EIEF WORKING PAPER SERIES

EINAUDI INSTITUTE FOR ECONOMICS AND FINANCE

EIEF Working Paper 12/10
June 2010

The Heterogeneous Thresholds Ordered Response Model: Identification and Inference

by
Franco Peracchi
(University of Rome "Tor Vergata" and EIEF)
Claudio Rossetti
(LUISS)

The heterogeneous thresholds ordered response model: Identification and inference*

Franco Peracchi
Tor Vergata University and EIEF

Claudio Rossetti
LUISS

First draft: June 2010
Revised version: April 2012

Abstract

Although surveys routinely ask respondents to evaluate various aspects of their life on an ordered scale, there is concern about interpersonal comparability of these self-assessments. Statistically, the problem is one of identification in ordered response models with heterogeneous thresholds. As a solution to the identification problem, King et al. (2004) proposed using anchoring vignettes, namely brief descriptions of hypothetical people or situations that survey respondents are asked to evaluate on the same scale they use to rate their own situation. While vignettes have been introduced in several social surveys and are increasingly employed in a variety of fields, reliability of this approach hinges crucially on the validity of the assumptions of response consistency and vignette equivalence. This paper proposes a joint test of these key assumptions based on the fact that the underlying statistical model is overidentified if the two assumptions hold. Monte Carlo results show that the proposed test has good size and power properties in finite samples. We apply our test to self-assessment of pain using data from the first wave of the Survey of Health, Ageing and Retirement in Europe. We find that, when using only one of the three available vignettes, or when the test is carried out separately by subgroups of respondents, the overidentifying restrictions are less likely to be rejected.

Key words: Ordered response models; Reporting heterogeneity; Differential item functioning; Anchoring vignettes; Minimum distance methods; Self-assessment of health

JEL codes: C35, C51, I10, J14, J16

* We thank Karim Abadir, Anne Case, Valentino Dardanoni, Angus Deaton, Chris Paxson, Frank Vella, an Associated Editor, three anonymous referees, and seminar participants at ECARES, ICEEE 2011, NBER, Princeton University, the University of Alicante, and the University of Naples for helpful comments. Franco Peracchi also thanks the Center for Health and Wellbeing at Princeton University for generous hospitality during Fall 2010.

1 Introduction

Surveys respondents are often asked to evaluate various aspects of their life on an ordered scale. Examples include questions on life satisfaction and self-rated health in household surveys, or questions on customer satisfaction in consumer surveys. Although such questions are widely used, there is a concern that different people may interpret and answer them differently. This is especially true when comparing subjective assessments across groups characterized by different culture, nationality, socio-economic status, age or gender. For example, when asked to rate their own health on a given categorical scale, people may answer differently because their true or perceived health differs, but also because they interpret differently the various levels of the scale. As a consequence, differences in self-reports between otherwise similar individuals may depend on differences in response style, namely the mapping of true or perceived health into reported health (Sen 2002). Lack of interpersonal comparability of responses to subjective survey questions is often referred to as "differential item functioning" (DIF), a term originated in the educational testing literature (Holland and Wainer 1993) where a test question is said to have DIF if equally able individuals have unequal probabilities of answering the question correctly. From the view point of statistical modeling, the DIF problem is essentially one of identification in ordered response models where the observed responses are derived from latent continuous random variables discretized through a set of heterogeneous thresholds or cutoff points.

Following the seminal paper of King et al. (2004), anchoring vignettes have been developed as a new component of survey instruments that may be used to solve the DIF problem. They are brief descriptions of hypothetical people or situations that survey respondents are asked to evaluate on the same scale they use to rate their own situation. Because the people or situations described in the vignettes are the same for all respondents, vignettes have the potential to identify individual variation in subjective thresholds. A number of social surveys such as the Survey of Health, Ageing and Retirement in Europe (SHARE), the U.S. Health and Retirement Study (HRS), the English Longitudinal Study of Ageing (ELSA), and the World Health Organization's World Health Surveys (WHS) have introduced specific modules with vignette questions. However, introducing anchoring vignettes implies substantial costs in terms of survey design and reduces the time available for collecting other information. Of course, anchoring vignettes would not be necessary in a survey if one is willing to apply the response-scale correction from a different survey under the assumption that the DIF problem is the same.

Vignette questions have been applied to a variety of problems including comparison of health

(Salomon, Tandon and Murray 2004, King and Wand 2007, Bago D'Uva, O'Donnel and van Doorslaer 2008, Bago D'Uva et al. 2008, Peracchi and Rossetti 2009), health system responsiveness (Rice et al. 2012), political efficacy (King et al. 2004), work disability (Kapteyn, Smith and van Soest 2007), life satisfaction (Angelini et al. 2008), and job satisfaction (Kristensen and Johansson 2008). In most cases, evidence of reporting heterogeneity is found and corrections on the comparisons of interest are made using the vignettes.

Although vignettes are increasingly employed by researchers in various fields, reliability of this approach hinges crucially on the validity of two key assumptions (King et al. 2004). The first assumption ("response consistency") is that individuals use the available response categories in the same way when assessing their own situation and the hypothetical situations in the vignettes. The second assumption ("vignette equivalence") is that the hypothetical situation in a vignette is perceived by all respondents in the same way and on the same uni-dimensional scale, apart from random error. As pointed out by Deaton (2010), the vignette approach replaces the assumption that there are no differences in the way people rank themselves on a subjective scale with the alternative assumption (response consistency) that there are no differences in their capacity for empathy with other people's conditions. In addition, vignette equivalence assumes that there are no systematic differences in the way people perceive the situations represented in each vignette. The latter is also a very strong assumption, for example because of problems with translation of the same vignette in different languages. Hence, testing these two key assumptions turns out to be a critical step in evaluating the validity of the vignette approach.

One approach to testing for response consistency relies on the availability of some objective measure of the concept of interest. This approach, which rests on the maintained assumption of vignette equivalence, is used by King et al. (2004) and van Soest et al. (2011) to provide evidence supporting the assumption of response consistency. Other evidence, however, is less supportive (Datta Gupta, Kristensen and Pozzoli 2010, Bago D'Uva et al. 2011, Voňková and Hullegie 2011). The main problem with this approach is that objective measures of the concept of interest are typically only available in *ad-hoc* studies. Recently, Kaptyen et al. (2011) propose a different test of response consistency based on longitudinal data where respondents are shown vignettes that are descriptions of their own health collected in a previous interview. They find that response consistency is satisfied only for one of the five health domains considered, namely sleep.

Far less attention has been paid to vignette equivalence. King et al. (2004) suggest an informal test based on the ordering of the answers to different vignette questions on the same domain. A

more formal approach is adopted by Bago D'Uva et al. (2011), who test the necessary condition of no systematic variation across individuals by allowing vignette evaluations to depend on observed personal characteristics. This test does not require objective measures but maintains the assumption of response consistency and needs at least two vignettes questions for each concept of interest.

In this paper we propose a simple joint test of the two key assumptions of response consistency and vignette equivalence. The proposed test exploits the fact that, as pointed out by Deaton (2010), the statistical model is overidentified under these two assumptions. Our test offers several advantages. First, it does not require the availability of some objective measures and can be carried out using any dataset containing at least one vignette question for each concept of interest. Second, it does not require embedding the restricted model that imposes response consistency and vignette equivalence into a larger encompassing model. Third, it only requires a consistent and asymptotically normal estimator of the estimable parameters in the model. This is an advantage, both computationally and because the test can easily be extended to models with sample selection and to semiparametric settings where strong distributional assumptions are relaxed. Fourth, because it exploits the mapping between the estimable parameters and the full set of model parameters, imposing additional restrictions on the model is particularly transparent and simple. Of course, as typical with tests of parametric or semi-parametric models, our test is conditional on some other assumptions. Thus, it may reject the overidentifying restrictions for other reasons than failure of response consistency and vignette equivalence, for example because of failure of parametric restrictions or because relevant variables have been omitted from the model.

We investigate the finite sample performance of the proposed test through a Monte Carlo study. We find that the test has good size and power properties in finite samples. Specifically, the test has no size distortion and no overrejection is reported when the number of overidentifying restrictions increases.

