,8,\ldots,14\}$	6.72	5.35	15.23	12.24
$\{5,6,7,15,\ldots,19\}$	2.19	1.29	3.53	4.85
		concentration		
$\{1\}$	35.80	33.17	39.44	42.12
$\{2\}$	20.95	24.01	25.06	21.67
$\{3,4,8,\ldots,14\}$	28.34	28.44	25.17	25.50
$\{5,6,7,15,\ldots,19\}$	14.91	14.38	10.32	10.70
		depression		
$\{1\}$	57.02	56.80	58.15	59.62
$\{2\}$	13.91	16.56	18.78	16.69
$\{3,4,8,\ldots,14\}$	20.42	18.15	17.52	16.95
$\{5,6,7,15,\ldots,19\}$	8.65	8.49	5.55	6.74
		mobility		
$\{1\}$	65.90	64.25	61.73	62.33
$\{2\}$	11.96	14.96	13.98	14.69
$\{3,4,8,\ldots,14\}$	13.54	12.89	15.86	13.98
$\{5,6,7,15,\ldots,19\}$	8.60	7.89	8.43	8.99
		pain		
$\{1\}$	33.04	30.07	35.24	37.74
$\{2\}$	20.55	23.36	25.70	22.45
$\{3,4,8,\ldots,14\}$	26.06	27.34	24.53	24.44
$\{5,6,7,15,\ldots,19\}$	20.35	19.24	14.52	15.37
		sleep		
$\{1\}$	64.16	64.28	59.76	62.15
$\{2\}$	7.29	10.64	12.19	10.26
$\{3,4,8,\ldots,14\}$	17.82	14.95	17.49	16.61
$\{5,6,7,15,\ldots,19\}$	10.73	10.12	10.56	10.99

Note: Predicted distributions are presented for the extended CHOPIT model with unobserved heterogeneity.

B Self-assessment questions and vignettes

Vignettes are in increasing order of the seriousness of the health problems. All self-assessments and vignettes were rated on the 5-point scale: none, mild, moderate, severe and extreme.

Self-assessment questions

breathing: In the last 30 days, how much of a problem did you have because of shortness of breath?

concentration: Overall in the last 30 days how much difficulty did you have with concentrating or remembering things?

depression: Overall in the last 30 days, how much of a problem did you have with feeling sad, low, or depressed?

mobility: Overall in the last 30 days, how much of a problem did you have with moving around?

pain: Overall in the last 30 days, how much of bodily aches or pains did you have?

sleep: In the last 30 days, how much difficulty did you have with sleeping such as falling asleep, waking up frequently during the night or waking up too early in the morning?

Vignettes

breathing

v_1: Mark has no problems with walking slowly. He gets out of breath easily when climbing uphill for 20 meters or a flight of stairs. In the last 30 days, how much of a problem did Mark have because of shortness of breath?

v_2: Paul suffers from respiratory infections about once every year. He is short of breath 3 or 4 times a week and had to be admitted in hospital twice in the past month with a bad cough that required treatment with antibiotics. In the last 30 days, how much of a problem did Paul have because of shortness of breath?

v_3: Henri has been a heavy smoker for 30 years and wakes up with a cough every morning. He gets short of breath even while resting and does not leave the house anymore. He often needs to be put on oxygen. In the last 30 days, how much of a problem did Henri have because of shortness of breath?

concentration

v_1: 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?

v_2: Sue is keen to learn new recipes but finds that she often makes mistakes and has to reread several times before she is able to do them properly. Overall in the last 30 days, how much difficulty did Sue have with concentrating and remembering things?

v_3: Eve cannot concentrate for more than 15 minutes and has difficulty paying attention to what is being said to her. Whenever she starts a task, she never manages to finish it and often forgets what she was doing. She is able to learn the names of people she meets. Overall in the last 30 days, how much difficulty did Eve have with concentrating or remembering things?

depression

v_1: Karen enjoys her work and social activities and is generally satisfied with her life. She gets depressed every 3 weeks for a day or two and loses interest in what she usually enjoys but is able to carry on with her day-to-day activities. Overall in the last 30 days, how much of a problem did Karen have with feeling sad, low, or depressed?

v_2: Maria feels nervous and anxious. She worries and thinks negatively about the future, but feels better in the company of people or when doing

something that really interests her. When she is alone she tends to feel useless and empty. Overall in the last 30 days, how much of a problem did Maria have with feeling sad, low, or depressed?

v_3: Anna feels depressed most of the time. She weeps frequently and feels hopeless about the future. She feels that she has become a burden on others and that she would be better dead. Overall in the last 30 days, how much of a problem did Anna have with feeling sad, low, or depressed?

mobility

v_1: Rob is able to walk distances of up to 200 metres without any problems but feels tired after walking one kilometre 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?

v_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?

v_3: 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?

pain

v_1: Paul has a headache once a month that is relieved after taking a pill. During the headache he can carry on with his day-to-day affairs. Overall in the last 30 days, how much of bodily aches or pains did Paul have?

v_2: Henri has pain that radiates down his right arm and wrist during his day at work. This is slightly relieved in the evenings when he is no longer

working on his computer. Overall in the last 30 days, how much of bodily aches or pains did Henri have?

v_3: Charles has pain in his knees, elbows, wrists and fingers, and the pain is present almost all the time. Although medication helps, he feels uncomfortable when moving around, holding and lifting things. Overall in the last 30 days, how much of bodily aches or pains did Charles have?

sleep

v_1: Alice falls asleep easily at night, but two nights a week she wakes up in the middle of the night and cannot go back to sleep for the rest of the night. In the last 30 days, how much difficulty did Alice have with sleeping, such as falling asleep, waking up frequently during the night or waking up too early in the morning?

v_2: Karen wakes up almost once every hour during the night. When he wakes up in the night, it takes around 15 minutes for her to go back to sleep. In the morning she does not feel well-rested. In the last 30 days, how much difficulty did Karen have with sleeping such as falling asleep, waking up frequently during the night or waking up too early in the morning?

v_3: Maria takes about two hours every night to fall asleep. She wakes up once or twice a night feeling panicked and takes more than one hour to fall asleep again. In the last 30 days, how much difficulty did Maria have with sleeping, such as falling asleep, waking up frequently during the night or waking up too early in the morning?