Finally, we apply our test to self-assessment of pain using data from Release 2 of the first (2004) wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). Release 2 of the data also includes the answers to vignettes questions in a self-administered questionnaire submitted to a randomly selected subsample of respondents. We find that the overidentifying restrictions are less likely to be rejected when using only one of the three available vignettes, or when the test is carried out separately by subgroups of respondents.

The remainder of this paper is organized as follows. Section 2 presents the heterogeneous thresh-

olds ordered response model, discusses its identification, and proposed a test of the overidentifying restrictions implied by the assumptions of response consistency and vignette equivalence. Section 3 presents the results of a Monte Carlo study to assess the finite sample performance of the proposed test. Section 4 illustrates the use of our test through an empirical application to self-assessment on various health domains. Finally, Section 5 offers some conclusions.

2 The heterogeneous thresholds ordered response model

Let Y_0 denote the answer by a randomly chosen individual on some concept of interest, and let $Y_1, \ldots, Y_J$ denote the answers given by the same individual to J vignette questions on the given concept. For concreteness, we think of Y_0 as the assessment of own health on some domain and of Y_j, $j = 1, \ldots, J$, as the assessment of health on the same domain in the jth vignette. We assume that the elements of the $(J+1)$-vector of observed responses $Y = (Y_0, Y_1, \ldots, Y_J)$ are all categorical and take values $r = 0, 1, \ldots, R$.

Each observed categorical response Y_j is assumed to depend on an underlying continuous latent variable Y_j^* through the observation rule

$$Y_j = \sum_{r=0}^{R} r \, 1\{\xi_{j,r-1} < Y_j^* \leq \xi_{jr}\}, \qquad j = 0, 1, \ldots, J,$$

where $1\{\cdot\}$ is the indicator function, and the ξ_{jr} are $R+2$ individual-specific thresholds or cutoff points satisfying $\xi_{j,r-1} < \xi_{jr}$, with $\xi_{j,-1} = -\infty$ and $\xi_{jR} = \infty$. We refer to Greene and Hensher (2010) for a history and an extensive review of this type of models.

The statistical problem is how to use the sample information in order to learn about the conditional distribution of Y_0^* given a vector of observable regressors. The vignette information is not of direct interest, but is used instrumentally in order to control for the fact that the cutoffs ξ_{jr} may vary across individuals depending on observable regressors and, possibly, unobservable individual effects.

2.1 Model specification

We assume that the continuous latent variables Y_j^* obey linear models of the form

$$Y_j^* = \alpha_j + \beta_j^\top X_j + \sigma_j U_j, \qquad j = 0, 1, \ldots, J, \tag{1}$$

where X_j is a vector of observable exogenous regressors, possibly specific to the jth latent variable, α_j, β_j and σ_j are unknown parameters, and U_j is an unobservable random error distributed

independently of X_j with mean zero and distribution function F. We could easily generalize this model by representing Y_j^* as additively separable in X_j and U_j, as in Cunha, Heckman and Navarro (2007), that is, by assuming that $Y_j^* = \varphi_j(X_j) + \sigma_j U_j$, where φ_j is an unknown function.

To account for observed heterogeneity in response scales, we let the thresholds depend on a vector W_j of observable exogenous regressors, possibly specific to the jth latent variable, that is

$$\xi_{jr} = \begin{cases} -\infty, & \text{if } r = -1, \\ \kappa_{jr}(W_j), & \text{if } r = 0, 1, \ldots, R-1, \\ \infty, & \text{if } r = R, \end{cases}$$

for $j = 0, 1, \ldots, J$, where the κ_{jr} are unknown functions. To guarantee monotonicity of the thresholds, that is $\xi_{j,r-1} < \xi_{jr}$ for all r, the functions κ_{jr} must be monotonically increasing. Unobserved heterogeneity may easily be accommodated by including in W_j an unobserved individual effect, as in Rossi, Gilula and Allenby (2001). This offers a simple way of allowing for correlation between self-assessment and vignette responses conditional on the observed regressors.

A parametric specification of the κ_{jr} functions is the so-called compound hierarchical ordered response model of King et al. (2004), where

$$\kappa_{jr}(W_j) = \begin{cases} \gamma_{j0} + \delta_{j0}^\top W_j, & \text{if } r = 0, \\ \kappa_{j,r-1} + \exp(\gamma_{jr} + \delta_{jr}^\top W_j), & \text{if } r = 1, \ldots, R-1. \end{cases} \tag{2}$$

This specification guarantees monotonicity of the thresholds, that is, $\xi_{j0} < \cdots < \xi_{j,R-1}$. In addition, the nonlinearities in (2) provide weak (through functional form) identification of the model when W_j includes the same variables as X_j. An alternative parametric specification, originally proposed by Terza (1985), is

$$\kappa_{jr}(W_j) = \gamma_{jr} + \delta_{jr}^\top W_j, \qquad r = 0, 1, \ldots, R-1. \tag{3}$$

This specification does not guarantee monotonicity of the thresholds, but is computationally simpler than (2) and has the advantage of making the identification issues more transparent.

To avoid identification via functional form restrictions, we adopt the linear model (3) for the cutoffs and consider the extreme but very relevant case of no exclusion restrictions, where $X_j = W_j = X$ for all j, with X containing k exogenous regressors. Pudney and Shields (2000) also specify the thresholds as linear functions of observed regressors but achieve identification through exclusion restrictions, by excluding some of the variables in the threshold equations from those in the latent linear model (1). Since model (3) puts no constraints on the threshold parameters, we cannot ensure monotonicity of the thresholds. As a result, although the probabilities sum to one by construction, there is no guarantee that they are positive.

Under this model specification, the likelihood contribution of the self-assessment component is

$$\mathcal{L}_1(\theta_1; X, Y_0) \propto \prod_{r=0}^{R} \left[F\left(\frac{\xi_{0r} - \alpha_0 - \beta_0^\top X}{\sigma_0} \right) - F\left(\frac{\xi_{0,r-1} - \alpha_0 - \beta_0^\top X}{\sigma_0} \right) \right]^{Y_{0r}},$$

where $Y_{0r} = 1\{Y_0 = r\}$ and the vector θ_1 consists of the parameters in α_0, β_0, σ_0, $\gamma_0 = (\gamma_{00}, \ldots, \gamma_{0,R-1})$ and $\delta_0 = (\delta_{00}, \ldots, \delta_{0,R-1})$. The total number of parameters in θ_1 is equal to $(k+1)(R+1)+1$. The likelihood contribution of the vignette component is

$$\mathcal{L}_2(\theta_2; X, Y_1, \ldots, Y_J) \propto \prod_{j=1}^{J} \prod_{r=0}^{R} \left[F\left(\frac{\xi_{jr} - \alpha_j - \beta_j^\top X}{\sigma_j} \right) - F\left(\frac{\xi_{j,r-1} - \alpha_j - \beta_j^\top X}{\sigma_j} \right) \right]^{Y_{jr}},$$

where $Y_{jr} = 1\{Y_j = r\}$ and the vector θ_2 consists of the parameters in all the α_j, β_j, σ_j, $\gamma_j = (\gamma_{j0}, \ldots, \gamma_{j,R-1})$ and $\delta_j = (\delta_{j0}, \ldots, \delta_{j,R-1})$. The total number of parameters in θ_2 is equal to $J[(k+1)(R+1)+1]$. The full likelihood for a single observation is

$$\begin{aligned} \mathcal{L}(\theta; X, Y) &= \mathcal{L}_1(\theta_1; X, Y_0)\, \mathcal{L}_2(\theta_2; X, Y_1, \ldots, Y_J) \\ &\propto \prod_{j=0}^{J} \prod_{r=0}^{R} \left[F\left(\frac{\xi_{jr} - \alpha_j - \beta_j^\top X}{\sigma_j} \right) - F\left(\frac{\xi_{j,r-1} - \alpha_j - \beta_j^\top X}{\sigma_j} \right) \right]^{Y_{jr}}, \end{aligned} \tag{4}$$

where $\theta = (\theta_1, \theta_2) = \{(\alpha_j, \beta_j, \sigma_j, \gamma_j, \delta_j), j = 0, \ldots, J\}$ and we write (θ_1, θ_2) as a shorthand for $(\theta_1^\top, \theta_2^\top)^\top$. The total number of parameters in θ is equal to $[(k+1)(R+1)+1](J+1)$.

2.2 Identification

Identification of the model parameters requires location and scale restrictions, plus restrictions linking the self-assessment and the vignette contributions to the likelihood. After substituting the model for the cutoffs (3) into (4), the full likelihood for a single observation becomes

$$\mathcal{L}(\theta; X, Y) \propto \prod_{j=0}^{J} \prod_{r=0}^{R} \left[F\left(\frac{(\gamma_{jr} - \alpha_j) + (\delta_{jr} - \beta_j)^\top X}{\sigma_j} \right) - F\left(\frac{(\gamma_{j,r-1} - \alpha_j) + (\delta_{j,r-1} - \beta_j)^\top X}{\sigma_j} \right) \right]^{Y_{jr}}.$$

In the absence of prior restrictions, the parameters in θ are clearly not separately identifiable. The identifiable parameters are the following functions of the parameters in θ

$$\gamma_{jr}^* = \frac{\gamma_{jr} - \alpha_j}{\sigma_j}, \qquad \delta_{jr}^* = \frac{\delta_{jr} - \beta_j}{\sigma_j},$$

with $r = 0, 1, \ldots, R-1$ and $j = 0, 1, \ldots, J$. We shall refer to these parameters as the reduced-form parameters. The reduced form of the model corresponds to a set of $J+1$ ordered response models with outcome specific parameters, a model first proposed by Pudney and Shields (2000) and

referred to as the generalized ordered response model. Because the total number of parameters in the reduced form is equal to $R(k+1)(J+1)$, the number of restrictions needed to exactly identify the parameters in θ from the identifiable reduced-form parameters is equal to

$$[(k+1)(R+1)+1](J+1) - R(k+1)(J+1) = (k+2)(J+1).$$

Standard location and scale restrictions, namely the $2(J+1)$ restrictions $\gamma_{j0} = 0$ and $\sigma_j = 1$, $j = 0, \ldots, J$, are not enough to identify the parameters in θ, so $k(J+1)$ additional restrictions are needed.

In the absence of vignette information ($J = 0$), the $(k+1)(R+1)+1$ parameters of the model for the self-assessment cannot be obtained from the $R(k+1)$ identifiable parameters of the reduced form because we only have 2 normalization restrictions ($\gamma_{00} = 0$ and $\sigma_0 = 1$). In this case, k additional restrictions would be needed to exactly identify the model. This means that we cannot separately identify the coefficients β_0 on the exogenous regressors in the latent regression for Y_0^* model from the coefficients δ_{0r} in the thresholds.

One way of achieving exact identification of the model is to exclude exogenous regressors from one threshold (Terza 1985). This gives the k additional restrictions needed. In this case, however, only deviations from the cutoff from which the regressors are arbitrarily excluded can be identified. Alternatively, a standard practice in ordered response models is to assume homogeneous thresholds, that is $\delta_{0r} = 0$, $r = 0, 1, \ldots, R-1$, which corresponds to a set of Rk restrictions. Because only k restrictions would be needed to identify the model, when there are more than two response categories ($R > 1$) we have $(R-1)k$ overidentifying restrictions that allow us to test the assumption of homogeneous thresholds. This test corresponds to the Wald test proposed by Brant (1990) for testing the proportional odds restriction in the ordered logistic regression.

If vignette information is available ($J > 0$), King et al. (2004) proposed to identify the model by linking the self-assessment and the vignettes through the following assumptions:

A.1 (Response consistency): $\gamma_{jr} - \gamma_{0r} = \delta_{jr} - \delta_{0r} = 0$, $r = 0, 1, \ldots, R-1$, $j = 1, \ldots, J$.

A.2 (Vignette equivalence): $\beta_j = 0$, $j = 1, \ldots, J$.

The first assumption is that each individual uses the response categories for a particular survey question in the same way when providing self-assessment and when assessing each of the hypothetical situations in the vignettes. The second assumption is that the level of the variable represented in each vignette is perceived by all respondents in the same way and on the same uni-dimensional

scale, apart from random measurement error. Imposing A.1 and A.2 provides $[R(k+1)+k]J$ restrictions. Because in this case self-assessment and vignettes are linked together, location and scale can be fixed by setting the constant term of the first (common) threshold $\gamma_{00} = 0$ and the variance of the self-assessment $\sigma_0 = 1$. Alternatively, location and scale can be fixed by setting the constant terms of the extreme vignettes $\alpha_1 = 0$ and $\alpha_J = 1$ (King, Lau and Wand 2009). Imposing assumptions A.1 and A.2, together with location and scale normalization of the self-assessment ($\gamma_{00} = 0$ and $\sigma_0 = 1$), gives a total of $[R(k+1)+k]J + 2$ restrictions.

To illustrate, in the special case of three response categories ($R = 2$) and one exogenous regressor ($k = 1$), the model contains $7(J+1)$ parameters, namely $\{(\alpha_j, \beta_j, \gamma_{j0}, \delta_{j0}, \gamma_{j1}, \delta_{j1}, \sigma_j), j = 0, 1, \ldots, J\}$. The reduced-form parameters are only $4(J+1)$, namely

$$\gamma^*_{j0} = \frac{\gamma_{j0} - \alpha_j}{\sigma_j}, \qquad \delta^*_{j0} = \frac{\delta_{j0} - \beta_j}{\sigma_j}, \qquad \gamma^*_{j1} = \frac{\gamma_{j1} - \alpha_j}{\sigma_j}, \qquad \delta^*_{j1} = \frac{\delta_{j1} - \beta_j}{\sigma_j},$$

with $j = 0, 1, \ldots, J$. In this case, $3(J+1)$ restrictions are needed to exactly identify the model.

Without vignettes ($J = 0$), the 7 parameters in the model (α_0, β_0, σ_0, γ_{00}, δ_{00}, γ_{01}, δ_{01}) cannot be obtained from the 4 reduced-form parameters (γ^*_{00}, δ^*_{00}, γ^*_{01}, δ^*_{01}) because we only have 2 normalization restrictions ($\gamma_{00} = 0$ and $\sigma_0 = 1$). The model is exactly identified under the additional assumption that $\delta_{00} = 0$. Nonetheless, in this case only deviations from δ_{00} can be identified. Another possibility to achieve identification is to assume homogeneous thresholds ($\delta_{00} = 0$ and $\delta_{01} = 0$). In this case, there is one overidentifying restriction that would allow testing the homogeneous thresholds hypothesis.

With vignettes ($J > 0$), the assumption of response consistency gives $4J$ restrictions

$$\gamma_{j0} - \gamma_{00} = \gamma_{j1} - \gamma_{01} = \delta_{j0} - \delta_{00} = \delta_{j1} - \delta_{01} = 0, \qquad j = 1, \ldots, J,$$

while the assumption of vignette equivalence gives J restrictions

$$\beta_j = 0, \qquad j = 1, \ldots, J.$$

Because these two sets of restrictions, together with location and scale normalization ($\gamma_{00} = 0$ and $\sigma_0 = 1$), provide a total of $5J + 2$ restrictions, we have a total of $2J - 1$ overidentifying restrictions.

For example, with only one vignette ($J = 1$) we have 14 model parameters (α_0, β_0, σ_0, γ_{00}, δ_{00}, γ_{01}, δ_{01}, α_1, β_1, σ_1, γ_{10}, δ_{10}, γ_{11}, δ_{11}) and 8 reduced-form parameters (γ^*_{00}, δ^*_{00}, γ^*_{01}, δ^*_{01}, γ^*_{10}, δ^*_{10}, γ^*_{11}, δ^*_{11}). Under the 2 normalization restrictions ($\gamma_{00} = 0$ and $\sigma_0 = 1$) and the 5 restrictions implied by A.1 and A.2 ($\gamma_{10} = \gamma_{00}$, $\gamma_{11} = \gamma_{01}$, $\delta_{10} = \delta_{00}$, $\delta_{11} = \delta_{01}$ and $\beta_1 = 0$) the model is

overidentified (it has 1 overidentifying restriction). With two vignettes ($J = 2$) we have 21 model parameters and 12 reduced-form parameters. In this case, with 2 normalization restrictions and 10 restrictions implied by A.1 and A.2, we have 3 overidentifying restrictions. Finally, with three vignettes ($J = 3$) we have 28 model parameters and 16 reduced-form parameters. In this case, with 2 normalization restrictions and 15 restrictions implied by A.1 and A.2, we have 5 overidentifying restrictions.

2.3 Inference

With vignettes ($J \geq 1$) and more than two response categories ($R \geq 2$), overidentification of the restricted model that imposes A.1, A.2 and the location and scale normalizations provides the basis for testing the key assumptions A.1 and A.2.

One way of approaching the problem of testing is to use a minimum distance (MD) approach. Let θ be the vector of $s = [(k+1)(R+1)+1](J+1)$ model parameters and let π be the vector of $q = R(k+1)(J+1)$ reduced-form parameters. Also let ψ be the subvector of θ containing the "free" parameters, namely those not subject to the restrictions implied by A.1, A.2 and the location and scale normalizations. Since the number of these restrictions is equal to $[R(k+1)+k]J+2$, the number of "free" parameters in ψ is equal to $p = k+R(k+1)+2J$, so the number of overidentifying restrictions is equal to

$$q - p = R(k+1)(J+1) - [k + R(k+1) + 2J] = k(JR-1) + J(R-2).$$

When there are more than two response categories ($R \geq 2$) and at least one vignette ($J \geq 1$), we have that $q - p \geq 1$ (assuming that $k \geq 1$). In the binary response case ($R = 1$), we still have overidentifying restrictions if either $J = 2$ and $k \geq 3$, or $J \geq 3$ and $k \geq 2$.

Let π_0 and ψ_0 be the values of π and ψ in the population. Because ψ_0 includes the scale parameters σ_j, for $j = 1, \ldots, J$, the relationship between π_0 and ψ_0 is nonlinear. We write this relationship as

$$\pi_0 = g(\psi_0),$$

where $g\colon \Re^p \to \Re^q$ is a differentiable function with Jacobian matrix G. For (local) identifiability, we need $G(\psi)$ to be of full rank in an open neighborhood of ψ_0. Appendix A presents the structure of g and G.

Given a sample of size n from the joint distribution of (X, Y), let $\hat{\pi}_n$ denote the estimator of π_0 obtained by fitting $J+1$ generalized ordered response models, one for each categorical variable

in Y. This estimator is very easy to compute, and is $\sqrt{n}$-consistent and asymptotically normal under general conditions. Given $\hat{\pi}_n$, the MD method suggests estimating the vector ψ_0 of "free" parameters by picking the element in the parameter space Ψ such that the difference $\hat{\pi}_n - g(\psi)$ is the smallest possible. The resulting estimator of ψ_0 is consistent and asymptotically normal under general conditions (Ferguson 1996).

An asymptotically optimal MD estimator of ψ_0 is the solution $\hat{\psi}_n$ to the problem

$$\min_{\psi \in \Psi} Q_n(\psi) = [\hat{\pi}_n - g(\psi)]^\top \hat{V}_n^{-1} [\hat{\pi}_n - g(\psi)], \tag{5}$$

where the $q \times q$ matrix $\hat{V}_n$ is a positive definite estimate of the asymptotic variance of $\hat{\pi}_n$. Under general conditions,

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \Rightarrow \mathcal{N}(0,\, (G_0 V_0^{-1} G_0^\top)^{-1})$$

as $n \to \infty$, where $G_0 = G(\psi_0)$ denotes the $p \times q$ Jacobian matrix of g evaluated at ψ_0 and V_0 denotes the asymptotic variance of $\hat{\pi}_n$.

Computation of $\hat{\psi}_n$ is straightforward using an iterative procedure. Starting from an initial estimate $\hat{\psi}^{(0)}$, the updated estimate at the $(h+1)$th iteration is given by

$$\hat{\psi}^{(h+1)} = (\hat{G}_h \hat{V}_n^{-1} \hat{G}_h^\top)^{-1} \hat{G}_h \hat{V}_n^{-1} (\hat{\pi}_n - \hat{g}_h + \hat{G}_h^\top \hat{\psi}^{(h)}), \qquad h = 0, 1, \ldots,$$

where $\hat{G}_h = G(\hat{\psi}^{(h)})$ and $\hat{g}_h = g(\hat{\psi}^{(h)})$. This corresponds to a GLS regression of the transformed reduced form estimates $\hat{\pi}_n - \hat{g}_h + \hat{G}_h^\top \hat{\psi}^{(h)}$ on the columns of $\hat{G}_h$ with weighting matrix $\hat{V}_n^{-1}$.

When $J \geq 1$, the model that imposes A.1 and A.2 is overidentified so, under the null hypothesis that both assumptions hold,

$$nQ_n(\hat{\psi}) \Rightarrow \chi^2_{q-p}$$

as $n \to \infty$, where $q - p = k(JR - 1) + J(R - 2)$ is the number of overidentifying restrictions. This result provides the basis for asymptotic tests that reject the key assumptions A.1 and A.2 for large values of the statistic $nQ_n(\hat{\psi}_n)$.

A test of this type offers several advantages. First, it can be performed with any dataset containing vignette questions (one vignette is enough) on a given concept of interest and does not require additional information like objective measures. Second, it does not require embedding the restricted model that imposes response consistency and vignette equivalence into a larger encompassing model. Third, it only requires a consistent and asymptotically normal estimator of the reduced-form parameters. This is an advantage, both computationally and because the test can

easily be extended to models with sample selection and to semiparametric settings where strong distributional assumptions are relaxed. Fourth, because we exploit the mapping g between the "free" parameters and the reduced form parameters, imposing additional restrictions is particularly simple and transparent. A potential disadvantage of our test is that it may reject the overidentifying restrictions for other reasons than failure of response consistency and vignette equivalence, for example because of failure of linear index restrictions or because relevant variables have been omitted from the model.

2.4 Power of the test

There are a few special cases in which the proposed test lacks power. The first case is when

$$\gamma_{jr} - \gamma_{0r} = 0$$

and

$$\delta_{jr} - \delta_{0r} - \beta_j = \delta_{ls} - \delta_{0s} - \beta_s,$$

for all vignettes j, l and all thresholds r, s. This is the unlikely case when (i) there is no violation of A.1 due to differences in the intercepts, and (ii) the violations of A.1 and A.2 due to the differences in the slopes are exactly the same for all thresholds and all vignettes, so they all cancel out.

For example, with three response categories ($R = 2$), one exogenous regressor ($k = 1$) and one vignette ($J = 1$), the vector of model parameters is $\theta = (\alpha_0,\ \beta_0,\ \sigma_0,\ \gamma_{00},\ \delta_{00},\ \gamma_{01},\ \delta_{01}\ \alpha_1,\ \beta_1,\ \sigma_1,\ \gamma_{10},\ \delta_{10},\ \gamma_{11},\ \delta_{11})$ while the vector of reduced-form parameters is $\pi = (\gamma^*_{00},\ \delta^*_{00},\ \gamma^*_{01},\ \delta^*_{01},\ \gamma^*_{10},\ \delta^*_{10},\ \gamma^*_{11},\ \delta^*_{11})$. In this case, if

$$\delta_{10} - \delta_{00} - \beta_1 = \delta_{11} - \delta_{01} - \beta_1 = \Delta \neq 0,$$

then the vector $\tilde{\psi} = (\alpha_0,\ \tilde{\beta}_0,\ \tilde{\delta}_{00},\ \gamma_{01},\ \tilde{\delta}_{01},\ \alpha_1,\ \sigma_1)$, with $\tilde{\beta}_0 = \beta_0 + \Delta$, $\tilde{\delta}_{00} = \delta_{00} + \Delta$ and $\tilde{\delta}_{01} = \delta_{01} + \Delta$, also solves the minimization problem (5) and satisfies the restrictions A.1 and A.2.

The second case is when

$$\gamma_{jr} - \gamma_{0r} = \gamma_{js} - \gamma_{0s} \neq 0,$$

for any vignette j and all thresholds r, s. This is the unlikely case when the violations of A.1 due to differences in the intercepts are exactly the same for all thresholds, so they all cancel out. Notice that in this case the violation of response consistency only affects the intercepts α_j in the vignette equations but does not affect the parameters of interest α_0 and β_0.

Consider again the example with three response categories ($R = 2$), one exogenous regressor ($k = 1$) and one vignette ($J = 1$). In this case, if

$$\gamma_{10} - \gamma_{00} = \gamma_{11} - \gamma_{01} = \Delta \neq 0,$$

then the vector $\tilde{\psi} = (\alpha_0,\ \beta_0,\ \delta_{00},\ \gamma_{01},\ \delta_{01},\ \tilde{\alpha}_1,\ \sigma_1)$, where $\tilde{\alpha}_1 = \alpha_1 - \Delta$, is also a solution to the minimization problem (5) and satisfies the restrictions A.1 and A.2. Note that in this case the parameters of interest α_0 and β_0 are not affected.

3 Monte Carlo results

In this section, we investigate the finite sample performance of our test of the overidentifying restrictions implied by A.1 (response consistency) and A.2 (vignette equivalence) through a Monte Carlo study. Our setup is as follows.

1. We set the number of thresholds or cutoffs to $R = 2$.
2. We set the number of exogenous regressors to $k = 1, 2$.
3. We set the number of vignettes to $J = 1, 2$.
4. We set the sample size to $n = 250$, 500 and 1000.
5. For all j, we draw the errors U_j from a standard normal distribution.
6. The first regressor X_1 is drawn from a $U(0, 1)$ distribution, while the second regressor X_2 is a 0-1 indicator equal to one with probability .50. Considering the case of a binary regressor is useful because researchers are often interested in comparing subjective assessments across groups.
7. The null hypothesis H_0 corresponds to the case when A.1 and A.2 both hold. As for the alternatives, we consider three cases: i) A.1 holds but A.2 fails (hypothesis H_1), ii) A.2 holds but A.1 fails (hypothesis H_2), and iii) both A.1 and A.2 fail (hypothesis H_3).
8. We choose the model parameters to have an approximately even distribution of reports in each category under the null hypothesis H_0.
9. Each Monte Carlo experiment consists of 1,000 runs using antithetic pseudo-random numbers.

The reduced-form parameters are estimated by maximizing the log-likelihood of $J+1$ generalized ordered probit models using the Newton-Raphson method with analytical first and second derivatives. The routines that compute the estimates of the reduced-form and the "free" parameters are all written in Mata, the matrix programming language of the statistical package Stata (version 11).

Table 1 shows the Monte Carlo rejection frequencies for tests of asymptotic 5% level. The row labeled H_0 reports the observed size of our test, which should be compared with its asymptotic value of 5%. Already for $n = 250$, rejection frequencies are close to nominal under the null. Thus, our test shows no evidence of size distortion in finite samples. On the other hand, the size of the test remains stable when the number of overidentifying restrictions increases from 1 to 6.

The block labeled H_1 reports the power of our test when response consistency holds but vignette equivalence fails. The rejection frequencies are presented for increasing values of β_1, which is the coefficient on the first regressor in the linear index for the first vignette. As discussed in Section 2.4, our test has essentially no power in the case of only one vignette, but its power increases with β_1 in the case of two vignettes. The block labeled H_2 reports the power of our test when vignette equivalence holds but response consistency fails. The first four rows present rejection frequencies for increasing values of the difference $\delta_{11} - \delta_{01}$, while the last four rows present rejection frequencies for increasing values of the differences $\delta_{11} - \delta_{01}$ and $\gamma_{11} - \gamma_{01}$. In this case, the power curves are always increasing except when $J = k = 1$ and the shift is only in the slope ($\delta_{11} - \delta_{01}$ is different from zero). Although the Monte Carlo was explicitly designed to detect differential power properties of our test when only one alternative fails, our results suggest that the test is in fact rather "symmetric". Finally, the block labeled H_3 reports the power of our test when both assumptions fail. In this case, the results are qualitatively similar to the case of H_2 but now the power of our test is higher in all experiments.

With $n = 500$ and $n = 1000$, the results are qualitatively similar to the case of $n = 250$, but the power increases with the sample size in most experiments.

4 Empirical application

Women tend to report worse health than men at all ages, although they are less likely to die than men and are less likely to be hospitalized than men at ages when pregnancy-related hospitalization is no longer an issue. As argued by Case and Deaton (2005), "this pattern ... by gender is close to universal around the world." This paradox could have various explanations, not necessarily mutually exclusive. One is that gender differences in self-assessment of health reflect systematic

differences in the prevalence of chronic conditions, for either biological or behavioral reasons. For example, Case and Paxson (2005) show that, in the U.S., gender differences in self-rated general health are almost entirely due to the differences in the distribution of reported chronic conditions, with hardly any role for gender differences in the mapping from chronic conditions to reported poor health. Another explanation is that gender differences in self-assessment of health reflect systematic differences in the way respondents locate themselves on subjective scales (Lindeboom and van Doorslaer 2004). Anchoring vignettes offer one way of controlling for such differences.

4.1 Data

Our data are from Release 2 of the first (2004–05) wave of the Survey of Health, Ageing and Retirement in Europe (SHARE), a multidisciplinary and cross-national bi-annual household panel survey, that is nationally representative of the population aged 50+ living in private households in Europe. The first wave covers about 19,500 households and about 28,500 individuals in 11 European countries (Austria, Belgium, Denmark, France, Germany, Greece, the Netherlands, Spain, Sweden and Switzerland). For a detailed description, see Börsch-Supan and Jürges (2005).

SHARE collects detailed information on demographic and economic variables, health, psychological variables, and social support variables. In particular, respondents are asked to use a 5-point ordered scale to rate their own health in general and to assess their health on six domains, namely pain, sleeping problems, mobility problems, concentration problems, shortness of breath and depression. In eight countries (Belgium, France, Germany, Greece, Italy, the Netherlands, Spain and Sweden), a random subsample of the respondents is also asked to answers to vignette questions on the six health domains. The vignette questions are presented in a random order after the self-assessment questions. For each domain, respondents are presented three hypothetical situations, corresponding to people with low, moderate and serious health problems. They are instructed to evaluate the hypothetical persons on exactly the same 5-point ordered scale used for the self-assessments and to assume that the hypothetical persons in the vignettes have their same age and background.

4.2 Descriptive statistics

We restrict attention to men and women aged 50–80 for whom the vignette information is available and there are no missing data on any of the variables that we use. Because the fraction with missing data is small (less than 3% for self-assessment questions and less than 5% for vignette questions), we work with the subsample with complete data and ignore selection issues. This gives

a sample of 3,458 observations (1,631 men and 1,827 women), that represents about 16% of the full SHARE sample in the relevant age group. Table 2 compares the composition of our working sample with that of the full SHARE sample and the vignette sample for the age group 50–80. Country differences in the importance of the vignette sample are mainly due to differences in sampling design and funding availability. We account for such differences by using the survey weights specifically provided for the vignette sample.

Table 3 shows the correlation between self-rated general health and self-assessments on the six health domains. Self-reported problems on these domains are all positively correlated with general health and with each other. The correlation with general health is highest for pain and mobility problems (.44), while the correlation between domains is highest for pain and mobility problems (.53). The last column of Table 3 shows the estimated coefficients from an ordered probit model for general health on self-assessments of health in the six domains. Because the estimated coefficient is highest for pain (.513), we focus on this particular health domain.[1] Appendix B reports the various vignettes for pain, where the labels "Vignette 1", "Vignette 2" and "Vignette 3" do not represent the order in which the three vignettes are presented (which is random) but instead refer to the severity of the hypothetical situation (low, moderate and serious).

Figure 1 shows the histograms of the self assessment of pain and the answers to the vignette questions by gender. Women are more likely to report severe or extreme pain than men. The distribution of the answers to the vignette questions confirms that on average respondents tend to rank the three vignettes from least to most severe pain.

4.3 Results

We estimate a fully parametric version of our model by assuming normality of the latent errors in (1). The reduced form of our model corresponds to a set of $J+1 = 4$ ordered probit models with outcome specific parameters. To avoid increasing too much the number of overidentifying restrictions, we merge the response category "Mild" with "Moderate", and "Severe" with "Extreme". This gives $R = 2$ thresholds.

Because no credible exclusion restriction is available, we allow W_j to contain exactly the same regressors as X_j for all j. In our baseline specification, the regressors include a female indicator, age, an indicator for college education completed, the logarithm of per-capita household income, an indicator for reporting at least one diagnosed chronic condition, and hand grip strength. The latter

[1] Results for the other five health domains are available from the Authors upon request.

is consider an objective measure of health and is known to be a good predictor of future medical problems (Rantanen et al. 1999). It is measured here as the maximum of up to four measurements taken by the interviewer, two for each hand. The baseline specification includes $k = 6$ regressors, so the number of reduced-form parameters is equal to $q = R(k+1)(J+1) = 56$, the number of "free" parameters is equal to $p = k + R(k+1) + 2J = 26$, and the number of overidentifying restrictions is equal to $q - p = k(JR-1) + J(R-2) = 30$. Appendix C presents the MD estimates of the model parameters under the assumptions of response consistency and vignette equivalence. Note that predicted probabilities are always positive in this specification of the model.

Table 4 presents the results of the χ^2 test of the overidentifying restrictions implied by our two key assumptions. Using the full sample, three vignettes ($J = 3$ and three response categories ($R = 2$), the overidentifying restrictions are rejected at any conventional significance level. The remainder of the table shows the results obtained when the test is carried out using different subsets of the vignettes ($J = 1$ or 2) or using all five original response categories ($R = 4$). The overidentifying restrictions are not rejected at the 5% level when using only one vignette, especially when using the first or the second. They are also not rejected when using the second and the third vignette together, but not when using the first and either the second or the third. Our results are consistent with those of Voňková and Hullegie (2011), who also find that the vignette method is sensitive to the choice of the vignette. When the test is carried out using all five original response categories (last row of the table), the overidentifying restrictions are again strongly rejected.

Table 5 shows the results obtained when the test is carried out separately for various subgroups of respondents. Specifically, we group respondents by gender, age group (50–64 vs. 65–80), health status (no self-reported chronic condition vs. some conditions), educational attainments (less than secondary vs. secondary or post-secondary), and region of residence (Mediterranean vs. non-Mediterranean country). Now the overidentifying restrictions are rejected for women, people aged 50–64, people reporting no chronic condition, people with less than secondary education, and for residents in non-Mediterranean countries, but are not rejected for men, people aged 65–80, people reporting some chronic conditions, more educated people, and for residents in Mediterranean countries. The fact that, when splitting the sample in two subgroups of similar size, the results of the test may be quite different suggests three things. First, failure to reject is not simply due to a smaller sample size. Second, since response consistency is a within-respondent property while vignette equivalence is a between-respondent property, our evidence suggests that the assumption of vignette equivalence is perhaps more problematic. Third, some of our subgroups may still be

too heterogeneous for vignette equivalence to hold. In fact, when we further distinguish by region and gender (last four rows of the table), the overidentifying restrictions are never rejected at the 5% level, and are rejected at the 10% level only for non-Mediterranean women.

5 Conclusions

Vignette questions have been introduced in several household surveys (SHARE, HRS, ELSA, WHS) and are increasingly used in various fields as an instrument to anchor response scales and allow comparisons across individuals. Reliability of this approach hinges crucially on the validity of the key assumptions of response consistency and vignette equivalence (King et al. 2004). In this paper we introduce a simple joint test of these two assumptions by exploiting the fact that, as pointed out by Deaton (2010), the statistical model is overidentified under these two assumptions. Our Monte Carlo results show that the proposed test has good size and power properties in finite samples.

Using data from the first wave of the Survey of Health, Ageing and Retirement in Europe (SHARE), we apply our test to self-assessment of pain. We find that, in several cases, the overidentifying restrictions imposed by the assumptions of response consistency and vignette equivalence are rejected. This typically occurs when we use more than one vignette question or, as also argued by Rice et al. (2012), when the model specification is not rich enough to fully account for individual heterogeneity. These results suggest that the assumption of vignette equivalence is perhaps more problematic, but also that care is needed with model specification because vignette equivalence may be violated because of failure to properly control for heterogeneity across respondents. In fact, when we carry out the test separately for subgroups of respondents distinguished by gender, age group, health status, education and region, the evidence against the overidentifying restrictions becomes weaker, especially for men and for people who are less healthy, more educated, or live in Mediterranean countries.

Overall, our results confirm the importance of testing the validity of the vignette approach used for identifying and correcting interpersonal incomparability of answers to subjective survey questions. Our results also point to the fruitfulness of exploring new research directions. One direction is vignette design, in particular how to minimize the risk that the vignettes may be interpreted differently. Another direction is extensions to semi-parametric or nonparametric settings. Relaxing distributional assumptions will also avoid the risk that the test rejects because of problems with the assumed parametric specification.

References

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," University of Padua, mimeo.

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

Bago d'Uva T., Lindeboom M., O'Donnell O., and van Doorslaer E. (2011). "Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity," *Journal of Human Resources*, 46: 872–903.

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

Brant R. (1990). "Assessing proportionality in the proportional odds model for ordinal logistic regression," *Biometrics*, 46: 1171–1178.

Case A., and Deaton A. (2005), "Broken down by work and sex: How our health declines", in Wise D.A. (ed.), *Analyses in the Economics of Aging*, Chicago: University of Chicago Press.

Case A., and Paxson C. (2005). "Sex differences in morbidity and mortality", *Demography*, 42: 189–214.

Cunha F., Heckman J.J., and Navarro S. (2007). "The identification and economic content of ordered choice models with stochastic thresholds," NBER Technical Working Paper 340.

Datta Gupta N., Kristensen N., and Pozzoli D. (2010). "External validation of the use of vignettes in cross-country health studies," *Economic Modelling*, 27: 854–865.

Deaton A. (2010). "Comment on 'Work Disability, Work, and Justification Bias in Europe and the U.S.'," in David A. Wise (ed.), *Explorations in the Economics of Aging*, University of Chicago Press, forthcoming.

Ferguson T.S. (1996). *A Course in Large Sample Theory*, Chapman & Hall, London.

Greene W.H., and Hensher D.A. (2010). *Modeling Ordered Choices. A Primer*, Cambridge University Press, New York.

Holland P.W., and Wainer H. (1993). *Differential Item Functioning*, Lawrence Erlbaum, Hillsdale (NJ).

Kapteyn A., Smith J., and van Soest A. (2007). "Vignettes and self-reports of work disability in the United States and the Netherlands," *American Economic Review*, 97: 461–473.

Kapteyn A., Smith J., van Soest A., Voňková H. (2011). "Anchoring vignettes and response consistency," RAND Working Paper 840.

King G., 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: 191–207.

King G., Lau O., and Wand J. (2009). "Anchors: Software for anchoring vignette data," *Journal of Statistical Software*, forthcoming.

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.

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

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

Pudney S., and Shields M. (2000). "Gender, race, pay and promotion in the British nursing profession: Estimation of a generalized ordered probit model," *Journal of Applied Econometrics*, 15: 367–399.

Rantanen T., Guralnik J.M., Foley D., Masaki K., Leveille S.G., Curb J.D., et al. (1999), "Midlife hand grip strength as a predictor of old age disability", *Journal of the American Medical Association*, 21: 558–560.

Rice N., Robone S., and Smith P.C. (2012). "Vignettes and health systems responsiveness in cross-country comparative analyses," *Journal of the Royal Statistical Society*, 175: 1–21.

Salomon J.A., Tandon A., and Murray C.J.L. (2004). "Comparability of self rated health: Cross sectional multi-country survey using anchoring vignettes," *British Medical Journal*, 328: 258–260.

Sen A. (2002). "Health: Perception versus observation," *British Medical Journal*, 324: 860–861.

Terza J. (1985). "Ordered probit: A generalization," *Communications in Statistics*, 14: 1–11.

van Soest A., Delaney L., Harmon C., Kapteyn A., and Smith J.P. (2011). "Validating the use of anchoring vignettes for the correction of response scales differences in subjective questions," *Journal of the Royal Statistical Society–Series A*, 174: 575–595.

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

Table 1: Monte Carlo rejection frequencies for tests of asymptotic 5% level. The number of thresholds is $R = 2$ and the number of runs is 1,000 per experiment.

	$k=1, J=1$ $q-p=1$			$k=2, J=1$ $q-p=2$			$k=1, J=2$ $q-p=3$			$k=2, J=2$ $q-p=6$		
	n			n			n			n		
	250	500	1000	250	500	1000	250	500	1000	250	500	1000
H_0	.057	.059	.050	.050	.042	.053	.056	.059	.043	.053	.052	.052
H_1												
$\beta_1 = .1$	.062	.052	.051	.051	.047	.052	.056	.065	.053	.052	.059	.052
$\beta_1 = .2$	.070	.060	.063	.043	.051	.052	.055	.057	.079	.053	.071	.052
$\beta_1 = .4$	.057	.079	.055	.068	.056	.052	.080	.128	.185	.070	.081	.148
$\beta_1 = .6$	.067	.062	.053	.051	.057	.057	.107	.207	.428	.076	.158	.328
$\beta_1 = .8$	.055	.057	.050	.051	.053	.054	.172	.337	.663	.142	.291	.587
$\beta_1 = 1$	.068	.073	.070	.065	.046	.045	.254	.543	.903	.188	.463	.774
H_2												
$\delta_{11} - \delta_{01} = .1$	.055	.054	.054	.054	.059	.057	.052	.068	.050	.054	.058	.056
$\delta_{11} - \delta_{01} = .2$	.059	.067	.067	.064	.076	.100	.047	.046	.059	.057	.067	.059
$\delta_{11} - \delta_{01} = .4$	.059	.074	.047	.102	.139	.244	.061	.091	.143	.065	.088	.129
$\delta_{11} - \delta_{01} = .6$	.074	.067	.080	.124	.257	.425	.103	.137	.263	.071	.130	.210
$\delta_{11} - \delta_{01} = .8$	.059	.066	.065	.181	.372	.654	.104	.227	.369	.096	.190	.378
$\delta_{11} - \delta_{01} = 1$	.056	.064	.055	.256	.501	.808	.151	.260	.518	.103	.240	.518
$\delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = .1$	.099	.125	.204	.051	.062	.062	.049	.054	.061	.046	.057	.066
$\delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = .2$	.181	.254	.407	.067	.096	.129	.054	.062	.088	.054	.069	.129
$\delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = .4$	.296	.495	.754	.108	.146	.317	.065	.078	.142	.091	.153	.341
$\delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = .6$	.377	.600	.875	.155	.273	.458	.092	.140	.224	.135	.309	.622
$\delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = .8$	.344	.629	.911	.148	.318	.632	.087	.186	.331	.187	.418	.805
$\delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = 1$	.281	.595	.905	.174	.360	.671	.088	.223	.438	.214	.559	.904
H_3												
$\beta_1 = \delta_{11} - \delta_{01} = .1$	.065	.059	.061	.050	.053	.063	.049	.061	.059	.057	.061	.059
$\beta_1 = \delta_{11} - \delta_{01} = .2$	.074	.063	.064	.070	.073	.096	.059	.062	.104	.065	.074	.065
$\beta_1 = \delta_{11} - \delta_{01} = .4$	.068	.078	.074	.110	.144	.260	.095	.170	.270	.089	.123	.232
$\beta_1 = \delta_{11} - \delta_{01} = .6$	.081	.078	.075	.145	.261	.464	.149	.298	.562	.119	.239	.490
$\beta_1 = \delta_{11} - \delta_{01} = .8$	.063	.072	.092	.189	.381	.705	.243	.458	.786	.202	.423	.780
$\beta_1 = \delta_{11} - \delta_{01} = 1$	.057	.084	.098	.291	.555	.854	.312	.628	.941	.269	.590	.899
$\beta_1 = \delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = .1$	.094	.125	.186	.058	.053	.071	.052	.062	.078	.048	.059	.074
$\beta_1 = \delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = .2$	.168	.262	.413	.066	.103	.137	.069	.099	.147	.070	.103	.195
$\beta_1 = \delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = .4$	.307	.525	.809	.105	.148	.324	.125	.218	.444	.139	.275	.548
$\beta_1 = \delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = .6$	.419	.641	.915	.175	.286	.524	.184	.409	.719	.254	.572	.879
$\beta_1 = \delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = .8$	.467	.758	.963	.192	.367	.681	.251	.591	.880	.382	.724	.985
$\beta_1 = \delta_{11} - \delta_{01} = \gamma_{11} - \gamma_{01} = 1$	.485	.801	.972	.244	.459	.760	.324	.654	.958	.474	.860	.996

Table 2: SHARE sample size by country and gender (people aged 50–80). The full sample includes all 50–80 respondents, the vignette sample includes all 50–80 respondents who answer the vignette questions, the working sample includes the respondents in the vignette sample with no missing data on any of the variables used in our analysis.

	Full sample		Vignette sample		Working sample	
	Men	Women	Men	Women	Men	Women
Belgium	1,602	1,791	234	291	201	244
France	1,270	1,484	352	451	301	368
Germany	1,323	1,448	211	264	168	210
Greece	1,154	1,277	317	298	285	254
Italy	1,077	1,295	189	229	149	184
Netherlands	1,272	1,402	242	257	213	216
Spain	900	1,194	185	238	173	202
Sweden	1,285	1,439	186	203	141	149
Total	9,883	11,330	1,916	2,231	1,631	1,827

Table 3: Correlation between self-rated general health and self-assessments on the various health domains. The last column shows estimated coefficients from an ordered probit (OP) model for self-rated general health on self-assessments of health on the six domains.

	Self-rated health	Pain	Sleeping problems	Mobility problems	Concentr. problems	Shortness of breath	Depression	OP coeff.
Self-rated health	1.000							
Pain	.443	1.000						.513 (.040)
Sleeping problems	.292	.415	1.000					.109 (.033)
Mobility problems	.446	.537	.371	1.000				.487 (.039)
Concentr. problems	.245	.340	.304	.339	1.000			.046 (.037)
Shortness of breath	.281	.306	.241	.383	.298	1.000		.178 (.038)
Depression	.291	.378	.399	.353	.391	.329	1.000	.112 (.035)

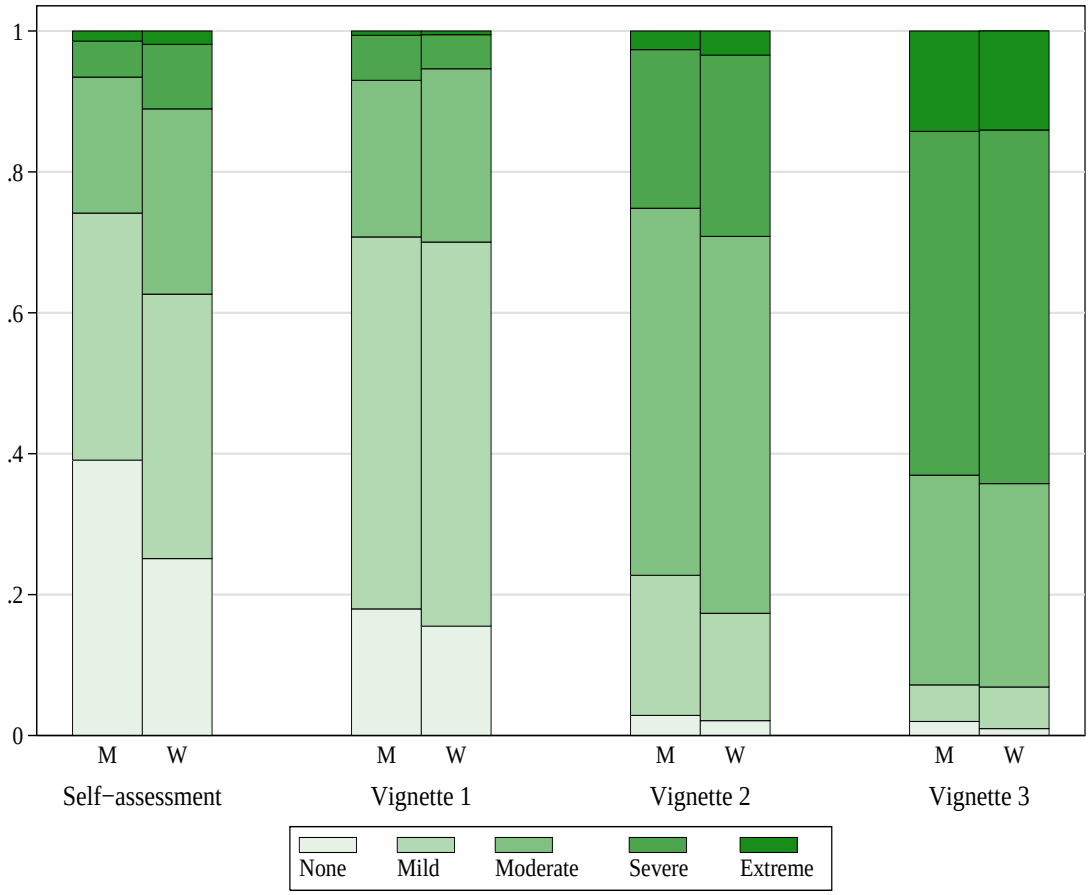
Table 4: Tests of response consistency and vignette equivalence. All respondents ($n = 3,458$).

	k	R	J	q	p	$q-p$	χ^2	p-value
All vignettes	6	2	3	56	26	30	67.4	.000
Only vignette 1	6	2	1	28	22	6	8.7	.188
Only vignette 2	6	2	1	28	22	6	8.6	.200
Only vignette 3	6	2	1	28	22	6	12.6	.051
Vignette 1 and 2	6	2	2	42	24	18	43.2	.001
Vignette 1 and 3	6	2	2	42	24	18	52.0	.000
Vignette 2 and 3	6	2	2	42	24	18	20.5	.305
All 5 categories	6	4	3	112	40	72	468.6	.000

Table 5: Tests of response consistency and vignette equivalence. Subgroups of respondents.

	n	k	R	J	q	p	$q-p$	χ^2	p-value
Men	1,631	5	2	3	48	23	25	33.1	.129
Women	1,827	5	2	3	48	23	25	40.7	.025
Aged 50–64	2,152	6	2	3	56	26	30	48.0	.020
Aged 65–80	1,306	6	2	3	56	26	30	32.4	.347
No conditions	995	5	2	3	48	23	25	38.2	.044
Any condition	2,463	5	2	3	48	23	25	31.3	.180
Less than secondary educ.	1,834	5	2	3	48	23	25	49.8	.002
Secondary and post-sec. educ.	1,624	5	2	3	48	23	25	22.8	.591
Mediterranean countries	1,247	6	2	3	56	26	30	36.1	.204
Non-Mediterranean countries	2,211	6	2	3	56	26	30	49.2	.015
Mediterranean men	607	5	2	3	48	23	25	28.3	.294
Mediterranean women	640	5	2	3	48	23	25	12.1	.985
Non-Mediterranean men	1,024	5	2	3	48	23	25	21.4	.671
Non-Mediterranean women	1,187	5	2	3	48	23	25	35.1	.087

Figure 1: Histograms of self-assessments and answers to the vignette questions on pain by gender.



Appendix

A Structure of the function g and its Jacobian matrix

Write the vector of $p = k + R(k+1) + 2J$ "free" parameters in the model as $\psi = (\rho, \sigma)$, where ρ is the $(p - J)$-subvector of ψ containing the parameters entering the function g linearly and $\sigma = (\sigma_1, \ldots, \sigma_J)$ is the J-subvector of ψ containing the scale parameters entering g non-linearly. Then, the relationship between the reduced-form parameters in π and the "free" parameters in ψ may be written

$$\pi = g(\psi) = A(\sigma)\,\rho,$$

where $A(\sigma)$ is a $q \times (p - J)$ matrix that does not depend on ρ. The $p \times q$ Jacobian matrix of $g(\psi)$ is then

$$G(\psi) = \frac{\partial g(\psi)}{\partial \psi} = \left[\begin{array}{c} \frac{\partial g(\psi)}{\partial \rho} \\ \frac{\partial g(\psi)}{\partial \sigma} \end{array}\right] = \left[\begin{array}{c} A(\sigma)^\top \\ \rho^\top A_1(\sigma)^\top \\ \vdots \\ \rho^\top A_J(\sigma)^\top, \end{array}\right].$$

where $A_j(\sigma) = \partial A(\sigma)/\partial \sigma_j$ is a $q \times (p - J)$ matrix.

To illustrate, in the special case of three response categories ($R = 2$), one exogenous regressor ($k = 1$) and one vignette ($J = 1$), the vector of $q = 8$ reduced-form parameters is

$$\pi = (\gamma^*_{00}, \delta^*_{00}, \gamma^*_{01}, \delta^*_{01}, \gamma^*_{10}, \delta^*_{10}, \gamma^*_{11}, \delta^*_{11}).$$

Let $\psi = (\rho, \sigma)$ be the vector of $p = 7$ "free" parameters, where $\rho = (\alpha_0, \beta_0, \delta_{00}, \gamma_{01}, \delta_{01}, \alpha_1)$ and $\sigma = \sigma_1$. In this case, the relationship between π and ψ can be rewritten as $\pi = g(\psi) = A(\sigma_1)\rho$, where

$$A(\sigma_1) = \left[\begin{array}{cccccc} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1/\sigma_1 \\ 0 & 0 & 1/\sigma_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/\sigma_1 & 0 & -1/\sigma_1 \\ 0 & 0 & 0 & 0 & 1/\sigma_1 & 0 \end{array}\right].$$

The 7×8 Jacobian matrix of $g(\psi)$ is then

$$G(\psi) = \left[\begin{array}{c} A(\sigma_1)^\top \\ \rho^\top A'(\sigma_1)^\top \end{array}\right],$$

where

$$A'(\sigma_1) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/\sigma_1^2 \\ 0 & 0 & -1/\sigma_1^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1/\sigma_1^2 & 0 & 1/\sigma_1^2 \\ 0 & 0 & 0 & 0 & -1/\sigma_1^2 & 0 \end{bmatrix}.$$

B Vignette questions for pain

The vignette questions for pain are the following.

1. *"Paul/Karen has a headache once a month that is relieved after taking a pill. During the headache he/she can carry on with his/her day-to-day affairs."*
2. *"Henri/Maria has pain that radiates down his/her right arm and wrist during his/her day at work. This is slightly relieved in the evenings when he/she is no longer working on his/her computer."*
3. *"Charles/Alice has pain in his/her knees, elbows, wrists and fingers, and the pain is present almost all the time. Although medication helps, he/she feels uncomfortable when moving around, holding and lifting things."*

C MD estimates for the ordered response model with heterogeneous thresholds

This appendix presents the MD estimates of the coefficients of the ordered response model with heterogeneous thresholds for pain under the assumptions of response consistency and vignette equivalence (* significant at 5%, ** significant at 1%).

	Self-assessment	Threshold 1	Threshold 2
Any condition	.527 **	-.085 **	-.033
Grip strength - 34.9	-.018 **	-.004 *	-.001
Age - 55	.001	.000	.003 *
Post-secondary education	-.177 **	-.145 **	-.047
Log household income	-.098 **	-.063 **	-.067 **
Female	.006	-.181 **	-.013
Constant	-.084	.000	1.776 **
	Vignette 1	Vignette 2	Vignette 3
Constant	.560 **	1.337 **	2.256 **
$\ln(\sigma)$	-.283 **	-.286 **	.057